وزارة التعليم العالي والبحث العلمي

Université Badji Mokhtar –Annaba

Badji Mokhtar – Annaba University

جامعة باجي مختار- عنابة -

Année 2017

Faculté des Sciences de l'Ingéniorat

Département d'Informatique

T H È S E

En vue de l'obtention du diplôme de

Docteur 3éme cycle

Analyse et recherche dans les réseaux sociaux:

Vers la caractérisation et l'identification significative d'une identité de structure noyau possible au sein d'un processus évolutionnaire décrivant la dynamique d'un réseau social

Filière : Informatique

Spécialité : Sciences et Technologies de l'Information et de la Communication

Préparée par

Bilel Hamadache

Jury :

Présidente:	Labiba Souici	Pr. Université Badji Mokhtar, Annaba.
Directeur de thèse :	Nadir Farah	Pr. Université Badji Mokhtar, Annaba.
Co-directrice de thèse :	Hassina Seridi-Bouchelaghem	Pr. Université Badji Mokhtar, Annaba.
Examinatrice :	Catherine Faron-Zucker	HDR. Université Sophia Antipolis, Nice (France)
Examinateur :	Yacine Lafifi	Pr. Université 08 Mai 1945, Guelma.
Examinateur :	Mohamed Tarek Khadir	Pr. Université Badji Mokhtar, Annaba.

À ma mère qui a sacrifié toute sa vie pour moi

“L'homme fort dit : je suis. Et il a raison. Il est.

L'homme médiocre dit également : je suis. Et lui aussi a raison. Il suit.”

- Victor Hugo –

"Les miracles commencent à se produire lorsque nous investissons autant d’énergie dans nos rêves que dans nos peurs".

- Richard Wilkins -

REMERCIEMENTS/ACKNOWLEDGEMENTS

Je fais partie des personnes qui ne croient pas aux miracles sans bosser, car "à vaincre sans péril, on triomphe sans gloire", mais tout en croyant aussi à la bonté de Dieu qui donne la chance de réussite, et je commence par Le remercier au terme de cette thèse.

Je tiens à remercier mon directeur de thèse monsieur le professeur Nadir Farah pour sa supervision et son soutien. Je remercie très chaleureusement ma co-directrice de thèse, madame la professeure Hassina Seridi pour son soutien, étant une source de motivation tout au long de ce travail, pour ses précieux conseils qui m'ont appris comment mener des travaux de recherche, et cibler des thèmes de recherche d'actualité, en transformant ainsi les difficultés rencontrées en une expérience enrichissante. Je lui suis reconnaissant de ses interventions et de rendre ce projet à terme.

Je dois exprimer aussi mes reconnaissances aux examinateurs qui ont accepté de siéger au jury de cette thèse et fourni des commentaires judicieux, me permettant de raffiner davantage certains points.

C'est le moment d'exprimer aussi mes plus sincères remerciements et gratitudes à certaines personnes, notamment les docteurs Hocine Bekkouche, Yacine Ayad, et Fayçal Hamdi, sans oublier aussi Mohamed Bayat, Zoheir Hamizi et Abdelhamid Loukil. Je leur suis redevable à leur soutien, accueil et leurs conseils, de près ou de loin.

J'adresse particulièrement des remerciements, plein d'amour et de reconnaissance à ma mère, mon père et ma sœur. Mais, les mots me manquent et je ne vois pas comment remercier ma chère maman pour les soutiens moral et psychologique indispensables pour maintenir la sérénité au travers des aléas de la vie et pour avoir cru en moi. Elle m'a constamment encouragé et consacré, à sa manière, énormément de temps et d'énergie, la lumière inestimable, qui m'a poussé pour aller vers l'avant dans ma vie et dans ce parcours académique jusqu'à ce projet qui n'aurait pas vu le jour sans elle.

Enfin en tout et pour tout, je remercie encore une fois Mon Dieu qui me guide et qui me protège.

Bilel.

Résumé

Les réseaux sociaux en ligne se prolifèrent et se diversifient dramatiquement dans différents environnements, permettant de se socialiser, être plus participatifs, de se regrouper, de partager et d'interagir. En parallèle, l'analyse des réseaux sociaux qui cherchait classiquement à découvrir les rôles de leaderships, les structures communautaires, etc. profite de la variété des données sociales et s'évolue également. Mais on s'attend à ce que ces nouveaux progrès soient au service de la gestion, l'innovation des organisations et les challenges des investigations, en se focalisant dans ce travail sur des réseaux qui émergent dans les intranets des organisations et les plateformes de collaborations. Même s'ils sont noyés dans le web social, ces réseaux sociaux doivent être spatialement ou contextuellement référencés. C'est notre point de départ à la recherche des interprétations plus significatives, plus bénéfiques. Nous cherchons dans cette thèse à sonder profondément dans ce type de réseaux afin de comprendre des phénomènes complexes, voir les structures sous-jacentes qui peuvent se produire derrière les besoins de partage d'informations, la durabilité des interactions, dynamique des groupes ainsi que leur sémantique. De ce fait, et à partir des motifs d'analyse ou de fouille, bien motivés, nous concevons des approches plus conceptuelles que statistiques. Nous ajoutons ainsi plus de dimensionnalité : Dynamicité temporelle ou richesse sémantique qui exigerons la définition de méta-modèles de réseaux sociaux. En particulier, nous essayons d'aller au-delà des conceptions statiques pour caractériser, modéliser et révéler une identité significative d'un noyau dominant le processus évolutionnaire d'un réseau social d'une organisation. Une classe élite qui affichera un comportement typique qui réunit tous les concepts de persistance, de centralité et de stabilité de centralité dans le temps. D'autre part, cela nous ouvre la voie pour étudier quelques aspects de modélisation sémantiques qui seront abordés dans le domaine métier du e-learning Social, au-delà des représentations topologiques. Nous verrons que le potentiel, la connectivité, l'esprit de collectivité seront paramétrés et varient sémantiquement selon différents points de vue.

Mot clés: Réseaux sociaux, analyse & fouille des réseaux sociaux, Pertinence des données sociales, centralité, Détection des communautés, Structure noyau sous-jacent, Dynamicité temporelle, dynamique des groupes, Chevauchement temporel, Durabilité, Stabilité de centralité, Identité significative de noyau, Sémantique des interactions, E-learning social, Méta-modèles, Analyse paramétrée, Sémantique des collectivités.

Abstract

Online social networks proliferate and diversify dramatically in different environments, where people are increasingly able to socialize, to be more participatory and to form groups, to share and interact. In parallel, the social network analysis that sought classically to find the leadership roles, community structures, etc., is benefitting from the variety of social data and is evolving also. But it is expected that the advances and prospects help the organization's management, innovation, and investigations, specifically when we are focused on networks emerging in corporate intranets and collaboration platforms. Even if they are drowned in the social web, these social networks should be spatially or contextually referenced. This is our starting point looking for more significant and beneficial interpretations. In this thesis, we seek to probe deeply in this kind of networks in order to understand complex phenomena including underlying structures resulting from the information sharing needs, sustainability of interactions, group dynamics and their semantics. In this respect and from serious and well-motivated analysis or mining grounds, we design much more conceptual approaches than statistic. We aim to add more dimensionality based on temporal dynamicity or semantic richness that will rather require the definition of meta-models. In particular, we try to move beyond static conceptions and characterize, model and reveal significant identity of a core structure dominating the evolutionary process of an organizational social network. This is an elite class displaying typical behavior such that all concepts of persistence of centrality and centrality stability are met over time. On the other hand, it opens the way to study some semantic aspects that we will discuss in the social e-learning context, beyond the topological representations. The individual potential, connectivity, community spirit will be shown parameterized and vary depending semantically on different viewpoints.

Key words: Social network; Social network analysis & mining (SNA/ SNAM); Relevant social data; Centrality; Community detection; Underlying structure; Temporal dynamicity; Group dynamics; Temporal overlap; Durability; Centrality stability; Significant core identity; Interaction semantics; Social e-learning; meta-models; Enriched analysis; Collectivity semantics.

ملخص

تتكاثرالشبكات الاجتماعية عبرالنت في بيئات مختلفة و تتنوع بشكل دراماتيكي، بحيث ان قدرة الاشخاص و المنظمات على التفاعل و التواصل الاجتماعي في تزايد مستمر، عِلاوة على تفعيل روح و عقلية المشاركة وتشكيل المجموعات. في موازاة ذلك، التحليل الكلاسيكي للشبكات الاجتماعية و الذي يمكن من الكشف على الأدوار القيادية، وكذا هياكل المجتمعات في الشبكة، وما إلى ذلك، يستفيد بدوره من مجموعة متنوعة من البيانات الاجتماعية مما يجعله أيضا بصدد التطور. لكن من المنتظر الان أن الآفاق الجديدة وتطور البحوث في هذا المجال تساعد تسيير وتحسين المنظمات ذات الصبغة الاجتماعية و كذا المساعدة في التحقيقات، تحديدا عندما يتعلق الامر بالشبكات الاجتماعية الناشئة في الشبكات الداخلية للشركات ومنصات التعاون. ان مثل هذه الشبكات حتى ولو كانت منغمرة في الويب الاجتماعي، يجب أن تكون لها مرجعية تتعلق بمحيطها و سياقها. نحن نعتمد على هذا المنطلق في سبيل البحث عن تفسيرات أكثر واقعية وفائدة. في هذه الأطروحة، نحن نسعى إلى تعميق مستوى البحث في هذا النوع من الشبكات على بعض الظواهر المعقدة و التركيبات الكامنة التي تنتج وراء الحاجة في تقاسم المعلومات، استدامة التفاعلات، ديناميكيات المجموعات و ما تخفي من دلالات. من اجل ذلك، و بناءا على التعريف بفكرة تحليل وتنقيب مبنية على دوافع موضوعية، نحن نصمم مناهج مبنية على مفاهيم أكثر منها احصائية. نحن نهدف ايضا إلى إضافة بعد ديناميكي زمني أو بعد الثراء الدلالي اللذان يتطلبان مستوى تصميم ونمذجة اعلى للشبكات. على وجه الخصوص، نود تجاوز التمثيل الثابت لبنية نواة بحيث نميز، نمثل و نكشف عن هوية ذات طابع زمني اكثر واقعية و تعبير لبنية نواة تهيمن في عمق العملية تطورية للشبكة الاجتماعية لمنظمة ما. الامر يتعلق بالبحث عن فئة نخبوية تظهر سلوك نموذجي يستوفي جميع مفاهيم الاستمرارية، المركزية والاستقرار في المركزية عبر الزمن. من ناحية أخرى، فإن ذلك يفتح لنا الطريق لدراسة ايضا بعض الجوانب الدلالية التي نناقشها في سياق التعلم الإلكتروني الاجتماعي متجاوزين التمثيل الطوبوغرافي الفقير دلاليا. سنبين ان لنشاط و مركزية الافراد، مدى الترابط والتجانس ، وعقلية المجموعات تتغير دلاليا حسب وجهات نظر مختلفة.

كلمات مفتاحية للبحث : الشبكات الاجتماعية، تحليل والتنقيب في الشبكات الاجتماعية، بيانات اجتماعية ذات مرجعية، المركزية، الكشف عن المجموعات، بنية كامنة ضمنية، الديناميكية الزمنية، ديناميكيات المجموعة، تداخل المجموعات عبر الزمن، الاستمرارية و المتانة، الاستقرار في المركزية، هوية اكثر تعبير للنواة، دلالات التفاعل، التعلم الإلكتروني الاجتماعي، تحليل مخصب، دلالات المجموعات

Publications

Hamadache B., Seridi-Bouchelaghem H. & Farah N. "A significant core structure inside the social network evolutionary process." Soc. Netw. Anal. Min. (Springer) (2016): (2016) 6: 38. doi:10.1007/s13278-016-0344-y.

Hamadache B., and Seridi-Bouchelaghem H. "How to analyse a semantic social network of learners, in a social learning environment?" Int. J. Web Based Communities: Open Web Communities for Social Evolution (2016): Vol. 12, No. 4.

Hamadache B., Seridi-Bouchelaghem H., and Farah N. "An elite grouping of individuals for expressing a core identity based on the temporal dynamicity or the semantic richness." Chapter in Social Network Analysis – Community Detection and Evolution (Springer) (2014–2015): pp.119–143.

Hamadache B., Seridi-Bouchelaghem H., and Farah N. "Toward Expressing a Preliminary Core Identity Significantly Characterized from the Social Network Temporal Dynamicity." Third International Conference, MEDI 2013, (September 25-27). Amantea, Calabria, Italy: Springer, 2013. pp: 149-161.

Hamadache B., Seridi-Bouchelaghem H., Farah N. "Toward characterizing a more significant identity of core structure within dynamic social network." The 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013 , (August 25-28). Niagara Falls, Canada: IEEE/ACM, 2013. Pages 1458-1459.

Hamadache B., Seridi-Bouchelaghem H., Farah N. "Analyse et recherche dans les réseaux sociaux." Les deuxièmes journées JDI 2012 - Session E-Technologies (18-19 Novembre). Guelma, Algérie, 2012. pp 33.

Hamadache B., Seridi-Bouchelaghem H., and Farah N., "Social Network Analysis and Mining, New dimensions and tendencies: Dynamics & Semantics", In Proc. Of JED, 2014.

Sommaire

Liste des Figures

Liste des Tableaux

Liste des Algorithmes

Introduction générale

Aujourd'hui, les sociétés vivent dans un monde connecté. Étant le modèle d'information et de connaissances le plus exploité dans ces dernières années, les réseaux constituent aujourd'hui la base de la société moderne, dans les systèmes, biologiques, physiques, sociaux et informatiques. Ces réseaux ou bien 'networked systems' sont typiquement des systèmes complexes, formés par des agents individuels qui interagissent souvent et affichent des caractéristiques et des comportements émergents. Les réseaux de communication notamment les réseaux sociaux sont étroitement liés à la vie quotidienne et font partie de ces réseaux complexes qui se présentent par des modèles mathématiques, des graphes sociaux, dont l'efficacité est tributaire des techniques et des méthodologies d'analyse de leur structure, leur évolution, etc. Le domaine d'analyse des réseaux sociaux, est une voie de recherche interdisciplinaire innovante, là où la théorie des graphes, la fouille de graphes, la statistique, la sociologie, la psychologie sociale, etc. se croisent. Beaucoup de métriques et algorithmes permettent de fournir formellement des réponses à la socialisation des individus, en découvrant leur potentiel (centralité) sur les flux de communication, ceux qui occupent des positions stratégiques et jouent des rôles de leaderships, ainsi que la connectivité du réseau social et ses structures communautaires, etc.

Mais aujourd'hui, on ne se contente pas seulement d'analyser statistiquement le réseau social. Les motifs d'analyse sont évolués, des tendances émergent et des nouveaux aspects sont ajoutés en se dirigeant plutôt vers une fouille, en parallèle de l'émergence, la prolifération et la diversité inarrêtable des réseaux sociaux en ligne. En effet, la communication médiatisée par ordinateur est devenue de manière générale une partie intégrante de la socialisation et le développement de l'individu et son environnement. Les nouvelles technologies et outils collaboratifs, les plateformes, services de réseautage et médias sociaux sur le web sont révolutionnaires mais surtout socialement ouverts, amplifient les interactions sociales entre les personnes, les communautés, etc. Cet aspect social moderne est devenu aussi omniprésent et déployé au sein des organisations, entreprises, environnements d'apprentissage collaboratifs, entre les employés, les collaborateurs, etc., en favorisant le caractère participatif de leurs activités afin atteindre efficacement les objectifs tracés. Cependant, si la disponibilité des données sociales est aujourd'hui en croissance, leur pertinence et richesse qui sont des notions plus profondes et ambitieuses ne sont pas forcément assurées.

Pour qu'on puisse multiplier le bénéfice informationnel et tirer le maximum de l'analyse et la fouille des réseaux, les contributions présentées dans cette thèse se basent tout d'abord sur la pertinence qui signifie que les données doivent être spatialement référencées, sachant que les comportements sociaux des individus sont directement influencés par leurs contextes sociaux, leur l'environnement. Notre orientation de recherche s'intéresse à des réseaux sociaux plus implicites, sous-jacents, évoluant dans des organisations, issus des corpus d'emails ou produits sur des plateformes de collaboration. Ils sont tout de même au

service de l'innovation, la recherche dans la sociologie des organisations économiques, politiques, etc., et les investigations menées sur les réseaux frauduleux criminelles et dans la lutte anti-terroriste, etc. C'est dans ce sens que nous voulons sonder profondément en cherchant à construire des méta-modèles pour fournir des interprétations précises et plus réalistes au sujet des phénomènes plus complexes notamment l'émergence des classes dominantes, le comportement dynamique et durabilité, sémantique des collectivités et des élites comme une structure noyau sous-jacente. Par conséquent, nos approches proposées qui exprimerons plus de conceptualisation que de scalabilité exigent plus de richesse de données. La dynamique temporelle ou la richesse sémantique de ces réseaux sociaux seront adoptés pour sortir du cadre classique (statique ou topologique) de l'analyse en ajoutant plus de dimensionnalité à nos études et arriver ainsi à obtenir des interprétations plus significatives. Nous sommes conscients du contexte dynamique endogène qui laisse apparaitre l'évolution du réseau social comme un processus de développement dans le temps. Nous synthétisons les concepts liés selon la capacité des modèles d'intégrer la composante temporelle et de reproduire et comprendre les propriétés d'évolution des individus ainsi que la dynamique des groupes.

En s'inspirant de cette dynamique temporelle, nous proposons une approche méthodologique pour caractériser, modéliser et identifier une identité plus significative d'un noyau dans un réseau social (lié au scandale d'Enron) évoluant dans le temps.

Cette dernière va montrer comment la dynamique temporelle topologique, notamment des collectivités, peut constituer une passerelle pour aborder une première ligne de sémantique qui règne derrière. Devant la multiplicité de contextes des relations, des intérêts, etc., la topologie d'un réseau ne présente que le flux d'information et les patterns de relations. À cet égard, nous présenterons des modèles sémantiquement plus explicites, avant de proposer une autre approche de modélisation sémantique pour paramétrer l'analyse d'un réseau social.

Scénario de motivation et cadre du travail, problématiques & objectifs

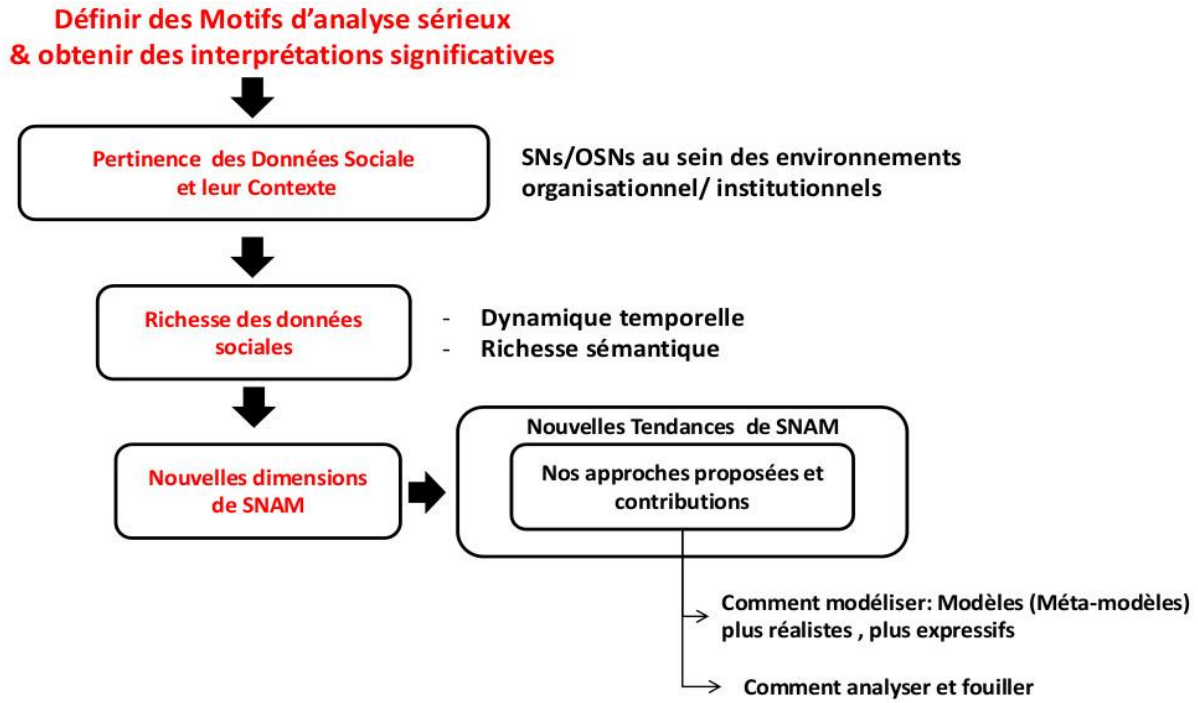


Figure 1. Squelette du scénario de motivation générale derrière le cadre et l'orientation du travail de recherche, à la lumière de nos contributions

La **Figure 1** représente les étapes principales pour le scénario de motivation derrière notre travail de recherche dans le sens des nouvelles tendances du SNAM. Chaque étape dégage une ou des problématiques et challenges. L'objectif global consiste à obtenir des interprétations plus significatives partant essentiellement de la recherche et la définition des motifs d'analyse ou de fouille. Des motifs qui nous poussent à explorer le réseau social (SN) en profondeur suivant une orientation conceptuelle plutôt que statistique pour ne pas diminuer notre capacité à sonder profondément dans le réseau, dans son processus évolutionnaire, dans ses groupes, dans ses structures sous-jacentes (structure noyau), et même dans ses aspects sémantiques (derrière les relations, des collectivités, les comportements dynamiques, etc.).

Nous nous sommes intéressés au départ par la recherche d'une identité plus significative d'une structure noyau dans un SN. Etant un sujet phare et une notion commune dans les systèmes complexes, le noyau est évoqué dans la reconnaissance des élites, comme une classe leader qui a certain contrôle et de régulation, qui donne au système sa cohérence, son existence. Mais en réalité, moins d'attention est accordée à l'étude d'une structure noyau dans les SNs, organisationnels notamment. En économie par exemple, la détection d'une classe élite, comme le noyau, est importante dans les réseaux de directions ou conseils d'administrations imbriqués 'interlocking directorates'. En effet, sous la forme principale des sociétés commerciales d'aujourd'hui, un actionnaire appelé 'multiple director' peut faire partie du conseil d'administration de plusieurs entreprises. Le noyau est dans ce cas l'ensemble de 'multiple directors' qui n'ont pas seulement une influence économique mais aussi nommés à des postes gouvernementaux. C'est en d'autre terme la classe dominante en politique moderne. En outre, la détection de noyau est attrayante pour les enquêtes menées sur les réseaux illégaux ou criminels et leurs centralisations, des réseaux qui préparent des

attaques terroristes ou qui infiltrent des sociétés et dissimulent des comportements frauduleux. C'est un sujet à aborder aussi dans les réseaux des citations scientifques, etc.

Mais tout d'abord, la pertinence, le contexte des données sociales (**Figure 1**) est un maillon critique qui peut amplifier le bénéfice informationnel pour toute contribution que nous souhaitons apporter. La pertinence sera tirée de l'environnement là où le SN émerge ou évolue. Quels acteurs ? Pour quels usages ? Les liens sont-ils pertinents ? Cette notion nous empêchera de tomber sur la probabilité croissante que les OSNs d'aujourd'hui sont envahis par des interactions plus ou moins pertinentes, comprenant de plus en plus des liens de connaissances éphémères. Nous avons ainsi tendance à nous concentrer sur des SNs/OSNs (même implicites) qui émergent au sein des environnements organisationnels. Par exemple en économie, en politique, réseaux illégaux, là où les obligations institutionnelles et sociales sont fusionnées, ou dans les plateformes de collaboration comme les environnements de e-learning social, etc., où l'acte de collaboration est plus orienté, acquis et conditionné par des compétences et contextes sociaux. Par conséquent nous serons en mesure de donner plus de fond à nos contributions et approfondir la compréhension du motif d'analyse ou de la fouille.

Malgré les progrès qu'il a réalisé le domaine de SNA, rares sont les travaux qui se sont intéressés à l'étude de tel phénomène dans tels réseaux, mis à part certaines conceptions plus formelles mais qui restent statiques. Avec le 'blockmodeling', un noyau a été distingué par exemple par un sous-ensemble d'individus densément liés, chacun susceptible d'avoir une sorte de centralité individuelle plus élevés. Mais sous une vue d'ensemble, un noyau s'est présenté comme une zone de chevauchements de communautés, ou dans une configuration core-and-whiskers. Malgré ça, on aurait pu profiter du bénéfice informationnel des centralités de groupes pour évaluer le potentiel d'un noyau en tant qu'une collectivité. La problématique ne s'arrête pas ici.

Nous pensons d'abord que l'étude de cette structure noyau mérite d'être entourée par les concepts, indices et représentations de haut niveau en SNA. Mais le plus important, est que nous avons tendance à ajouter aussi plus de dimensionnalité à nos études partant du fait que les progrès, perspectives (Scott 2011) et les ambitions en SNAM dépassent aujourd'hui les cadres statiques (ou topologiques). De ce fait, nous nous appliquons sur des données sociales plus riches (**Figure 1**), notamment des données temporelles (ou sémantiques) dont la disponibilité ou l'accessibilité est un autre challenge car elle n'est pas à la hauteur des données ordinaires. Pourtant, cette richesse nous parait indispensable dans notre orientation de recherche, et nous inspire de suivre des nouvelles dimensions d'analyse : dimension dynamique temporelle (ou même sémantique), afin de dépasser les inconvénients des conceptions statiques (ou topologiques structurelles) qui sont trompeuses (ou sémantiquement pauvres). Par conséquents, d'autres challenges émergent et exigent avant tout des modèles, voir des méta-modèles plus réalistes, plus expressifs et des approches d'analyse, d'identification plus performantes, en mesure d'exploiter les nouvelles composantes en faveur des interprétations plus significatives (**Figure 1**). Cela aidera de manière générale un modérateur ('SN/Community Manager', enquêteur, etc.) à la gestion, la prise de décision et recommandation plus intelligentes, etc.

Plus précisément, tant que le SN évolue dans le temps, sa dynamique temporelle (notamment la dynamique des groupes) nous inspire à révéler une identité plus significative d'un noyau. Nous proposons une approche méthodologique en trois phases: Caractérisation, Modélisation et identification, basées sur trois principales caractéristiques: cohésion, dominance et résistance qui décrivent théoriquement cette identité. La phase de caractérisation va

conceptuellement mouler cette identité dans le concept groupe, de telle sorte que nous explorons respectivement ses caractéristiques via des paramètres dérivés de la dynamique des groupes: cohésion interne, persistance, centralité du groupe, la stabilité de sa centralité dans le temps. Nous cherchons ainsi des structures sous-jacentes pertinentes par rapport à ces paramètres : Pour combien de temps un groupement d'individus sous-jacent peut persister, s'il joue un rôle central/ efficace dans le réseau et dans quelle mesure son rôle/influence est stable. De ce fait, nous formaliserons dans la phase de modélisation, un méta-modèle d'un SN évoluant dans le temps sous forme de processus évolutif de groupes dans des 'time-steps' successifs. Un arc pondéré impliquera un chevauchement temporel entre deux groupes. La fonction de pondération, étant la composante critique incarnera les paramètres précédents, axés sur ce type de chevauchements. C'est-à-dire, le poids lourd se réfère à une structure (un chevauchement) sous-jacente pertinente : Plus large et plus centrale). Dans une phase d'identification une recherche de patterns 'critical pattern-based research' sera effectuée pour détecter le pattern critique d'évolution: Un chemin couvrant, le plus lourd qui inclut une succession de chevauchements temporels pertinents. C'est là où un regroupement qui peut persister profondément à l'intérieur, affichera un comportement typique de la plus grande composition, durable qui joue un rôle central, le plus stable possible au fil du temps. C'est le pattern d'évolution d'une structure noyau profondément dans le SN.

Ces études vont nous faire penser à quelques aspects sémantiques implicites qui règnent autour de ces comportements dynamiques. Cela va nous guider suivant les mêmes étapes de ce scénario de motivation, vers un autre objectif qui viendra en second plan et qui aborde la sémantique des SNs. Nous somme intéressés par l'un des SNs de collaboration qui émerge dans un environnement différent, dans le contexte e-learning social, là où le potentiel individuel, l'esprit de collectivité peuvent varier sémantiquement et au même temps. Mais, moins d'attention a été accordée à la sémantique des interactions sociales entre les apprenants par à rapport à la sémantique des matériaux dans ces environements. Mêmes dans une configuration dynamique, les représentations topologiques tracent que des flux de circulation d'information et ne permettent pas d'étudier des aspects sémantiques. Donc, cette fois-ci, dans un contexte plutôt statique, nous proposons une méthode de modélisation et d'analyse suivant une dimension sémantique. Nous explorons les propriétés sémantiques des interactions sociales entre apprenants pour identifier les rôles de leadership, les meilleurs collaborateurs à recommander par exemple pour un apprenant donné selon sa positivité, le type de relation préféré. Après avoir modélisé la sémantique de ce SN par un graphe social RDF, son analyse relèvera un autre défi. D'où, nous proposerons une application capable d'exécuter un processus de 'mapping' vers une représentation équivalente qui conserve la même richesse exprimée et l'analyser sans avoir besoin d'outils intermédiaires. Nous serons capables de paramétrer les métriques de centralité et avoir ainsi plusieurs interprétations du profil social de l'apprenant, de son rôle et de ses affiliations (structures modulaires) suivant le type de relations.

Le travail se compose de deux parties principales (**Figure 2**). La première partie contient deux chapitres. Nous synthétisons dans le premier les concepts de représentations, les métriques et algorithmes d'analyse des réseaux sociaux, suivis par la prolifération et la catégorisation des données sociales. Dans le deuxième, nous abordons la dyamique temporelle des SNs et la capacité des modèles d'intégrer la composante temporelle, de reproduire et comprendre les propriétés d'évolution des individus ainsi que la dynamique des groupes. Nous parlons aussi de la sémantique des SNs comme une autre dimenion d'analyse mais elle sera plus étoffée dans les annexes. La deuxième partie sera réservée aux nos contributions, dans le troisième et le quatrième chapitre qui sont séparés par une petite partie intermédiaire.

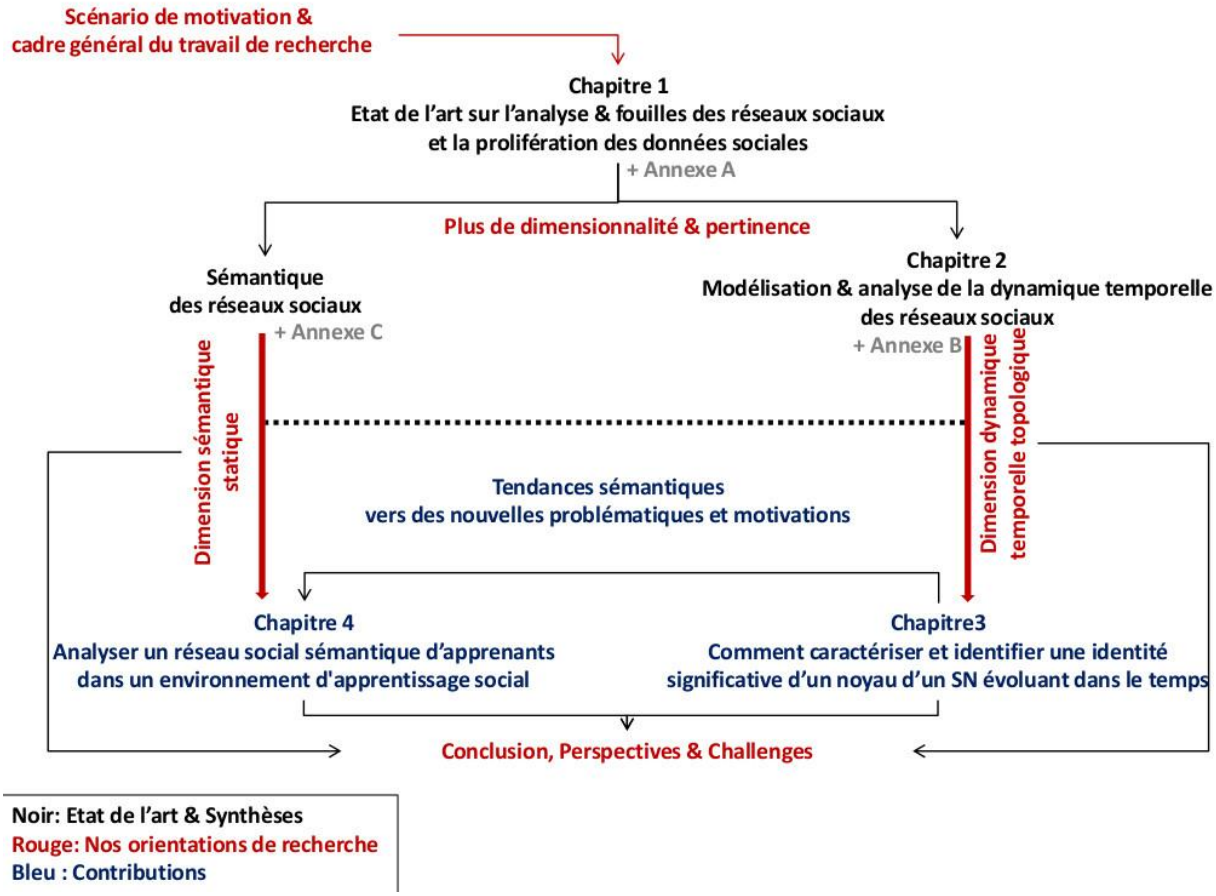


Figure 2. Architecture globale de la thèse

PARTIE I : ÉTAT DE L'ART & SYNTHESES

Chapitre 1 : Etat de l'art sur l'analyse classique & fouille des réseaux sociaux et la prolifération des données sociales

1. Introduction

Le réseau social (SN) est avant tout une représentation d'une structuration sociologique qui s'est développée avec des formalisations mathématiques lisible en mémoire informatique, basées sur la théorie des graphes. Une théorie qui aide à modéliser des liens sociaux détectés et pas seulement des interactions déclarées entre les acteurs d'une population. L'analyse des graphes sociaux a montré qu'il s'agit des graphes non-aléatoires qui affichent des caractéristiques particulières, des distributions suivant une loi de puissance, phénomène de petit monde. Nous allons aborder d'abord le carde classique : statique et topologique de SNA comprenant les métriques, indicateur et algorithme les plus connus selon deux points de vue majeurs. D'un point de vue local, une collection de métriques basées sur le concept de centralité permet d'estimer le potentiel, le rôle, l'influence d'un acteur, d'un lien sur les flux de communication et de révéler ainsi des rôles leaderships, des positions stratégiques individuelles. Les mesures locales seront également catégorisées ainsi que raffinées. Certaines s'appliquent au voisinage d'un nœud, certaines d'autres évaluent sa proximité, son intermédiarité, etc., sur l'ensemble du réseau.

La localité de ce traitement d'analyse n'empêche pas d'avoir à partir des positions stratégiques ou anormales qu'ils occupent certains individus, une vue sur l'organisation globale du SN, sur sa centralisation et sa dépendance avec les acteurs centraux, sur sa connectivité et sa structure modulaire. Dans ce sens, nous verrons que la centralité d'intermédiarité présente un atout majeur dans la détection des individus qui dominent les flux de communication entre les communautés. Les structures communautaires constituent aujourd'hui une propriété significative dans les SNs et améliore l'efficacité de communication, la confiance, etc. Il n'y'a pas un accord complet sur la définition d'une communauté ou un groupe, mais plutôt, une définition de plus en plus flexible, étant des patterns de sous graphes : clique, n-clique, n-clan, k-plex, etc. jusqu'à avoir incarné le concept général de cohésion. Une communauté est identifiée comme une région cohésive dont la qualité standard est topologiquement basée sur le rapport entre liens intracommunautaires et intercommunautaires. Le problème de découverte de communautés a été formulé et traité par une diversité d'algorithmes comme un problème de partitionnement de graphe. Ces algorithmes seront classifiés selon différent critères : la fonction objective à optimiser qui reflète la qualité de décomposition (modularité, KL, etc.), le contrôle de granularité, performance, etc. Nous trouvons par exemple l'extraction agglomérative et plusieurs autres variantes et heuristiques, mais les algorithmes de divisions : 'Girvan & Newman', 'Louvain Method', etc., sont parmi les approches les plus efficaces.

Aujourd'hui il existe une panoplie de données sociales à analyser et fouiller par ces techniques de SNA. Ces jeux de données partant des données collectés par les sociologues, celles extraites par le Web Mining, les collections des réseaux sociaux en lignes (OSN) sur les plateformes sociales jusqu'aux données extraites depuis les usages de web sociales seront classifiés selon des critères. À quel point ces données sont implicites, leur pertinence et richesse sont des critères clés qui peuvent rendre les interprétations d'une étude analytique plus significatives.

2. Réseaux sociaux et représentations

2.1. Qu'est-ce qu'un réseau social dans le monde réel

Un réseau social est communément connu pour représenter des interactions et des structures sociales, des structures organisationnelles et même des proximités physiques (Tantipathananandh et al 2007), (Sociogrammes). Il est composé d'un ensemble de participants, des acteurs : individus, organisations, etc., et des liens mutuels désignant des relations sociales et interdépendances entre participants (Erétéo 2011) (Takes 2011) (Kazienko et al 2011) (Nettleton 2013) (Sudeshna & Birinder 2009). Selon les contextes sociaux et les types d'acteurs, les relations sont de différents catégories et nature et se distinguent spécifiquement entre humains et généralement entre acteurs (Erétéo 2011). Les relations entre humains sont explicites et comprennent toutes liaisons entre personnes (liens familiaux, d'amitié, simple connaissance, Co-travailleur, etc., ou entre personnes et organisations (employées, membres, propriétaires…etc.) (Erétéo 2011). Les relations prennent un sens plus large et peuvent décrire des interactions plus ou moins abstraites (Tantipathananandh et al 2007) ou des affiliations entre acteurs. Les interactions sociales sont des actions d'échanges là ou au moins 2 acteurs sont impliqués: Discussion, collaboration (co-auteur), proximités physiques dans les groupes d'animaux, etc. Tandis que les affiliations sont basées plus sur la similarité entre acteurs: en partageant le même attribut, les mêmes objets, un intérêt commun, les mêmes activités ou organisations, etc., (Erétéo 2011) (Takes 2011) (Sudeshna & Birinder 2009).

Un réseau social peut servir à représenter des phénomènes du monde réel pas nécessairement sociaux: Dans les réseaux électriques, les appels téléphoniques, la propagation des virus informatique (Loiacono 2011), etc. C'est un moyen important pour expliquer certains processus de diffusion d'information, des innovations, etc. ou les épidémies (Tantipathananandh et al 2007). Les réseaux sociaux attitrent de plus en plus l'attention des épidémiologistes, des sociologues, les biologistes (les interactions animales), des communautés de renseignement et enquêteurs (les réseaux frauduleux et terroristes) et notamment des scientifiques informatiques (Tantipathananandh et al 2007).

2.2. Histoire de développement en représentation et analyse

Partant des origines sociologiques, les réseaux sociaux de nos jours est une histoire de développement en représentation, en prolifération et audience ainsi qu'en analyse.

Les premières théories structurelles en sociologie sont définies par (Simmel 1903) (Farganis 1993), fondées sur des structures des triades et le développement de l'individualisme: sous-groupes de 3 personnes A, B, et C, tel que A, a une relation directe

avec C et une autre indirecte avec C via B. B peut agir en influençant la relation entre A et C (Nettleton 2013). Dans les années 1930, Jacob Levy Moreno est le premier qui a développé une représentation graphique « sociogramme » (Nettleton 2013) (Erétéo 2011). C'est un diagramme présentant les personnes par des points et les relations sociales par des lignes entre eux (Nettleton 2013) (Erétéo 2011), introduisant l'aspect de la toile d'araignée (web) (Erétéo 2011). La visualisation d'une structure sociale en réseau a facilité sa compréhension et son analyse, en identifiant certaines caractéristiques locale, globales et des participants influents (Nettleton 2013) (Erétéo 2011) (Wasserman & Faust 1994). La personne ayant le grand nombre de connections a été par exemple désigné par le concept étoile «Star » (Erétéo 2011). Le sociogramme désigné par le terme du réseau social provenant de (Barnes 1954), était une première étape et le contexte idéal pour développer les théories mathématiques de l'analyse des réseaux sociaux.

Des formalisations mathématiques ont été développées pendant les années 1950, fondées sur les concepts de la théorie des graphes par des mathématiciens comme Harry, Norman, etc., (Erétéo 2011) (Nettleton 2013). L'aspect sociologique du monde réel a été donc saisi par l'aspect formel mathématique. La seconde moitié du XXe siècle a connu les premières applications des approches de la théorie des graphes pour analyser des graphes sociaux (Scott 2000) (Erétéo 2011). Par conséquent, les structures de graphes sont devenus le modèle formel principal pour les réseaux sociaux dans les sciences comportementales et sociales modernes, en informatique et l'économie ainsi que pour l'analyse complexe des réseaux en général (Erétéo 2011) (Cuvelier & Aufaure 2011) (Nettleton 2013). Le modèle mathématique adopté permet jusqu'à présent de formaliser de mieux en mieux l'analyse des réseaux sociaux en proposant des techniques et algorithmes de calcule et de détection des patterns et caractérisant des structures sociales.

2.3. Le modèle de représentation de base

2.3.1. Les premières structures sociales capturées et représentées

La recherche des modèles pour représenter des interactions sociales au sein d'une population a un large éventail d'applications (Berger-Wolf & Saia 2006) non seulement en sociologie et informatique: Dans la modélisation des maladies et leur propagation (l'épidémiologie), la transmission d'informations culturelles l'intelligence et la surveillance, la gestion des affaires, la biologie de conservation et l'écologie du comportement (Berger-Wolf & Saia 2006). D'abord, les bases sur lesquelles on s'appuie pour détecter des liens varient selon la nature des acteurs et le contexte social. Les relations d'amitié ont fait l'objet des premiers réseaux sociaux manipulés et ont été détectés à travers des questionnaires. On demande par exemple à des personnes d'une population de citer leurs amis. Le célèbre réseau d'amis d'un club de karaté aux États-Unis de l'anthropologiste Zachary (Zachary 1977) (Parthasarathy et al 2011) en 1977 était une très bonne illustration au milieu universitaire (Zachary 1977). Le réseau est composé de 34 membres observés durant 3 ans et fait l'objet de plusieurs études (Parthasarathy et al 2011). Le réseau est bien connu notamment par une fission qui s'est produite suite à l'apparition des conflits internes (opinions différentes entre 2 membres) dans le club, ce qui a engendré des ruptures dans la cohésion sociale (Zachary 1977) (Parthasarathy et al 2011).

Cependant, la socialisation de l'homme n'est pas limitée seulement à des relations explicites déclarée (Erétéo 2011), elle dépend également de plusieurs facteurs permettant de créer et maintenir ses liens. Des réseaux sont captés, formés par des relations implicites qui ont été fondamentalement identifiées entre des personnes similaires. En d'autres termes, les premières hypothèses sur des structures sociales ont été influencées par l'idée que les personnes tendent à s'associer avec d'autres similaires à eux. La similarité a été basée sur des facteurs généralement démographiques (Race, ethnicité, sexe, âge, religion, etc.) (Nettleton 2013), et qui produisent souvent des regroupements sociaux puissants. Des facteurs secondaires comme l'occupation, le comportement ou des valeurs intra-personnelles, etc., peuvent être introduits.

La façon la plus commune pour extraire des informations sur ces interactions et entités sociales est un réseau modélisé par un graphe (Berger-Wolf & Saia 2006). Maintenant il est intuitif de dire qu'un graphe est la représentation la plus classique d'un réseau social.

2.3.2. Bases formelles de représentation

Dans un graphe ou encore un graphe social qui modélise un réseau social, les nœuds représentent les acteurs et les bords représentent des relations. Cette structure de base permet de déduire sans doute que la représentation ainsi que l'analyse des réseaux sociaux sont fondées sur la théorie des graphes. Les concepts et notions de bases nécessaires pour formuler un réseau social (SN) seront détaillés:

- **Définition** 1. Nœud : Un nœud ou un sommet est l'unité de base qui représente un acteur, un agent ou même une ressource dans un SN.
- **Définition** 2. Lien : Un lien, une arête ou un arc est une connexion entre une paire de nœuds, et représente souvent une interaction sociale qui a eu lieu dans un point quelconque dans le temps entre les 2 individus concernés (Cuvelier & Aufaure 2011) (Berger-Wolf & Saia 2006).

La distinction entre les notions d'arc ou arrête devient importante pour modéliser les types de relations symétrique et asymétrique selon le taux de richesse informationnelle d'une relation, que porte par le lien correspondant. Une arrête est bien adaptée pour décrire une relation symétrique. Elle devient orientée d'un nœud source vers un nœud cible et c'est le cas d'un arc quand il s'agit d'une relation asymétrique.

- **Définition** 3. Graphe Social: Etant la représentation d'un réseau social, un graphe est considéré ici comme un type de données abstrait plus qu'une entité mathématique. Un graphe G (V, E) est défini par un ensemble de sommets V et un ensemble de liens E.
- **Définition** 4. Sous-graphe: Un sous-graphe G'(V,'E') de G est un graphe, tel que V' $\subset$ V, E' $\subset$ E, et $\forall$ {v, u} $\in$ E' $\rightarrow$ v, u $\in$ V. Un sous-graphe G(C) = (C, E(C)) dit induit peut être défini par un sous-ensemble C $\subset$ V où E(C) = {{v, u} $\in$ E | v, u $\in$ C}.
- **Définition** 5. Chemin: C'est une séquence ordonnée des nœuds liés, un après l'autre dans le graphe. Un chemin P est un sous-graphe P (V (P), E (P)), V(P) $\subset$ V et E(P) $\subset$ E, tel que V (P),= $\{v_{i0},.., v_{ik}\}$. E (P) = $\{\{v_{i0}, v_{i1}\},\ldots \{v_{ik-1}, v_{ik}\}\}$, k est la longueur du

chemin. S'il n y a aucun sommet répété, le chemin est appelé simple (Cuvelier & Aufaure 2011).

S'il existe au moins un chemin entre v et u, alors ils sont connectés.

- **Définition** 6. Géodésique: C'est le chemin détecté comme étant le plus court entre 2 nœuds données.
- **Définition** 7. Diamètre: La longueur de la géodésique la plus longue dans le réseau.
- **Définition** 8. Degré: le degré d'un nœud est le nombre des liens adjacents.
- **Définition** 9. Graphe connexe: C'est un graphe où il y a entre chaque paire de nœuds un chemin. $\forall v, u \in V, \exists P$ connectant v, u.

  Un sous-graphe connexe est une composante connectée (Cuvelier & Aufaure 2011).
- **Définition** 10. Graphe complet: C'est un graphe ayant un lien entre chaque paire de nœuds. Un graphe ayant $|V| = n$ $(n \in N)$, est dit complet si $|E| = n(n - 1)/2$.

  $\forall v, u \in V, \exists e \{v, u\} \in E$.

L'un des anciens exemples les plus connus des réseaux sociaux représentés du monde réel est celui de la famille « Medici » et ses liens avec d'autres familles dans le XVe siècle à Florence (Padgett 1994) (Breiger & Pattison 1986). Différentes types de relations de mariages, business, alliances, etc. ont été prise en comptes (Padgett 1994) (Nettleton 2013). ((Lusseau et al 2003)) représentent le SN de mariage entre 16 familles par un graphe simple non-orienté avec des relations symétriques.

Autres exemples simple peuvent être tirés depuis des études en science comportemental là où les graphes sociaux décrivent des relations basées souvent sur des proximités physiques (écologie comportemental) entre animaux ainsi que créatures marines. La figure suivante montre un réseau social des associations fréquentes (159 liens) entre 62 dauphins vivant hors de "Doubtful Sound" au sud de la Nouvelle-Zélande, étudié par (Lusseau et al 2003), (Parthasarathy et al 2011).

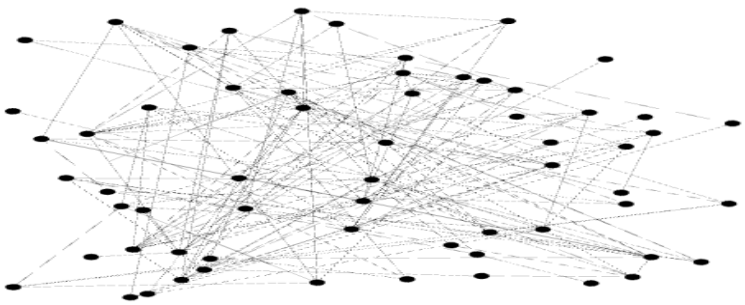


Figure 3. Graphe social des associations fréquentes entre des dauphins vivant au sud de la Nouvelle-Zélande

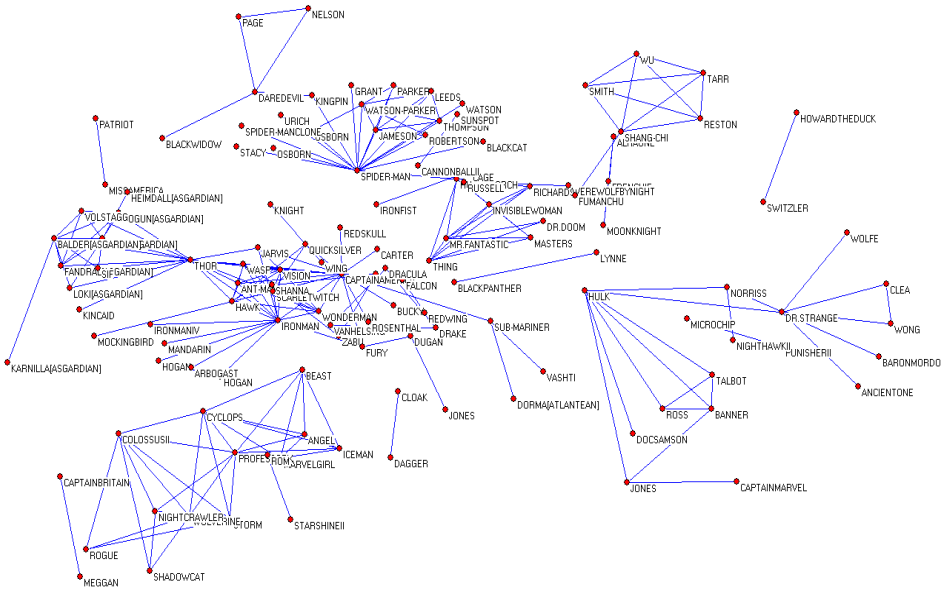


Figure 4. Le graphe social de l'univers de Marvel

Le troisième exemple présente un réseau social construit par Cesc Rosselló, Ricardo Alberich, et Joe Miro de l'université des îles Baléares dans (Alberich et al 2002) pendant leurs études sur l'univers Marvel comme un graphe social réel. Il s'agit d'un réseau de collaborations sociales entre des acteurs des bandes dessinées, sur lequel des études analytiques ont été réalisées afin de découvrir comment le réseau s'est développé comme un réseau réel et affiner les modèles de graphes sociaux qui ont été utilisés jusqu'à nos jours. Le SN des personnages dans le roman ' les Misérables' de Victor Hugo (Knuth 1993) est aussi l'un des anciens SNs populaires qui a été testé dans plusieurs expérimentations. Selon la co-apparence de 2 personnages donnés dans une scène (dans un chapitre), une relation (d'affiliation) est créée entre les 2 (Knuth 1993).

Au-delà des nœuds et des liens, plusieurs types de graphes se distinguent pour supporter des informations supplémentaires : Liens orientés (arcs), pondérés, étiquetés, etc., à considérer dans un SN. Par exemple des propriétés supplémentaires et fonctions peuvent être ajoutées et permettent de définir le type du graphe social.

- **Définition** 11. Graphe orienté: C'est un graphe avec un ensemble de liens orientés: arcs.

Un arc de u vers v se décrit par une paire ordonnée ou encore le couple (u, v). L'ordre déterminé des extrémités reflète une information significative dans l'interaction sociale entre u et v (symétrique ou asymétrique). Par conséquent, Le cas orienté va imposer plus de restrictions sur beaucoup de concepts manipulés dans la représentation sociale. Des notions comme le chemin seront plus pointues. Un chemin orienté est un chemin ou une séquence d'arcs depuis un nœud source (de départ) vers un nœud d'arrivée. L'un des premiers graphes sociaux orientés à citer est celui formé à travers des interactions émergeantes par des lettres envoyées dans l'expérience de « Milgram » (Miller 1986) entre 1960-1963.

- **Définition** 12. Graphe acyclique: C'est un graphe qui ne possède pas un circuit ou un cycle et représente une hiérarchie ou une forêt respectivement dans le cas orienté ou non-orienté.
- **Définition** 13. Lien pondéré: Un lien pondéré est un lien associé à une valeur appelé un poids ou encore affinité qui décrit l'intensité, l'importance, force, coût, fréquence, etc., d'une relation correspondante.
- **Définition** 14. Graphe pondéré: C'est un graphe avec un ensemble de liens pondérés. Il est défini par G (V, E, W), dans le cas général $W : E \rightarrow \mathbb{R}$, est une fonction de pondération des liens.

Par exemple, le réseau des équipes universitaires de football américain « US College Football » (Girvan & Newman 2002) est une illustration des premiers graphes sociaux pondérés où les nœuds sont les équipes et les poids sur les liens représentent le nombre de matchs joués entre 2 équipes (Nettleton 2013) au cours d'une saison régulière (Automne 2000). Le graphe des co-apparences des personnages du roman ' les Misérables' ((Knuth 1993)) est également un réseau pondéré tel que: La valeur de chaque lien représente le nombre de co-apparences entre une paire de personnages.

- **Définition** 15. Lien étiqueté: Un lien étiqueté est associé à un label (une étiquette). cette propriété permet souvent de distinguer le type ou la signification d'une relation.
- **Définition** 16. Graphe étiqueté: C'est un graphe avec un ensemble de liens étiquetés. Il est défini par G (V, E, L), dans le cas général $L : E \rightarrow lbs$, est une fonction d'étiquetage des liens à partir d'un ensemble d'étiquette « lbs » bien défini.
- **Définition** 17. Coupure (Cut): Une coupure c(C, V\C) est une partition de V en 2 ensembles C, V\C. Sa taille est le nombre des liens reliant les sommets de C avec les sommets de V\C (Cuvelier & Aufaure 2011).

$$c(C, V\backslash C) = |\{\{u, v\} \in E \mid u \in C, \quad v \in V\backslash C\}| \tag{1}$$

Plusieurs autres structures de graphes ont été abordées par (Newman 2003) pour supporter des réseaux de plus en plus complexes à travers un inventaire des définitions de topologies et de métriques. On trouve par exemple la notion de l'hyperlien qui est défini comme étant un lien qui connecte plus de 2 nœuds. Un hypergraphe est donc défini par un graphe qui contient un ou plusieurs hyperliens (Newman 2003) (Berge 1985), etc. Si les relations d'un SN présentés par des liens jouent un rôle majeur pour décider le type du modèle de graphe, les éléments de la population en interaction font également l'objet de distinction. Un SN peut inclure des individus ainsi que différents types de ressources manipulées par les acteurs et qui sont le support d'interactions (**Réseaux hétérogènes**). Dans ce cas, différent type de nœuds se distinguent dans le graphe correspondant. Les notions liées aux graphes multipartites sont très adaptées pour modéliser tels SN.

Réseaux sociaux hétérogènes

La plupart des études analytiques sur les SNs, s'appliquent sur des modèles de graphe social homogène. Cependant, une structure sociale dans le monde réel peut être formée par des nœuds, des liens de types différents. Cette diversité constitue d'une part une opportunité pour des informations précieuses à extraire (Parthasarathy et al 2011). D'autre part, elle représente

un challenge, car la diversité des types de nœuds et des liens n'est pas évidente à supporter et analyser dans un graphe social. Dans ce contexte, la base de données en ligne IMDb (Internet Movie Database ou encore : Base de données cinématographiques d'Internet) est une bonne illustration. Un réseau peut être formé par des entités : films, réalisateurs, acteurs, etc., ainsi que des liens : acted-in, directed-by, co-acted-in, etc., (Parthasarathy et al 2011). Maintenant, les SNs hétérogènes émergent en ligne clairement depuis l'usage des applications spéciales sur le web (Erétéo 2011)

- **Définition** 18. Graphe multipartite: C'est un graphe dont l'ensemble des nœuds est décomposé en k parties. Chaque partie contient un type unique de nœuds et les liens connectent les nœuds de différentes parties. Si k = 2 le graphe est appelé bipartie, k = 3 le graphe est triparti, etc.

Un graphe biparti est utilisé par exemple pour modéliser un réseau d'affiliation entre les acteurs comme un premier type de nœuds et les objets d'affiliation comme le deuxième type de nœuds. Ce type de réseau s'appelle également 'Two-mode Networks' ou en mode 'Two-mode ties' ((Latapy et al 2008)). C'est l'image par exemple d'un réseau d'affiliation d'un ensemble d'administrateurs à un ensemble d'entreprises où les liens associent par exemple les entreprises à leurs directeurs. Une diversité de ressources peut être impliquée (auteur, papier, évènement, etc.) (Kang et al 2007) dans des SN plus compliqués (relations de collaborations : co-auteur)) faisant appel à des notions comme l'hyperlien et hypergraphe. L'analyse de tels modèles de réseaux nécessite souvent des transformations vers des structures plus simples. Par exemple, les graphes sociaux bipartis 'Two-mode Networks' peuvent être transformés (des projections) en 2 réseaux simples pondérés en mode 'Weighted one-mode networks'.

Les notions définies ci-dessus sont des concepts nécessaires qui montrent le fondement théorique pour formuler un SN basé sur la théorie des graphes. Toutefois, les détails ne seront utiles que selon les besoins car il ne faut pas oublier le contexte le sociologique et être juste théoriquement limité

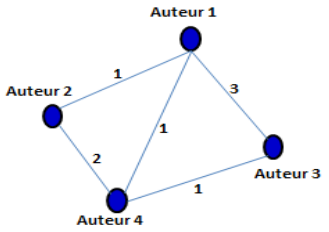
2.3.3. Représentation en mémoire informatique

La modèle de représentation des SN soulève évidement une question clé sur comment représenter un graphe en général en mémoire informatique devant les coûts de calcules potentiellement élevés de plusieurs opérations de haut niveau en SNA. Les structures des données les plus populaires sont des listes ou des matrices en adjacence ou incidence mais il y a d'autres approches ensemblistes moins connues dans (Scott 2000).

- **Matrice ou liste d'adjacence**

La matrice est l'objet mathématique le plus utilisé pour saisir et manipuler les composantes d'un graphe, notamment un graphe social. La matrice d'adjacence A d'un graphe G (V, E) : $|V| = n,\ |E| = m$ est un tableau à deux dimensions $n \times n$ de valeurs binaires $a_{ij} = 1$ ou des poids (dans le cas d'un graphe pondéré) si et seulement, $\{v_i, v_j\} \in E$ sinon $a_{ij} = 0.\ i, j = 1..n$. Dans le cas non-orienté $a_{ij} = a_{ji}$ ce qui donne une matrice

symétrique. La figure suivante montre un exemple de matrice représentant un SN non-orienté, formé par des interactions de collaborations (relations symétriques) entre des auteurs sur des papiers scientifiques.



	Auteur 1	Auteur 2	Auteur 3	Auteur 4
Auteur 1	0	1	3	1
Auteur 2	1	0	0	2
Auteur 3	3	0	0	1
Auteur 4	1	2	1	0

Figure 5. Matrice d'adjacence d'un réseau social de collaborations scientifiques entre auteurs.

Dans le cas d'un SN orienté, un arc (v_i, v_j) se représente par la cellule aij tel que la ligne i et colonne j correspondantes représentent respectivement la source et la cible de cet arc. Pour (v_j, v_i) il est possible d'insérer la valeur négative tel que : $a_{ji} = -a_{ij}$

Dans le cas d'une liste d'adjacence, le graphe est représenté par un tableau T à une dimension de taille n où T[i] contient la tête de la liste des successeurs (nœuds adjacents – cas non orienté) d'un nœud v_i. En général, les successeurs peuvent apparaitre dans n'importe quel ordre. Si v_i est isolé la liste T[i] est vide. Les pointeurs sont souvent le moyen le plus efficace pour se structurer sous forme des listes. En résumé, le graphe apparait comme un tableau de pointeurs sur des listes.

- **Matrice ou liste d'incidence**

Les lignes et les colonnes dans une matrice d'adjacence représentent le même ensemble de nœuds. Cependant, la matrice d'incidence semble être la plus adaptée pour représenter un graphe social formé par 2 types de nœuds G (V, U, E). C'est un tableau : $n \times m$ tel que la cellule aij représente une relation entre $v_i \; et \; u_j, i = 1..n, j = 1..m$ suivant un principe identique en cas orienté ou de pondération. La figure suivante montre qu'une matrice d'incidence est convertible en deux matrices d'adjacence (2 SN) chacune avec un seul type de nœuds : $n \times n$ et $m \times m$. La valeur de chaque cellule dans une matrice d'adjacence produite, est obtenue à travers le nombre de connexions partagées entre 2 nœuds du même type dans la matrice source d'incidence.

	Chercheur 1	Chercheur 2	Chercheur 3	Chercheur 4
Conférence 1	0	1	1	1
Conférence 2	1	0	0	1
Conférence 3	1	0	0	1

Matrice d'incidence d'un réseau social des chercheurs participant à des conférences scientifiques

	Conférence 1	Conférence 2	Conférence 3
Conférence 1	0	1	1
Conférence 2	1	0	2
Conférence 3	1	2	0

Matrice d'adjacence des conférences : Les poids représentent le nombre des chercheurs participant partagés

	Chercheur 1	Chercheur 2	Chercheur 3	Chercheur 4
Chercheur 1	0	0	0	2
Chercheur 2	0	0	1	1
Chercheur 3	0	1	0	1
Chercheur 4	2	1	1	0

Matrice d'adjacence des chercheurs: Les poids représentent le nombre des conférences partagés

Figure 6. Matrice d'incidence convertible en 2 matrices d'adjacence

Une liste d'incidence apparait également comme un tableau T de taille n où chaque T[i] pointe sur une liste des arrêtes (arcs) incidentes à v_i

On peut distinguer 2 éléments essentiels influant sur le choix de la structure la plus adaptée en pour économiser préalablement le coût de calcul en espace mémoire (l'occupation mémoire O()): la connectivité du graphe social (les liens) et les types de ressources (les nœuds).

Tableau 1. Les variantes de structurations d'un graphe social en mémoire

	Adjacence	Incidence
Matrice	- Dense connectivité - Nœuds de type unique ➔ $O(n^2)$	- Dense connectivité - 2 types de nœuds ➔ $O(n \times m)$
Liste	- Graphe clairsemé - Nœuds de type unique ➔ $O(n+m)$	- Graphe clairsemé - 2 types de nœuds ➔ $O(n+m)$

Généralement, les listes d'adjacence sont préférées lorsque le graphe est clairsemé en termes de connectivité, tandis qu'une matrice d'adjacence sera préférée si le graphe est dense.

- **Matrice Laplacienne (de Laplace)**

La matrice de Laplace '$\mathcal{L}$' ou de Kirchhoff est le type de représentation matricielle d'un graphe, qui est définie à travers le degré des nœuds. Ses valeurs sont affectées comme suivant ((Erétéo 2011)):

$$a_{ij} = \begin{cases} degré\ de\ i, & si\ i = j \\ -1\ si\ i \neq j\ et\ (v_i, v_j) \in E \\ 0\ autrement \end{cases}$$

2.3.4. Caractéristiques particulières du modèle de graphe social

Les graphes représentant des réseaux sociaux réels sont-ils aléatoires ? C'était une question de recherche en SN (McGlohon & Faloutsos 2008). Les réponses ont déjà démontré qu'un graphe social sur plan statique se distingue d'un graphe aléatoire ((Loiacono 2011)). Il est statistiquement caractérisé par des valeurs dérivées particulières extraites depuis les degrés des nœuds et la longueur des chemins (géodésiques), etc., ainsi qu'il présente des caractéristiques supplémentaires (Nettleton 2013) liées par exemple au diamètre des graphes, le nombre de triangles ou d'isomorphismes, Coefficient de clustering, etc. Des séries de propriétés de SNs ont été définies et montrées par plusieurs auteurs comme dans (Mislove et al 2007) mais les plus citées sont:

Loi de puissance dans la distribution des degrés

La distribution des degrés est le paramètre descriptif communément admis dans les recherches récentes de caractérisation et d'analyse des réseaux sociaux. Il s'agit des études de distributions statistiques de certaines valeurs (degré ou autre) pour déduire la distinction du modèle de SN en déterminant l'existence d'une structure invariante à l'échelle (Ducruet 2010). La distribution des degrés des nœuds d'un graphe modélisant un SN peut être visualisée par une courbe (bi logarithmique) résultant d'un croisement entre les degrés des sommets (les abscisses), et la fréquence des sommets (les ordonnées). La fréquence « d_k » désigne dans ce cas le nombre des nœuds ayant un même degré k. Il s'agit donc d'une distribution de fréquences visant à vérifier l'applicabilité de la loi de puissance qui permet en effet de décrire tous les phénomènes qui présentent une invariance à l'échelle. La loi de puissance ici définit la probabilité qu'un nœud aura un degré k est proportionnelle à $k^{-\gamma}$ pour un grand k et $\gamma > 1$. Dans la figure suivante, l'invariance à l'échelle pour les 2 exemples de graphe social se confirme visuellement au niveau de la pente de la courbe qui est généralement supérieur ou égale à 1 (Ducruet 2010), même si la littérature suggère des valeurs plus élevés (de 2 à 3) pour valider la structure en question. Le premier exemple montre une distribution dans le réseau social de l'univers Marvel (Figure 4). Dans le deuxième exemple illustre la distribution dans le réseau social du club de karaté de Zachary (Zachary 1977).

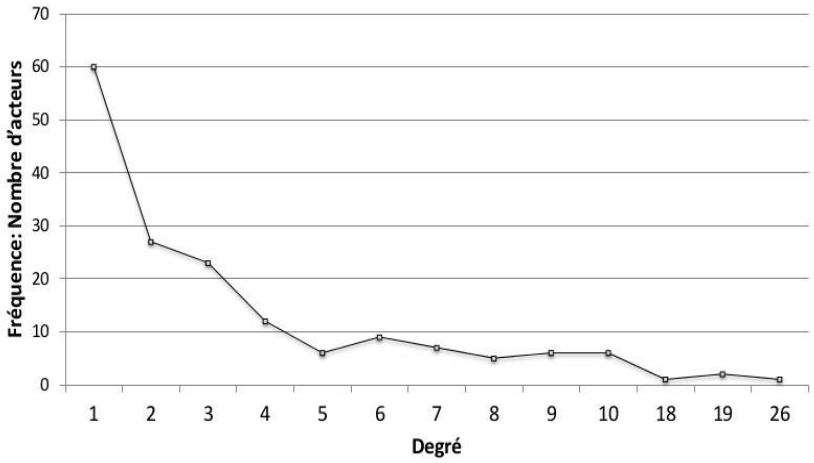


Figure 7. Distribution de degrés selon la loi de puissance (le graphe social de l'univers de Marvel)

Les 2 cas montrent que la distribution des degrés suit la loi de puissance classique. En d'autre terme, plus le degré « k » est grand, plus le nombre des acteurs ayant ce degré (la fréquence dk), diminue. C'est-à-dire qu'il y a peu de nœuds qui ont les degrés les plus élevés. Cette configuration est un paramètre descriptif souvent lié à la notion de '**l'attachement préférentiel**' qui veut dire que les liens se créent dans le graphe social avec des nœuds ayant un degré important. Cependant, elle n'est pas forcément validée dans le cas des graphes planaires modélisant les réseaux où la dimension hiérarchique est modérée (Ducruet 2010).

Phénomène du petit monde

Le phénomène du petit monde est une propriété globale dans un graphe social, qui explique le faite que même si le graphe est large, sa structure globale est déterminée par sa structure au niveau local (Nettleton 2013). Certains chercheurs affirment qu'il existe une certaine analogie entre le comportement local et la structure globale des SNs réels comme dans (Robins et al 2005) de telle sorte que le réseau est globalement dominé par des processus sociaux locaux. L'exemple donné dans ((Lusseau et al 2003)) montre que la famille «Medici» était un acteur clé dans le SN général de Florence au XVe siècle, car ses membres étaient au centre des structures des mariages et des alliances commerciales (Une structure semblable à une étoile « Star »). Ils avaient des relations simples plus rapides avec d'autres acteurs clés par rapport à leurs rivaux politiques et autres acteurs (Nettleton 2013).

Structure q-Star: La structure q-Star est un sous-graphe de (q + 1) nœuds dans laquelle un nœud « central » qui est connecté exactement à q nœuds.

En termes de relations humaines, il suffit donc d'un nombre réduit de connections (au moyenne, 6) pour permettre à 2 personnes fortement différenciées socio-économiquement et géographiquement d'y accéder l'un à l'autre dans un grand réseau. Les auteurs citent également 4 conditions clés pour qu'un petit monde se manifeste dans le réseau (Nettleton 2013):

- Les individus cherchent plus d'un seul partenaire dans le réseau.
- Devant le coût élevé pour maintenir plusieurs partenaires, il y a une tendance (cognitive, sociologique, etc.) qui limite une multitude de partenaires. Par conséquent, le diamètre dans les SNs larges sera également limité. La limitation à 6 a été tirée de la probabilité qu'un nœud donné tente de contacter un autre à une distance 6, est très faible (McGlohon & Faloutsos 2008) (Kleinberg 2000).
- Il existe une certaine tendance dans les partenaires du réseau de s'étendre sur d'autres partenaires, ce qui amène à un équilibre structural, l'apparition des triangles et 'clustering'.
- Si la troisième condition est excessivement appliquée, des structures de cliques vont se produire: Les liens seront suffisamment nombreux entre les nœuds pour rendre les chemins de plus en plus courts.

Triangles: L'existence des structures en triangles dans le graphe social incarne l'aspect de **transitivité** dans le SN, en se basant sur le postulat suivant : L'ami de mon ami est mon ami (Malek 2009). En général, Si un nœud A est connecté à un nœud B et que ce dernier est connecté à un autre nœud C, alors A et C ont une forte probabilité d'être également connectés : Deux nœuds liés à un même nœud ont une forte probabilité d'être liés entre eux.

A cause de ces structures locales répétées, l'effet du petit monde est l'une des caractéristiques essentielles qui différencie les graphes sociaux des autres graphes d'une façon générale. En outre, il reflète des propriétés sur les géodésiques, diamètre (réduit), les triangles, q-Stars, et la présence d'une grande tendance au « clustering » et donc des structures modulaires se construisent. Ainsi, un modèle de SN affiche souvent une **structure en communautés**.

Beaucoup de propriétés structurelles des réseaux complexes se présentent dans les SNs (Erétéo 2011) (Newman 2003). La distribution de degré selon la loi de puissance, l'effet de petit monde, la tendance au « clustering », etc., sont des caractéristiques typiques qui sont validées à travers des techniques et indicateurs de SNA. Dans ce sens, les auteurs affirment que les défis dans la modélisation des graphes réels comme les SNs, se répliquent autour des distributions qui vérifient la loi de puissance, mais aussi autres structures moins connues comme « Bowtie » ou « Jelly Fish » (Nettleton 2013) tout en maintenant le petit diamètre (l'effet du petit monde) du graphe. Cela n'a pas empêché certains auteurs d'essayer de rapprocher un modèle de SN réel à un graphe aléatoire en simulant à travers des formules probabilistes proposées, la création des liens dans les matrices d'adjacences correspondantes. Cependant, l'hypothèse d'avoir des liens indépendants dans les SNs est généralement invraisemblable (Nettleton 2013). En conséquence, certains chercheurs ont fait à appelle à des modèles paramétrés de dépendances Markoviennes pour supporter la dépendance entre les relations sociales et les paramètres, proportionnels à la fréquence de 4 structures: Simple lien, 2-Star, 3-Stars et triangles (Nettleton 2013). Donc, la compréhension de ces propriétés ne montre pas seulement que les graphes modélisant des SNs ne sont pas aléatoires mais permet également de définir des repères pour construire des modèles générateurs de graphes sociaux ainsi que pour les processus d'échantillonnage.

3. Analyse classique des réseaux sociaux (SNA/ SNAM)

3.1. Historique

SNA a près de 70 ans d'histoire. Une histoire divisée en 3 grandes périodes (Degenne & Forse 1994): La première période a connu les fondements de l'approche, les différents édifices construits entre les années 1940 et les années 1960 basés sur les textes classiques de certains auteurs de la fin du XIXème et du début du XXème siècle, comme ceux de Bouglé et de Simmel (Degenne & Forse 1994) (Nettleton 2013). La deuxième période est caractérisée par l'articulation de la méthode entre les années 1960 et 1970. Des recherches méthodologiques se sont développées, destinées à assurer la mise en œuvre rigoureuse (Degenne & Forse 1994). La dernière période réfère au développement actuel, depuis les années 1980 jusqu' à aujourd'hui là où les recherches ont été amendées et perfectionnées. En plus, cette période est caractérisée actuellement par des nouvelles pistes qui se sont ouvertes (Degenne & Forse 1994).

3.2. Le domaine de recherche et objectifs

La représentation et l'analyse des interactions sociales sont considérées comme un domaine de recherche important qui attire les chercheurs de nombreux domaines (Kazienko et al 2011). L'analyse des réseaux sociaux (SNA) apparait comme le domaine d'étude et de

recherche dans les SNs, qui vise à fournir généralement des réponses à la socialisation de l'homme. C'est un domaine interdisciplinaire là où se chevauche la théorie des graphes, les statistiques la psychologie sociale, la sociologie, etc. La recherche dans un SN signifie la recherche dans un modèle de graphe particulier: Une structure analysable. Par conséquent, la SNA est systématiquement fondée sur le « Graph Mining », une spécialisation de « Data Mining » qui a comme objectif global de traiter des données et extraire des connaissances de valeur, difficiles à interpréter significativement par l'être humain. Dans ce sens, SNA ou bien « SNA and Mining » (SNAM) vise à comprendre la structure du SN, le comportement de ses entités sociales, l'évolution, etc., et exploiter ses caractéristiques clés à fin de gérer leur cycle de vie et prédire leur évolution (Erétéo et al 2009). En outre, c'est l'ensemble des méthodologies: Des techniques, approches, algorithmes, métriques, etc., moulés dans un cadre théorique formel, proposé pour modéliser et analyser respectivement des SNs. Il est vrai que cet aspect théorique permet de concevoir la structure sociale sous forme de modèles mathématiques. Toutefois, avec l'aspect méthodologique, ces modèles sont traités beaucoup plus comme un type de données abstrait analysable en premier plan à travers des mesures de calcule sociométriques inspirés de la sociométrie. C'est une discipline qui a aussi contribué à l'essor de SNA par son apport empirique dû en partie à l'œuvre de Moreno, l'un des précurseurs de SNA et de la psychologie sociale (Degenne & Forse 1994).

Beaucoup de recherche en SNA ont mené à la définition de plusieurs indicateurs (métriques) et algorithmes permettant de caractériser le réseau localement ou globalement. Ils sont en général décomposés en 2 grandes catégories selon le niveau de traitement dans le réseau. Certains fournissent des informations au niveau local, sur le positionnement individuel des acteurs et leur impact sur la communication, leur prestige, etc. D'autres donnent des informations sur la structure globale du SN ainsi que sur sa structure modulaire (en groupes).

3.3. Métriques, indicateurs et algorithmes

3.3.1. Traitements d'analyse locale

L'étude de l'influence des acteurs dans les réseaux humains est une question de recherche importante en SNA (Tang et al 2010b). Le traitement d'analyse locale se fait à travers des métriques qui s'attachent à caractériser (quantifier) la situation/ l'influence d'un élément: un acteur ou encore un lien, dans le réseau par rapport aux autres éléments. Les mesure locales sont bien plus nombreuses et diverses que les mesures globales et peuvent être classées en deux grands types (Ducruet 2010). Il y a des mesures locales de voisinage ou d'ensemble. Les mesures locales en voisinages décrivent la situation d'un élément, généralement un acteur, par rapport à ses voisins immédiats, directement connectés ou adjacents. Tandis que les mesures d'ensemble prennent en compte la situation d'un élément par rapport à tous les autres éléments de même nature présents dans le réseau (Ducruet 2010). À cet égard, des sous-catégories et des extensions sont développés pour avoir des indices plus raffinés, basées souvent sur le concept de « centralité ».

3.3.1.1. Mesures de centralité: Positions stratégiques individuelles

Les mesures locales peuvent être interprétées, selon l'approche et la thématique en jeu comme des mesures de centralités et d'accessibilités. Intuitivement, une personne qui a

beaucoup de contacts et communiquent bien avec les autres par exemple dans un groupe de collaborateurs, est considérée plus importante qu'une autre personne ayant moins de contacts. L'exemple suivant montre un réseau en étoile (9-Star) où l'acteur (1) est plus central car il communique avec la majorité.

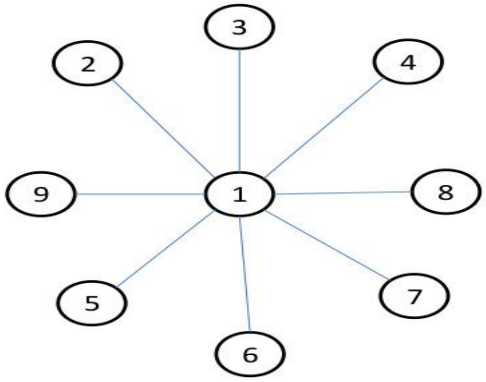


Figure 8. Un SN en étoile centré sur un acteur

Par conséquent, la problématique de la centralité est de justifier pourquoi un acteur est plus central qu'un autre. Elle définit le potentiel de l'acteur (Ducruet 2010), en identifiant les acteurs importants, influant et les positions stratégiques dans un SN de n nœuds. Différents critères peuvent être considérés pour définir une diversité de centralités individuelles et leurs normalisations: la portée de la mesure (au voisinage ou sur l'ensemble du SN), le cas orienté, la pondération, etc. Mais elles sont en fait basées sur 3 principales définitions (Freeman 1979) (Tommasini & Daolio 2010) (Erétéo et al 2008): centralité de degré, d'intermédiarité et de proximité.

3.3.1.1.1. Centralités de voisinage

Centralité de degré Cd (Degree Centrality)

Avec la centralité de degré, les acteurs centraux sont les nœuds ayant les degrés les plus élevés dans le graphe social. En effet, ces nœuds suscitent un grand intérêt, sont très visibles, plus actifs (Malek 2009) et ont un potentiel élevé pour faire circuler l'information, à travers cette forte connectivité. Elle peut répondre à des questions comme: quel est l'acteur le plus actif dans le réseau? Le plus central et celui qui a plus de liens ? (Tommasini & Daolio 2010). La centralité de degré d'un nœud $vi, i = 1..n$, n'est que le réflexe de son degré (di) qui est le nombre des liens (arrêtes/ arcs) adjacents.

$$\mathrm{Cd(i)} = \mathrm{d(i)} = \sum_j \{\mathrm{vi, vj}\} \qquad (2)$$

Remarque (Normalisation): La centralité de degré d'un acteur a l'avantage d'être normalisée selon deux façons possibles. La normalisation peut être le résultat d'une division par la centralisé maximale théorique de degré qui se présente dans un nœud central d'un réseau en étoile composé depuis les n nœuds du SN étudié ((n-1)-Star):

$$\mathrm{Cd(i)} = \frac{\mathrm{d(i)}}{\mathrm{n-1}} \qquad (3)$$

Comme il est possible de normaliser aussi en devisant par la centralité maximale de degré : L'acteur ayant la plus grande centralisé degré.

$$\mathrm{Cd(i)} = \frac{\mathrm{d(i)}}{\max_{\mathrm{j=1..n}} \mathrm{d(j)}} \qquad (\mathbf{4})$$

Évidemment, le cas général de calcule de cette centralité n'est pas différent dans un graphe orienté. Cependant, la distinction entre les liens entrants et sortants respectivement vers et depuis un nœud (vi) est très importante pour raffiner la métrique sur ces relations asymétriques du SN. Donc des versions comme « In_Degree » : degré entrant et « Out_Degree »: degré sortant, sont utiles pour caractériser le support et l'influence. Dans le cas d'un graphe pondéré, l'intensité de la relation (son poids) peut être prise en compte dans le calcul pour caractériser l'activité d'un acteur avec plus de fidélité. La centralité de degré pondéré « Weighted Degree » est définie comme étant la somme des poids des liens adjacents.

$$\mathrm{Cdw(i)} = \sum_j \mathrm{w(vi, vj)} \qquad (\mathbf{5})$$

Il existe autres définitions alternatives comme le « n-Degree » proposé par (Scott 2000). C'est une version de degré étendue sur des distances variables : sur des chemins. On note que la distance entre 2 acteurs est le nombre minimum des relations (distance géodésique) qui les relie. Par exemple, 2-Degree consiste à considérer tous les voisins à une distance inférieure ou égale à distance 2. En d'autre terme, c'est le nombre des voisins qui ne sont pas directement adjacents, mais qui se trouvent sur des chemins adjacents d'une distance égale ou parfois inférieure ou égale à n. Cependant et selon (Burt 1992), le n-Degree est rarement utilisé avec n supérieur à 2 qui est l'équivalent à la limite de la portée d'un acteur qui lui permet d'observer et / ou influencer dans son réseau ((Erétéo 2011)).

La centralité de degré n'est pas seulement la mesure la plus représentative comme mesure locale de centralité mais elle pèse également dans les mesures d'ensemble (voir plus loin).

Indice de contribution

Dans le contexte de communication par email, (Gloor & Zhao 2004) proposent une mesure appelée indice de contribution (CI) d'un acteur dans son réseau. Elle mesure l'activité de l'individu étant l'émetteur et le récepteur d'un ensemble de messages échangés au sein d'une équipe (Gloor & Zhao 2004).

$$\mathrm{CI} = \frac{\mathrm{messages_{sent}} - messages_{received}}{\mathrm{messages_{sent}} + messages_{received}} \qquad (\mathbf{6})$$

C'est la fréquence suivant laquelle il envoie (Out_Degree) et reçoit (In_Degree) les messages

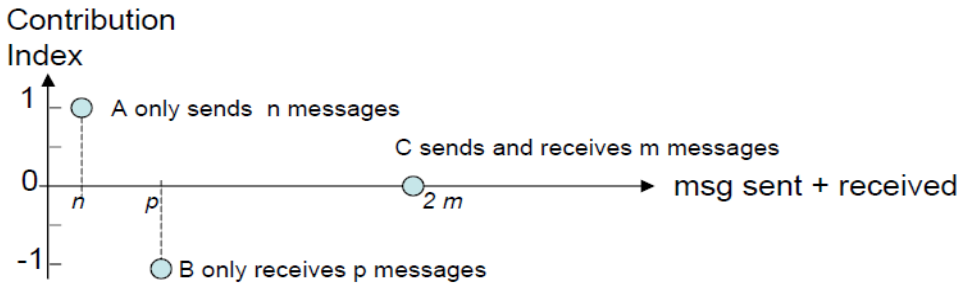


Figure 9. Valeurs de l'indice de contribution CI ((Gloor & Zhao 2004))

L'indice de contribution atteint la valeur 1 (le point A dans la Figure 9), si l'individu envoie n messages sans recevoir aucun message. Il touche le -1 (point B dans la Figure 9), si l'individu ne reçoit que p messages, sans envoyer aucun message. Il est égal à 0 (le point C dans la Figure 9), si l'individu a un comportement de communication totalement équilibré (Le même nombre de messages envoyés et reçus (Gloor & Zhao 2004).

3.3.1.1.2. Centralités d'ensemble

La plupart des mesures locales d'ensemble sont des mesures d'accessibilité, traitées selon différents angles mais basées essentiellement sur les notions des chemins, distance géodésique ou encore la distance euclidienne, etc. L'applicabilité de ces mesures dépasse le cadre des SNs, dans les réseaux de transport routier ou ferroviaire, visant à détecter des situations particulières des sommets (villes, gares, etc.). Mais le point commun se concentre sur le processus de localisation d'un nœud qui se réalise par rapport à tous les autres, à travers les chemins possibles, la longueur (totale, la plus courte, moyenne) et autres caractéristiques (poids, qualités, etc.).

3.3.1.1.2.1. Centralité de proximité Cc (Closeness Centrality)

Pour certains chercheurs la centralité en général a été considérée comme mesure de proximité dans le graphe social. La centralité de proximité mesure la capacité d'un nœud à se connecter et atteindre (cas orienté) rapidement les autres acteurs dans le réseau ((Erétéo 2011)), suivant la longueur des chemins sur lesquels il se trouve. La notion de proximité (Closeness) est liée à la distance. Dans ce cas, l'acteur central est celui qui se trouve sur des distances les plus courtes possibles (distance géodésique) pour interagir facilement avec les autres acteurs. Deux nœuds sont proches l'un à l'autre si leur distance géodésique est courte (Tang et al 2010b).

- **Définition** 19. Distance géodésique : C'est la plus courte distance d (i, j) de v_i vers v_j, définie par le nombre des liens sur le chemin le plus court (le chemin géodésique) entre v_i et v_j.

La centralité de proximité d'un nœud est inversement proportionnelle à la somme des distances géodésiques sur les chemins minimaux possibles (les plus courts) entre ce nœud et les autres (Tommasini & Daolio 2010).

$$Cc(i) = \frac{1}{\sum_j^n d(i,j)} \qquad (7)$$

La centralité de proximité suscite une importance statistique pour identifier des nœuds et régions clés. La réduction du coût de son calcul fait l'objet de certains travaux comme celui de: Eppstein et Wang (Nettleton 2013) qui ont proposé un algorithme d'approximation applicable sur des graphes du petit monde « Small world graphs », pondérés. L'algorithme estime la centralité de tous les nœuds selon une probabilité élevé et un coût en temps de O (m) (Nettleton 2013). Dans les graphes orientés, l'interprétation de la centralité de proximité est plus raffinée soit par les arcs sortants ou entrants pour représenter respectivement la capacité d'un acteur d'atteindre ou être atteint dans l'ensemble du réseau (voir plus loin). La figure suivante montre l'exemple où les acteurs: 1, 33, 34, sont les plus centraux en termes de degré et de proximité au même temps. Ils sont les plus proches aux autres nœuds à cause de leur degré supérieur.

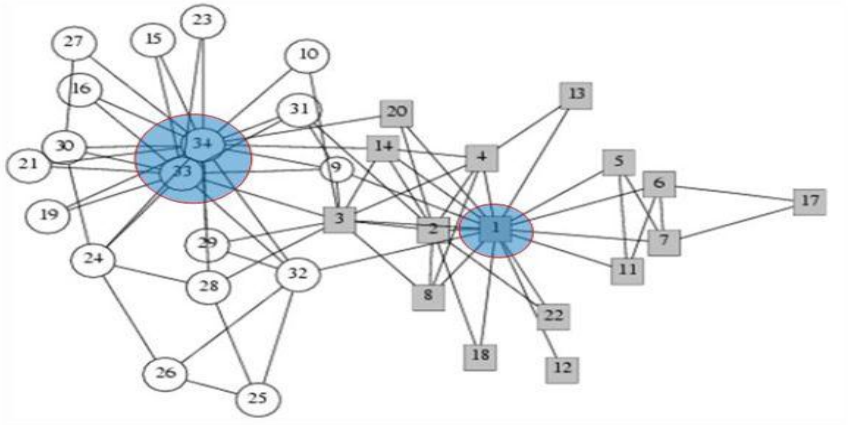


Figure 10. Les acteurs les plus centraux en termes de centralité de degré et de proximité dans le SN du club de karaté (Zachary 1977) ((Erétéo 2011))

Maximiser cette centralité de proximité c'est l'équivalent de minimiser la somme $\sum_j^n d(i,j)$, ce qui permet de répondre à la question qui cherche quel est l'acteur le moins loin des autres nœuds sur le SN, c.-à-d. le plus efficient dans la transmission de l'information et donc le plus central. Cependant, la minimisation de cette centralité permet de découvrir l'acteur le moins central. Ce qui permet d'avoir la tendance d'une mesure de « dé-centralité » qui s'appelle aussi « Indice de Shimbel » (Erétéo et al 2008). L'indice de Shimbel (Ai) d'un nœud v_i est l'inverse de la centralité de proximité. Il est connu également par la distance de Shimbel, l'accessibilité nodal, ou encore nodalité.

$$Ai = \sum_j^n d(i,j) \qquad (8)$$

Dans ce cas, les acteurs ayant des valeurs minimales sont les plus accessibles : Les plus centraux

Remarque (Normalisation): Remarquons que ces deux mesures de centralité permettent de répondre aux même questions : quel est l'acteur le plus central ou bien le plus indépendant ?, suivant 2 approches inversées. Cependant, la centralité de proximité est plus avantageuse vue que ses valeurs sont déjà normalisées entre 0 et 1. La valeur maximale 1 représente la plus forte accessibilité et centralité. De l'autre côté, si on veut appliquer systématiquement une

normalisation comme précédemment, la valeur de proximité sera plutôt divisée par la proximité minimale (l'inverse de la somme des distances maximales), ou encore par la proximité minimale théorique du nœud central dans un réseau en étoile (n-1)-Star. Cela amène également au principe de l'indice de Shimbel.

Cette liaison entre ces 2 indices de proximité et de Shimbel permet de dériver des variantes comme l'accessibilité **géographique**, notée A(G) obtenue depuis l'indice de Shimbel divisé par le nombre des nœuds présents sur les chemins géodésiques (Ducruet 2010). Cette variante peut être généralisée sur tous les n nœuds du graphe social G. Toutefois, l'accessibilité géographique de G reste une mesure locale et non pas globale, calculée comme suivant:

$$\mathrm{A(G)} = \frac{\sum_i^{\mathrm{n}} \mathrm{Ai}}{\mathrm{n}} \qquad (\mathbf{9})$$

La notion d'accessibilité a une applicabilité plus vaste non seulement en SNs, avec différentes variantes basées sur ces indices. C'est l'exemple de l'accessibilité potentielle dans le cas d'un SN composé par des nœuds ayant chacun un ensemble d'attributs (des caractéristiques) quantitatives. L'accessibilité potentielle d'un nœud est définie par une valeur d'attribut choisie (P) en fonction de la thématique traité (la population d'une ville, surface commerciale, la richesse, etc.) divisée par l'indice de Shimbel (Ducruet 2010).

$$\mathrm{Ai(P)} = \frac{\mathrm{Pi}}{\mathrm{Ai}} \qquad (\mathbf{10})$$

Noter bien que la centralité de proximité d'un acteur donné, basée sur le calcul des distances et chemins géodésiques sera beaucoup plus significative dans les graphes connexes. Ainsi, certains auteurs exigent d'avoir ce type de graphe pour autoriser le calcule de cette métrique. Cependant, les graphes sociaux modélisant des SNs sont rarement connectés. La centralité de proximité est par conséquent indépendante de la nature du SN modélisé.

3.3.1.1.2.2. Centralité d'intermédiarité Cb (Betweenness Centrality)

La centralité d'intermédiarité caractérise la capacité d'un nœud de servir d'intermédiaire dans un graphe social. Elle considère le nœud comme central tant qu'il se positionne sur beaucoup de chemins géodésiques entre les autres nœuds ((Erétéo 2011)). Dans ce cas le nœud central qui se trouve sur un chemin géodésique entre 2 parties du réseau, possède une position stratégique dans la cohésion, la circulation de l'information dans l'ensemble du réseau. Plus un acteur est intermédiaire plus il possède un fort contrôle sur la communication, plus le réseau dépend de lui, notamment lorsqu'il se trouve sur des chemins géodésiques uniques entre des parties (des groupes) éloignée ((Erétéo 2011)). Donc il a plus de pouvoir comme intermédiaire ou courtier ((Erétéo 2011)) qui peut choisir de lever ou abaisser le taux de communication entre les groupes et avoir un accès privilégié à l'information de chaque groupe (Burt 2004).

La centralité d'intermédiarité d'un nœud v_i est tout d'abord basée sur le concept d'intermédiarité ou encore l'intermédiarité partielle ((Erétéo 2011)), qui concerne son positionnement entre une paire de nœuds données v_j, v_k, non adjacents. Si vi, se localise sur une géodésique entre v_j et v_k, alors v_i aura un certain contrôle sur leur interaction mesurée par l'intermédiarité bjk(i):

$$\text{bjk(i)} = \frac{\text{gjk(i)}}{\text{gjk}} \qquad (\mathbf{11})$$

Tel que:

- gjk(i) : représente le nombre des chemins géodésiques possibles entre v_j et v_k, qui passe par v_i. Autrement dit, c'est le nombre d'occurrences de v_i sur ces chemins (Tang et al 2010b).
- gjk : représente le nombre total des chemins géodésiques entre les acteurs v_j et v_k, qui n'incluent pas v_i.

Noter bien qu'il peut exister plusieurs géodésiques entre 2 nœuds, en supposant que tous ont la même probabilité d'être utilisés. Si v_i est positionné comme un intermédiaire sur les chemins de plusieurs interactions, alors il sera qualifié par sa centralité d'intermédiarité comme un acteur central important (Malek 2009). En effet, la centralité d'intermédiarité Cb(i) d'un acteur v_i dans le SN est définie par le nombre gjk(i) des géodésiques possibles entre toute paire de nœuds v_j et v_k passant par v_i, normalisé par le nombre total des géodésiques gjk, entre v_j et v_k qui n'incluent pas v_i (Malek 2009), tel que $i \neq j,\ i \neq k$. En d'autres termes, c'est la somme des intermédiarités v_i de mesurées entre chaque paire de nœuds.

$$\text{Cb(i)} = \sum_j^n \sum_k^n \text{bjk(i)} \qquad (\mathbf{12})$$

Dans le cas où v_i ne figure sur aucun chemin géodésique, sa centralité d'intermédiarité atteint son minimum '0' (Malek 2009). Son maximum atteint (n-1) (n-2) / 2 lorsque toutes les géodésiques du réseau passent par v_i (graphe non orienté) ou (n-1) (n-2) avec des liens asymétriques (Graphe orienté). Dans ce cas, l'interprétation de la mesure est la même en conservant la même orientation dans un chemin ((Erétéo 2011)). Théoriquement cette valeur maximale se présente dans le nœud central d'un réseau (n-1)-Star. Dans l'exemple de la (Figure 8), l'acteur (1) apparait intuitivement comme le nœud central. Mais sa centralité d'intermédiarité a une justification plus théorique et formelle qui montre qu'il se trouve sur les 16 plus courts chemins entre les 6 autres acteurs, tel que Cb(1) = 15, tandis que Cb(2) = Cb(3) = Cb(4) = Cb(5) = Cb(6) = Cb(7) = 0. Dans le SN du club de karaté de Zachary (Zachary 1977), les acteurs ayant déjà des degrés et des proximités importants sont aussi centraux en termes d'intermédiarité. Sachant que la composition du réseau est presque scindée en 2 sous-ensembles, les acteurs 3, 9, 14, 20, 31 et 32 jouent un rôle intermédiaire central assez stratégique malgré leur degré inférieur ((Erétéo 2011)). Ils ne sont pas seulement centraux, mais leur absence ou la rupture de leurs liens fragilisera la cohésion du réseau et le coupera en deux parties ((Erétéo 2011)).

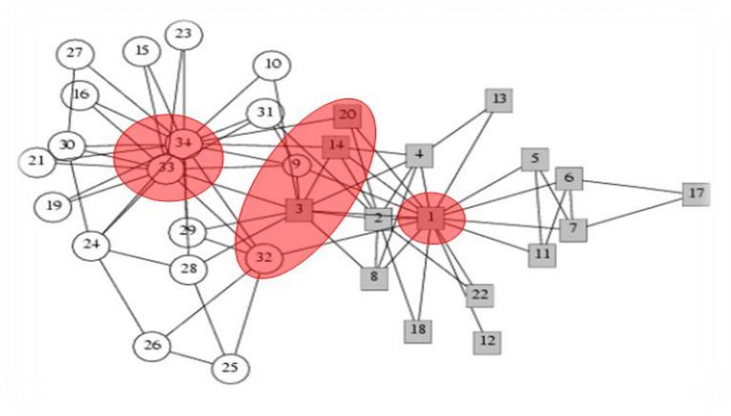


Figure 11. Les acteurs les plus centraux en termes de centralité d'intermédiarité dans le SN du club de karaté de Zachary (Zachary 1977) ((Erétéo 2011))

La suppression des nœuds centraux intermédiaires conduit généralement à ralentir les flux qui devraient passer par des chemins plus longs (Ducruet 2010).

Remarque (Normalisation): La centralité d'intermédiarité d'un nœud v_i se normalise entre 0 et 1, par la valeur maximale théorique (n-1) (n-2)/2 ou (n-1) (n-2) ou par la centralité d'intermédiarité du nœud le plus centrale (le plus intermédiaire) dans le SN.

$$\text{Normalisation (Cb(i))} = \frac{\text{Cb(i)}}{\max\limits_{j=1..n} \text{Cb(j)}} \quad (\mathbf{13})$$

3 types d'algorithmes sont distingués pour calculer la centralité d'intermédiarité ((Erétéo 2011)):

Tableau 2. Algorithmes pour la centralité d'intermédiarité

	Algorithmes exactes	**Heuristiques basées sur l'échantillonnage**	**Approches en parallélisme**
Aperçu	Cb calculée selon sa définition formelle littéralement. Ils sont basés sur les géodésiques	Des estimations de Cb. Certains calculent Cb sur un échantillon de nœuds distribué aléatoirement dans le réseau	Certain basé sur 2 algorithmes qui traitent le SN en parallèle ou un seul algorithme avec des procédures qui s'exécutent en parallèle ou selon une approche incrémentale
Type de graphes sociaux	Graphes, pondérés, non pondérés, graphes multipartites	Graphes simple : non étiquetés, non-orientés en acceptant la pondération	Graphes simple : non étiquetés, non-orientés en acceptant la pondération
Taille	Applicable sur des petits réseaux. Taille max est de l'ordre de 10^5 nœuds pour le plus	Applicable sur des réseaux plus grands avec un ordre de 10^6 nœuds	Ils peuvent traiter des SNs d'un ordre de million de nœuds

	performant		
Précision / performance	Efficace et précis	Moins précis et plus performants. La qualité dépend de la technique d'échantillonnage	En général des contributions majeures en performance. Des résultats précis ou des bonnes approximations
Complexité	Pour le plus efficace est : En espace : $O(n+m)$ En temps : $O(n.m)$ pour les graphes non-pondérés et $O(n.m+\log 2(n))$ pour le cas pondéré	-	Généralement pas d'estimation
Exemples	Le plus efficace est celui de Brandes et Newman (Erétéo 2011) (Brandes 2001)	-	Approche incrémental et parallèle de Santos (Erétéo 2011)

Généralement, un algorithme exact est performant, à partir de lequel, des alternatives et variantes ont été proposées pour calculer la centralité d'intermédiarité. Des considérations supplémentaires sont inclus comprenant par exemple l'importance de position du nœud sur une géodésique ou encore la longueur des géodésiques (Erétéo 2011), etc. Il existe même des alternatives focalisées sur l'intermédiarité des liens, des groupes de nœuds (Erétéo 2011), etc. Par exemple « **Edge Betweenness** » est une approche très importante pour mesurer la centralité des liens, utiles pour distinguer des structures modulaires, en se basant sur la même définition formelle. Cependant les approches qui visent à adapter la définition de la centralité d'intermédiarité sur les graphes sociaux multipartites ainsi qu'aux hypergraphes, restent insuffisantes dans la littérature (Erétéo 2011). D'autre part, plus l'échantillon est représentatif de l'ensemble du réseau plus les résultats des autres heuristiques sont plus précis. C'est pourquoi certains auteurs se sont basés sur l'exploitation de la propriété du petit monde des SNs (Erétéo 2011).

La centralité d'intermédiarité reste encore un concept attrayant de nombreux travaux récents. Par exemple en géographie, elle est utilisée par ((Rozenblat 2010)) pour comprendre les logiques de localisation des firmes multinationales dans les villes européennes. Elle est aussi utilisée par ((Comin 2009)) pour montrer la polarisation du réseau de collaborations scientifiques européennes par les grandes villes. Un réseau (de villes) est dit polarisé (intégral), s'il y a une ville (un nœud) principale qui domine toutes les autres villes (secondaires) de telle sorte que toutes les fonctions importantes se concentrent dans cette ville. C'est l'exemple du réseau urbain français des villes (Une infrastructure de liens: voiries, transport, canalisation, câblage, etc.) qui est polarisé autour de Paris (Ducruet 2010).

3.3.1.1.2.3. Autres centralités (Eigenvector Centrality)

Il existe des définitions de centralité moins connues par rapport les indices précédents comme : La centralité vectorielle, de flux, de pontage (Bridging Centrality), etc. Par exemple, la centralité de flux, 'Flow Centrality', est une définition plus vaste par rapport l'intermédiarité car elle est définie par le nombre de tous les chemins qui passent par un nœud (Nettleton 2013). Il y a aussi la centralité **égocentrique** qui mesure l'influence d'un acteur dans un sous réseau au voisinage (Sous-graphe : son réseau adjacent) ((Erétéo 2011)). En outre, 'Eigenvector Centrality' ou la centralité vectorielle est une mesure de centralité d'ensemble mais basée sur la centralité de degré : une version itérative de Cd. Dans ce cas, la centralité d'un nœud est proportionnelle à la somme des centralités de ses voisins directs. Généralement, un nœud ayant une centralité vectorielle élevée, est lié aux nombreux nœuds, chacun possède un degré important et ainsi de suite. Noter que cet indicateur est mal appliquée sur un SN avec des liens asymétriques et souvent remplacé par un indice raffiné (**PageRank**) qui donne des résultats plus significatifs.

3.3.1.1.3. Bilan sur les centralités individuelles

Les méthodes de calcul abordées de ces 3 mesures principales de centralité (Degré, proximité, intermédiarité) sont des standards. Ces méthodes selon Freeman (Freeman 1977) (Erétéo 2011) calculent l'indice en fonction de la taille du réseau pour mesurer l'influence et l'activité d'un nœud dans son réseau. Cependant, les versions normalisées sont des approches suggérées pour ne pas tenir en compte ce paramètre de taille, utilisées pour comparer des centralités entre les différents réseaux. En divisant par la valeur maximale du nœud central théorique ou réel, les normalisations abordées ci-dessus sont utiles pour rendre ces mesures indépendantes de la taille du réseau.

Tableau 3. Centralités individuelles Degré Vs. Proximité Vs. Intermédiarité

	Centralité de voisinage : Centralité de degré	**Centralité d'ensemble**	
		Centralité d'intermédiarité	**Centralité de proximité**
La portée	La centralité dépend de ses voisins (liens adjacents) : Potentiel individuel local	La centralité dépend des chemins géodésiques : Mesurer le potentiel individuel sur l'ensemble du réseau	
Connectivité	-	Calcul possible sur des graphes pas fortement connectés	Calcul discutable sur les graphes qui ne sont pas connexes
Complexité	Triviale	Complexité en temps $O(n.m)$ $O(m)$ pour de certains algorithmes d'approximation	

		(Cc)

Si l'influence des individus est estimée, certains auteurs comme Kempe et Kleinberg (Kempe et al 2003) ont cherché comment maximiser **la propagation de l'influence** dans le SN (Nettleton 2013). Le problème est NP-difficile car il consiste initialement à choisir le sous-ensemble d'individus qui peut influencer la prise d'une décision par exemple, pour acheter un produit ou un service d'un nombre maximal des autres individus dans l'ensemble du réseau et qui vont faire la même chose (Nettleton 2013). Cependant la propriété du petit monde dans les SNs implique qu'une bonne connectivité au voisinage reflète souvent l'influence globale du nœud dans le réseau. En d'autre terme, le degré est une bonne estimation (Heuristique de recherche) de la proximité ou de l'intermédiarité d'un nœud: Un degré élevé signifie que le nœud est plus proche plus intermédiaire. En allant plus loin, selon (Scott 2000), la centralité d'un nœud sera plus représentative si la centralité de ses nœuds adjacents est prise en compte, à l'image de 'Eigenvector Centrality'. Mais généralement, la centralité d'intermédiarité semble être la métrique la plus significative en SNA pour identifier des positions individuelles hautement stratégiques, soit sur les flux de communications ou par rapport **la résilience et la cohésion du réseau** ((Erétéo 2011)). Toutefois sa complexité de calcul de l'intermédiarité reste une limite pour appliquer ses approches et ses algorithmes sur **des SNs larges sur le web** qui peuvent contenir plusieurs des millions de sommets.

3.3.1.2. Mesures de prestige : Centralités raffinées (cas orienté)

L'extension des métriques de centralité sur le cas orienté représente une bonne illustration de certaines versions étendues, plus simples et plus informatives. Les relations asymétriques ou les arcs dans les graphes sociaux incorporent souvent une information implicite (la sémantique d'orientation). De ce fait, le sens des liens devient important dans la modélisation ainsi pour une analyse plus raffinée. Les chemins doivent être composés par des arcs ayant la même orientation, ce qui est primordial, par exemple, pour analyser et acheminer la propagation d'information dans un réseau. Par conséquent, l'interprétation des mesures de centralité est nuancée en évoquant la notion de prestige selon l'orientation des arcs: un arc sortant ou rentrant d'un nœud 'v' s'interprète respectivement comme l'influence ou le support de 'v'. Le prestige est donc une mesure de centralité raffinée (Malek 2009), qui concerne beaucoup plus le support du nœud.

Degré de prestige

2 indices sont distingués depuis la centralité de degré selon l'orientation des arcs. Le 'Out_Degree' mesure l'influence de l'activité d'un nœud. Alors que le 'In_Degree' mesure le support, plus précisément le prestige d'un acteur: C'est le degré de prestige (Malek 2009):

$$\mathrm{Pd(i)} = \frac{\mathrm{In\_Degree(i)}}{\mathrm{n}-1} \qquad (\mathbf{14})$$

Cette équation définit le degré de prestige d'un nœud v_i : le nombre des arcs rentrant, normalisé par le prestige maximal théorique. Un acteur prestigieux est un acteur ayant beaucoup d'arcs rentrants (Malek 2009).

Prestige de proximité

La capacité d'un acteur de se connecter avec un autre acteur dans le réseau, mesurée par la centralité de proximité se distingue en 2 définitions dans le cas orienté. Mais elle s'interprète souvent comme le prestige de proximité qui évalue la capacité d'un nœud v_i d'être atteint (son rapprochement des autres). C'est l'évaluation de son **accessibilité**

$$\mathrm{Pp(i)} = \frac{1}{\sum_j^n \mathrm{d(j,i)}} \qquad (\mathbf{15})$$

Tel que : $\mathrm{d(j,i)}$ est la distance géodésique sur les chemins exclusivement de v_j vers v_i. Le prestige de proximité est entre 0 et 1. Si la valeur tend vers 1, l'acteur est plus accessible, plus proche des autres.

Rang de prestige (PageRank)

C'est un algorithme appliqué spécialement dans les graphes orientés afin de mesurer et classer la réputation et l'importance des acteurs en choisissant l'acteur i (Malek 2009). Le principe de calcule se base sur un principe similaire de celui de la centralité vectorielle: l'importance d'un nœud est basée sur l'importance de ses voisins et ainsi de suite. En d'autre terme, le principe est inspiré de la même hypothèse de l'attachement préférentiel.

$$\mathrm{PR(i)} = \mathrm{A1i.PR(1)} + \mathrm{A2i.PR(2)} + \cdots + \mathrm{Ani.PR(n)} \qquad (\mathbf{16})$$

Le rang de prestige PR(i) d'un nœud v_i est évalué depuis les PR de ces nœuds voisins pondérés par 'Aij' (A est la matrice d'adjacence) qui vaut 1 ou plus, si v_i pointe sur v_j. Les valeurs de tri de n acteurs se représentent dans un vecteur P, (Malek 2009):

$$\mathrm{P} = (\mathrm{PR(1)}, \mathrm{PR(2)}, \ldots, \mathrm{PR(n)})^{\mathrm{T}} \qquad (\mathbf{17})$$

Selon PR(i) qui est pondéré par les valeurs de la ligne i dans la matrice d'adjacence A, alors :

$$\mathrm{P} = A^T P \qquad (\mathbf{18})$$

Noter bien que cet algorithme est très connu pour évaluer et trier la réputation des pages web par des moteurs de rechercher comme Google (algorithme PageRank), (Malek 2009). C'est la méthode de classement des pages web ayant le plus de succès

Bilan sur les indices de prestige

Certaines mesures de centralités peuvent être raffinées selon l'orientation des arcs sous forme d'indices de prestige, focalisés sur le support (la cible de l'arc). Ce sont des indices qui déterminent le prestige (la réputation) d'un acteur par rapport à ses liens rentrants: Des métriques plus nuancées.

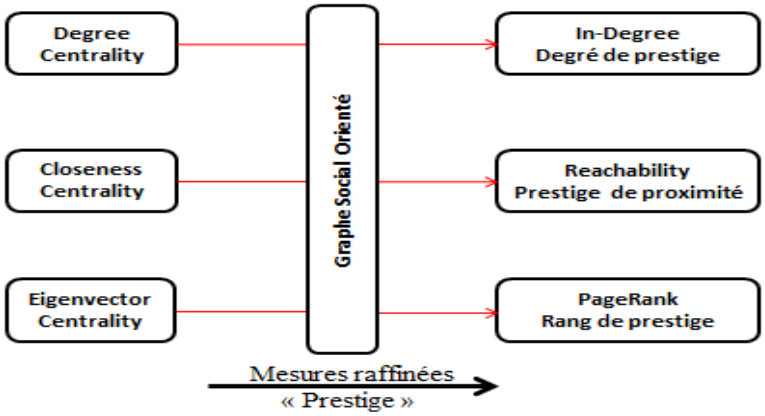


Figure 12. Des indices de prestiges : Des mesures de centralité raffinées par l'orientation des liens

La centralité d'intermédiarité qui mesure le contrôle d'un nœud donné sur l'échange d'information dans le réseau n'est pas vraiment extensible. Mais son calcul est légèrement modifié en considérant l'orientation des arcs ou encore le sens de circulation de l'information.

3.3.2. Analyse locale, une vue sur la structure globale

L'ensemble des métriques abordées ci-dessus, caractérise souvent les traitements les plus cités dans une analyse locale d'un SN. La localité ici se concentre sur le nœud lui-même. Sa pertinence et son influence est mesurée soit au niveau de son voisinage ou sur l'ensemble du réseau. Cependant, les études statistiques de ces métriques et les positions stratégiques individuelles détectées peuvent donner parfois des explications sur certaines lois d'organisation sur la structure globale des SNs.

3.3.2.1. Centralisation du SN et sa dépendance avec ses acteurs centraux

La dépendance d'une structure sociale de ses individus peut s'évaluer en termes de centralisation ou centralité globale (Erétéo 2011) (Freeman 1979) depuis certains acteurs centraux, dominateurs. Le concept de centralité globale ou centralisation d'un SN n'est qu'une généralisation des centralités locales des nœuds. Il y a deux approches pour la calculer : une approche 'variationnelle', et celle de Freeman (Elle prend la valeur 0 lorsque toutes les centralités sont identiques -réseau cycle- ou tend vers 1 (réseau dominé totalement par un seul individu -réseau en étoile-). Freeman fournit une formule de calcul qui peut s'adapter avec les trois indices principaux de centralité locale (Freeman 1979). Elle consiste à mesurer l'écart entre les valeurs de centralités des nouds et la valeur du nœud le plus central. Son calcul et sa signification dépend de la définition de la métrique locale, choisie en termes d'activité, accessibilité, intermédiarité, etc. La centralité globale reflète principalement :

- L'activité du réseau se concentre autour de certains acteurs (Erétéo 2011) avec la centralité de degré.
- La dépendance du réseau en termes de connectivité et efficacité de ses acteurs en utilisant les intermédiarités.
- La performance de communication et d'accessibilité en se basant sur les centralités de proximité (Erétéo 2011).

Selon ((Gil-Mendieta & Schmidt 1996)), *plus le réseau est centralisé, plus il y a des régions cohésives plus influentes*.

3.3.2.2. Résilience et cohésion du SN devant des positions locales exceptionnelles

La dépendance du SN ne s'estime pas seulement par sa centralité globale dominée par des acteurs centraux. Sa connectivité peut être également **sensible** ou non à certains acteurs ou liens (centraux intermédiaires), ((Erétéo 2011)).

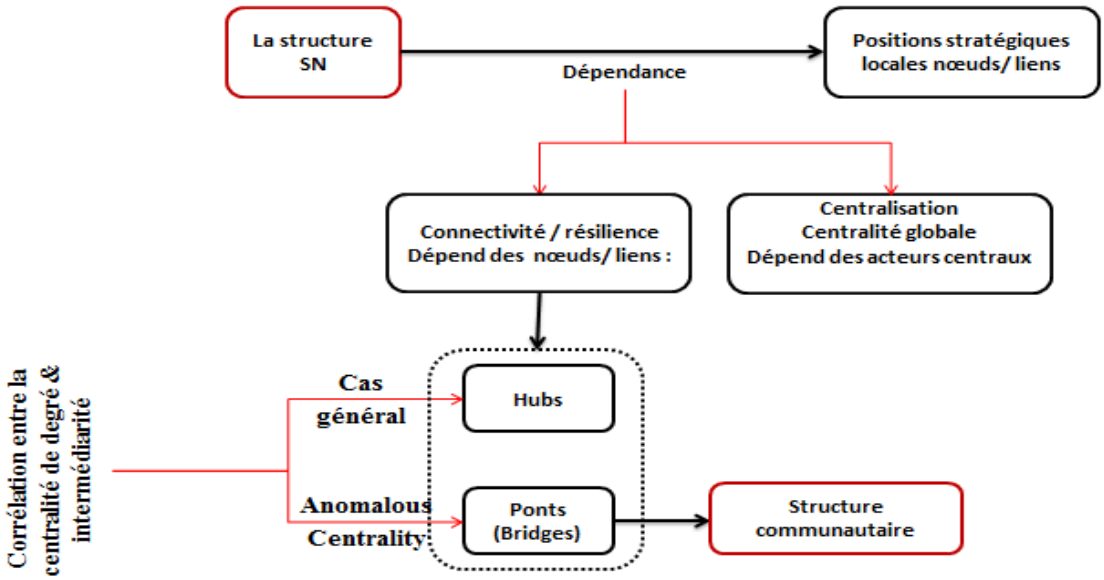


Figure 13. La structure SN et sa dépendance avec positions stratégiques nœuds/ liens

Des études comme la distribution statistique d'un indice de centralité de degré qui montre une configuration selon la loi de puissance est devenue un paramètre descriptif qui distingue la structure de graphe social d'une configuration aléatoire. En outre, les corrélations statistiques entre certains indices sont également assez informatives sur certaines configurations particulières en termes de cohésion qui ne sont pas perceptibles à partir les centralités locales soit au niveau des individus ou même des liens. Noter que cette corrélation entre certains indices (au voisinage face à l'ensemble) se justifie souvent par l'effet du petit monde dans les SNs. Dans les réseaux en général, la corrélation entre les centralités de degré et d'intermédiarité des nœuds est souvent forte. Un degré élevé est le synonyme d'une forte intermédiarité pour un nœud qui est également considéré comme un « **hub** » (Ducruet 2010). Par conséquent, cette corrélation reflète une cohésion du réseau qui se résume dans un seul module, alors qu'une structure modulaire en communautés ne se présente pas dans tous les réseaux. Par exemple les réseaux aériens étant plus favorables à la formation des communautés, que ceux du transport maritime (Ducruet 2010). Toutefois, le croisement entre ces 2 indices dans les SNs peut repérer des situations exceptionnelles de certains acteurs ayant une centralité désignée comme anormale « **Anomalous Centrality**». Le terme a été introduit par Guimera en 2005 (Ducruet 2010) et décrit souvent la situation d'un nœud ayant une forte centralité d'intermédiarité sur l'ensemble du réseau malgré sa faible connectivité au voisinage. Un tel nœud représente un connecteur, un pont (**Bridge**) qui serve par exemple dans un réseau aérien de relais entre des régions distantes comme Anchorage (Alaska) entre

l'Asie et l'Amérique du Nord (Ducruet 2010). Dans un SN, ce type de configuration caractérise une **structure communautaire** du réseau. Des régions cohésives apparaissent sous forme des groupes interconnectés par peu de sommets (ponts) (Figure 14).

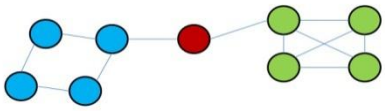


Figure 14. Nœud rouge en position de Bridge qui entretient des liens (non redondants) entre 2 groupes distincts.

Une telle configuration apparait par exemple dans le SN du club de karaté de Zachary dans la (Figure 11), où des nœuds comme 20, 14, 9, relient 2 régions cohésives quasi séparées. Par conséquent, l'absence des nœuds ou liens stratégiques: des hubs ou des ponts peut augmenter la longueur des géodésiques ou couper la communication respectivement, en divisant le réseau en modules non-connectés. La structure du SN et sa connectivité est affectée sérieusement lorsqu'un « retrait ciblée » de ces positions stratégiques (**Attaque ciblée du réseau**) est appliqué par rapport un retrait aléatoire ((Erétéo 2011)).

3.3.2.3. Pont, trou structural

Le trou structural (Structural hole) introduit par (Burt 2004) est une notion qui s'applique parfaitement sur les ponts (Bridges) entre groupes. C'est en effet une séparation entre des contacts non-redondants : qui ne sont pas des voisins directs. La redondance en SN est évoquée dans les voisins directs où les mêmes informations sont probablement partagées : **Des contacts redondants** ((Erétéo 2011)) qui forment souvent des groupes/ sous-groupes, et des nœuds ponts entre eux sont à identifier comme dans le travail de Weiss and Jacobson qui est considéré également comme une première version du concept de centralité d'intermédiarité popularisé par Newman (Parthasarathy et al 2011). Un acteur pont contrôle un trou structural qui présente un bénéfice informationnel dans le réseau suivant 2 atouts majeurs. Il permet un accès rapide en tant qu'un canal et une source fournissant des nouvelles informations non-redondantes pour **les régions cohésives** (pour les contacts redondants). D'autre part, les acteurs les plus proches des trous structuraux (**Des frontières**) sont rapidement et mieux informés, et plus susceptibles d'avoir des ' bonnes idées' grâce au bénéfice informationnel apportés par les trous structuraux ((Erétéo 2011)). Même si ce bénéfice informationnel est techniquement prouvé, selon certains chercheurs, les acteurs ponts ne sont pas permanents et n'ont pas une motivation pour manipuler stratégiquement le réseau (Nettleton 2013). Cependant, ces conclusions restent limitées par le contexte spécifique de l'étude (Données et environnement) et ne sont pas nécessairement généralisables. **Le pont : un nœud ou lien est donc une partie sensible du SN et offre au même temps un bénéfice informationnel.**

3.3.3. Métriques globales d'analyse de structure SN

Les mesures globales sont moins nombreuses et moins diverses par rapport aux indices locaux. Les indices les plus utilisés dans ce plan d'analyse des SNs permettent d'avoir l'information sur la structure globale du réseau.

Distances représentatives

La distance géodésique est souvent le concept fondamental de nombreuses dérivées ainsi que dans des mesures qui s'appliquent sur l'ensemble du réseau. Il y a par exemple la distance moyenne qui présente la moyenne des distances géodésiques entre toute paire d'acteurs. Ou encore le diamètre et sa particularité liée à la propriété du petit monde du SN. Ces distances seront plus significatives pour comparer les SNs dans le cas des graphes connectés. Cependant, ce n'est pas toujours le cas avec les graphes sociaux.

Densité

La densité n'est qu'une mesure utilisée pour décrire la connectivité et déterminer la cohésion du graphe social. C'est l'indicateur utilisé pour mesurer à quel point le graphe est complet (Santoro et al 2011). La définition la plus simple de la densité se base sur la quantité des liens par rapport la taille (le nombre des acteurs) du réseau (Padgett 1994) (Erétéo et al 2008).

Tableau 4. Densité et ses définitions

Densité (D) du graphe social		
Définition simple	Définitions étendues : Densité relative	
$D = m / n$. m : Le nombre des liens. n : Le nombre des nœuds.	$D = m / \|E^*\|$ E^* : C'est l'ensemble des liens théorique dont la taille est le nombre maximal des liens dans un graphe complet de n nœuds ((Nettleton 2013)). Par exemple : n (n-1) ou : n (n-1) / 2 dans le cas orienté ou non orienté respectivement	$D = Z / \|Z^*\|$ C'est le quotient du nombre des 0 dans la matrice d'adjacence, divisé par le nombre des 1 ((Nettleton 2013))

Une autre définition étendue s'appelle 'Densité relative' (Padgett 1994) qui est définit selon 2 métriques (Tableau 4). La densité montre également l'influence dans une analyse égocentrique autour d'un nœud donné : Densité de son sous-graphe (Cuvelier & Aufaure 2011) en voisinage et même dans une analyse socio-centrique (Scott 2000). **Contrairement aux graphes en général, la signification de la densité des SNs et son explication dépasse le cadre d'une métrique, étant liée avec plusieurs caractéristiques du SN**:

En effet, la densité dans un SN est limitée tout dépend du type des relations sociales dont l'établissement et la gestion sont couteux en temps, en ajoutant aussi le coût cognitif inhérent au maintien de ces relations ((Erétéo 2011)). Un réseau de relations professionnelles est beaucoup plus dense qu'un autre formé par des relations amoureuses notamment en raison des caractéristiques de ces liens (L'exclusivité des relations, le coût de maintien, etc.), ((Erétéo 2011)). De ce fait, le nombre des contacts qu'une personne peut les conserver est limité. En conséquence, la densité du réseau ne sera pas remarquablement influencée par rapport aux nombre des acteurs même si le SN est grand. C'est en effet l'image de **la loi de puissance** qui régit l'invariance à l'échelle d'un SN. Noter bien que la densité des liens qui reflète la cohésion du réseau montre également la tendance à s'organiser **en structures**

communautaires ou non. Plus le réseau est dense et cohésif, plus **les parties du réseau** sont de plus en plus réunis.

Les mesures abordées ci-dessus sont très diverses et cet inventaire est loin d'être exhaustif. **Le plan de l'analyse peut s'élargir sur l'ensemble du réseau non seulement à travers des métriques locales ou même globales, mais aussi avec des approches et des techniques pour étudier l'organisation du SN en structure communautaires**.

3.4. Structures communautaires

L'une des caractéristiques d'un modèle de SN est d'afficher des régions cohésives**: des groupes d'acteurs densément connectés,** connues comme des communautés là où les acteurs sont beaucoup plus liés entre eux. Même si dans la littérature, il n y a pas un accord complet pour définir le concept de communauté. Cependant, d'un point de vue original purement social:

« Un groupe est une communauté sociale dont les membres coopèrent plus souvent au sein du groupe qu'à l'extérieur » (Kazienko et al 2011).

« Un groupe peut être défini comme un sous-ensemble d'utilisateurs fortement connectés entre eux et moins connectées avec les membres d'autres groupes » (Kazienko et al 2011).

« C'est un ensemble d'individus qui s'interagissent fréquemment » (Tantipathananandh et al 2007). L'objet et le maintien des interactions dans un groupe qui porte souvent sur des sujets connexes ou des points de vue communs partagés par ses membres révèlent des propriétés intéressantes (Tantipathananandh et al 2007):

- Selon (Burt 1992) (Burt 2004), plus les contacts d'une personne sont reliés entre eux, plus son comportement est contraint par le réseau ((Erétéo 2011)).
- La communauté incarne la fermeture réseau et la redondance des contacts: Selon Coleman et Burt, l'information se propage rapidement dans une communauté et est partagé par la plupart de ses membres. D'où, ils connaissent l'information détenue par chacun ((Erétéo 2011)).
- Dans le contexte commercial, institutionnel éducatif, etc., la formation des groupes (des communautés) améliore l'efficacité de communication.

L'aspect de pénalisation comme la prise de confiance (une variable cachée) sont fréquents dans la redondance des contacts dans une communauté ((Erétéo 2011)). En effet, si le partage d'information rapide facilite la décision de faire confiance à quelqu'un, les erreurs d'une personne se propagent aussi rapidement (commençant par ses contacts directs) et engendrent ainsi sa sanction qui se traduit par l'isolement ou la perte (la révocation) de confiance dans le réseau (Erétéo 2011) (Massa et al 2009) ((Srivastava & DeLong 2013)). La pénalisation dans un groupe tend à éviter la diffusion de mauvaises informations et les mauvais comportements (Erétéo et al 2008). D'après certains auteurs (Tantipathananandh et al 2007), le terme groupe et communauté ne signifie pas toujours la même chose. Selon (Tantipathananandh et al 2007), un groupe capture un moment (time-step) d'interactions. Tandis qu'une communauté est un concept latent qui couvre la plupart des interactions mais pas nécessairement l'intégralité des

interactions (Tantipathananandh et al 2007). Cependant sur **un premier plan formel statique**, cette distinction peut s'effacer.

3.4.1. Contraintes de représentations et descriptions formelles

Le fait de modéliser un SN par **une structure topologique** de graphe, une communauté en théorie de graphe et d'un point de vue clustering est **un cluster de nœuds** (Cuvelier & Aufaure 2011). Une communauté apparait comme un sous-graphe ayant une connectivité interne remarquable (cohésif), qui peut être théoriquement saisi dans des patterns stricts : clique, etc., (Scott 2000) (Erétéo 2011), mais de plus en plus étendus en s'adaptant à certains critères. Même si les communautés restent vaguement définies (Tantipathananandh et al 2007), les auteurs essayent de mettre des contraintes sous lesquelles, un groupe peut exister, plutôt qu'une définition précise (Tantipathananandh et al 2007). Certains auteurs ont abordé un ensemble de critères qui permettent de qualifier une collection en une communauté (Cuvelier & Aufaure 2011). Par exemple les 4 critères de Wassermann sont:

a. Mutualité complète

Elle signifie que tous les membres d'un sous-graphe doivent être liés entre eux (Chacun doit être "l'amis" de tous les membres du sous-graphe). En théorie des graphes, la mutualité complète se manifeste dans une structure de clique (Cuvelier & Aufaure 2011).

- **Définition** 20. Composante : C'est un sous-graphe connexe, isolé ((Erétéo 2011)).
- **Définition** 21. Composante forte : C'est un sous-graphe (Composante) qualifié dans le cas orienté comme fortement connecté, dans lequel chaque chemin est formé par des arcs qui ne change pas de direction sinon c'est une composante faiblement connectée ((Erétéo 2011)).
- **Définition** 22. Clique: C'est un sous-graphe complet (composante) qui représente la structure classique de communauté (Nettleton 2013).

Certains SNs peuvent afficher des réseaux de cliques (Des réseaux à base de cliques). C'est l'image par exemple des réseaux des relations d'amitié, ou des acteurs d'un film cinéma, des collaborateurs scientifiques, etc. Dans un graphe étant une entité mathématique, ces définitions rigoureuses strictes présentent une base formelle pour décrire une communauté. Cependant ces patterns sont trop restrictifs pour un SN et les particularités de ses caractéristiques en tant qu'un graphe social ((Erétéo 2011)). Plusieurs définitions étendues de clique sont ainsi proposées pour couvrir un deuxième critère : Accessibilité (Nettleton 2013).

b. Accessibilité

C'est un critère qui exige l'existence d'au moins une géodésique dont la longueur ne dépasse pas n (sauts/ houblons), (Cuvelier & Aufaure 2011), entre chaque paire de nœuds. D'abord, l'accessibilité collective au sein d'un groupe de nœuds est définie par une composante connectée (Tang et al 2010a). Autrement dit, la connexité se remplace par la connectivité. Prenant le cas orienté, l'accessibilité d'un nœud i est plus raffinée et définie par 'in-component/ out-component'. C'est l'ensemble des nœuds qui peuvent atteindre (in-component) et être atteint (out-component) par i (Tang et al 2010a). D'où, l'accessibilité se voit comme une composante faiblement connectée ou

fortement connectée (où il existe à la fois un chemin de i à j et de j à i). Cependant, dans le cas non orienté, un nœud ne peut appartenir qu'à une seule composante connectée (Tang et al 2010a). L'exigence sur la longueur des chemins laisse l'accessibilité un critère qui fait appel à la notion de n-clique.

- **Définition** 23. n-Clique: C'est une composante, dans laquelle la distance maximale entre 2 nœuds donnés ne dépasse pas n.

Une clique classique est le synonyme de 1-clique. Cependant, une géodésique dans n-clique peut fonctionner à l'extérieur en couvrant des nœuds exclues de ce sous-graphe car son diamètre peut être supérieur à n. Dans ce cas, n-clan peut remplacer n-clique (Cuvelier & Aufaure 2011) (Erétéo 2011).

- **Définition** 24. n-Clan: C'est une composante dont tous les nœuds sont connectés par des chemins comprenant le diamètre dont la longueur ne dépasse n.

c. Degré nodal

C'est une contrainte imposée sur le nombre des nœuds adjacents dans le sous-graphe, qui peut être supportée par le pattern : k-plex.

- **Définition** 25. k-plex : C'est une composante, dans laquelle chaque nœud est adjacent à tous les nœuds à l'exception de k nœuds au maximum dans ce sous-graphe (Erétéo 2011). K-plex est généralement un pattern plus robuste.
- **Définition** 26. k-core : C'est un sous-graphe maximal dans lequel, chaque nœud est adjacent à, au moins, k nœuds dans cette composante. C'est l'inverse de k-plex.

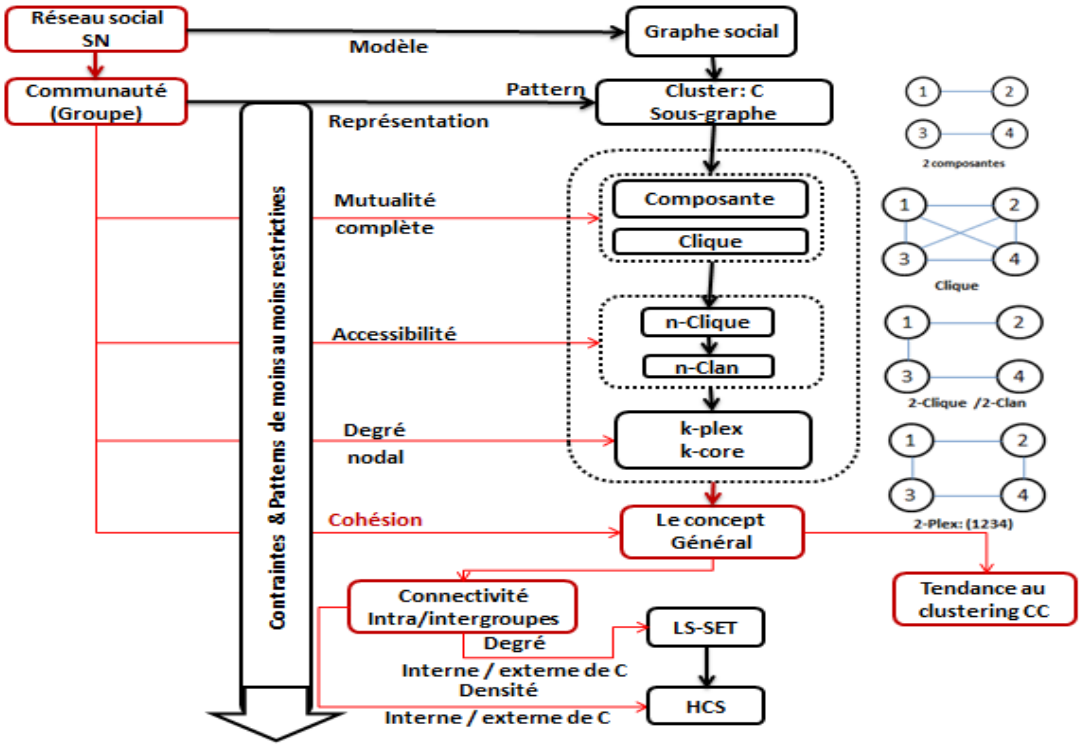


Figure 15. Le concept de communauté dans le graphe social représenté par des patterns de moins au moins restrictifs

La propriété commune de ces patterns, est la connectivité (la cohésion) qui caractérise le concept général de la communauté sur le graphe social.

d. Cohésion

La formation et la cohésion intragroupe dans le SN doit être expliquée formellement partant de la tendance des individus à former un groupe. Les SNs ont généralement une forte tendance à s'organiser en groupes, une **tendance au clustering**, liée à une caractéristique importante qui est la transitivité. La transitivité reflète en effet la tendance de l'individu à se socialiser en groupe

❖ Tendance au clustering

La transitivité dans le SN se représente formellement par un **cycle triadique** (triade) de longueur 3 connectant 3 nœuds différents. C'est un pattern très fréquent dans le graphe social, ce qui explique souvent une forte tendance au clustering mesurée par un coefficient de clustering « CC » d'abord de l'individu et du réseau globalement.

- **Définition** 27. Coefficient de clustering de v_i : C'est le rapport du nombre de triades existants sur le nombre maximum des triades possibles dont le nœud v_i fait partie.

$$\mathrm{CC\,i} = \frac{|\mathrm{TRIADES\ i}|}{|\mathrm{TRIPLETS\ i}|} \qquad (\mathbf{19})$$

|TRIPLETS i|, est le nombre théoriques des triades probables entre chaque triplet de nœuds (contenant vi), qui est noté également par|2_PATH i |. (Chemins à 2 liens), (Erétéo 2011). Le taux global de clustering du SN peut se calculer alternativement comme une moyenne de ces valeurs locales.

$$\mathrm{CC} = \frac{1}{\mathrm{n}} \sum_{i}^{n} \mathrm{CCi} \qquad (\mathbf{20})$$

Ou à travers la généralisation de la version individuelle, proposée par Watts et Strogatz en se basant sur le concept de transitivité.

$$\mathrm{CC} = \frac{3 \times |\mathrm{TRIADES}|}{|\mathrm{TRIPLETS}|} \qquad (\mathbf{21})$$

❖ Connectivité Intragroupe/ Intergroupes

La cohésion qui décrit le concept général d'une communauté (cluster : C dans un graphe G (V, E)) s'exprime sur le plan topologique par une connectivité minimale entre groupes et maximale au sein de chaque groupe. Une telle configuration peut être évaluée et prouvée en se basant sur plusieurs notions: Degré interne et externe du cluster, densité inter/intra-cluster, etc. Sachant que $C \subset V$ alors :

Le degré interne de C est le nombre de liens internes dont les 2 extrémités se trouvent dans C.

$$\mathrm{Deg_{int(C)}} = |\{e(vi, vj) \setminus vi, vj \in C\}| \qquad (\mathbf{22})$$

Le degré externe de C est le nombre de liens dont l'une des extrémités se trouve dans C, et l'autre en dehors de C (Cuvelier & Aufaure 2011).

$$\mathrm{Deg}_{\mathrm{ext(C)}} = |\{e(vi, vj) \setminus (vi \in C \wedge vj \in V\setminus C) \vee (vj \in C \wedge vi \in V\setminus C)\}| \quad (\mathbf{23})$$

Donc une forte communauté présentée par le sous-graphe cluster C est exprimée d'après le degré interne et externe de chaque membre: $\forall \boldsymbol{vi} \in \mathbf{C}$, $\mathbf{Deg(vi)_{int(C)} > Deg(vi)_{ext(C)}}$ **qui prouve la cohésion interne. Une forte communauté fait référence au concept « LS-SET »**

- **Définition** 28. LS-SET: C'est un sous-ensemble de sommets S tel que tout sous-ensemble de S (différent de S lui-même) a plus de liens vers son complément dans S qu'à l'extérieur de S (Erétéo 2011).

Si $\mathrm{Deg(vi)}_{\mathrm{ext(C)}} = 0$ alors $vi \in \mathrm{C}$ est une assignation parfaite. Si $\mathrm{Deg(vi)}_{\mathrm{int(C)}} = 0$ alors $vi \notin \mathrm{C}$ **La conductance d'une communauté** C, est ainsi définie comme le rapport entre la taille de la coupure (Cut) entre C et V\C et le minimum entre le degré total de C et son complément (Cuvelier & Aufaure 2011) :

$$\phi(\mathrm{C}) = \frac{\mathrm{c(C, V\setminus C)}}{\min\{\mathrm{deg(C)}, \mathrm{deg(V\setminus C)}\}} \quad (\mathbf{24})$$

La conductance atteint son minimum lorsque la taille de la coupure est assez faible et le degré total de C et son complément sont égaux (Cuvelier & Aufaure 2011). Dans les de réseaux du monde réel, les meilleures communautés évaluées par la conductance sont selon Leskovec (Leskovec et al 2009) (Parthasarathy et al 2011) des groupes notés 'Whiskers' liés uniquement par un seul lien avec le reste du graphe (ou 2 liens entre groupes : 2- Whiskers), (Parthasarathy et al 2011) (Leskovec et al 2009).

Une communauté ayant une quantité importante de liens internes (Kazienko et al 2011) doit avoir également une grande densité relative interne (Cuvelier & Aufaure 2011), calculée à travers la même définition de densité (Tableau 4) appliquée sur le sous-graphe correspondant C : **Densité intra-cluster**, notée$\mathrm{D}_{\mathrm{int(C)}}$. **La densité inter-cluster**, notée $\mathrm{D}_{\mathrm{ext(C)}}$ mesure selon (Cuvelier & Aufaure 2011) le nombre des liens externes dont l'une des extrémités est un nœud de C, et l'autre en dehors de C par rapport les liens totalement externes du graphe entre les nœuds $V\setminus C$

$$\mathrm{D}_{\mathrm{ext(C)}} = \frac{|\{\{\mathrm{v, u}\}\setminus (v \in C \wedge u \in V\setminus C) \vee (u \in C \wedge v \in V\setminus C)\}}{|\mathrm{C}| \times (|\mathrm{V}| - |\mathrm{C}|)} \quad (\mathbf{25})$$

Cependant une configuration composée par K clusters d'un graphe social selon une partition donnée $\{\mathrm{C}_1, \ldots, C_k\}$ n'est pas clairement impliquée. Selon ces notions qui décrivent la cohésion d'une communauté et sa connectivité interne et externe, la somme des densités intra-cluster est fortement supérieure de densité du graphe social G (Cuvelier & Aufaure 2011):

$$\sum_{k}^{K} D_{int(C_k)} \gg \ D(G) \qquad (\mathbf{26})$$

Au-delà des patterns moins au moins restrictifs, le concept général d'une communauté se décrit communément par un sous-graphe fortement connecté (**HCS: Highly Connected Subgraph**), (Cuvelier & Aufaure 2011). Toutefois la cohésion des communautés est définie sur un modèle de représentation topologique de graphe social alors qu'en effet les entités sociales se regroupent, se socialisent et se connectent en communautés selon **des tendances communes qui ne sont pas forcément explicites.** Dans certains œuvres, des auteurs ont allé plus loin en proposant des modèles de formation d**'équipe implicitement connectée en ligne sur des SNs,** qui s'appelle ICT «Implicitly Connected Teams » (Nettleton 2013). Un ICT est une équipe Q qui n'est pas nécessairement formée par un sous-graphe connexe (pas de connections directes). Il suffit d'avoir des chemins de communication entre les personnes de l'équipe, qui passent par d'autres individus de SN et qui ne sont pas forcément affiliés à Q (Nettleton 2013). Par ailleurs, des outils conceptuels analytiques comme 'Viewpoint Analysis' : VPA sont récemment introduits pour comprendre et définir des groupes influents à partir le point d'un nœud ou un sous-ensemble de nœuds dans les grands réseaux (Parthasarathy et al 2011).

Cependant tels modèles restent limités et prouvés dans un contexte spécifique de l'étude.

❖ Appartenance et Similarité

La similarité est une formalisation d'un l'élément déclencheur possible de la connectivité dans un groupe.

« *Un groupe est un ensemble d'objets à l'intérieur d'un cluster, qui doivent être plus semblables que des objets à l'extérieur du groupe: en maximisant la similitude intra-classe et en minimisant similitude interclasses »* (Cuvelier & Aufaure 2011). Donc, l'appartenance d'un individu à une communauté et la distance entre 2 ensembles (2 communautés), peuvent être évaluées par diverses mesures de similarité ou di-similarité (mesures de distances). Ce sont des opérateurs fréquemment utilisés en clustering, par exemple l'algorithme ou l'indice de Jaccard (Cuvelier & Aufaure 2011), distance euclidienne, etc., pour découvrir la formation des communautés. D'abord, les mesures qui évaluent l'appartenance d'un individu à un cluster (communauté) se trouvent en 2 catégories. Il existe des mesures de similarité « s » des mesures de di-similarité « d » (des distances) entre 2 individus u, v ayant probablement chacun un vecteur d'attributs.

Tableau 5. Appartenance et distance en termes de similarité

	Similarité « s »	**Di-similarité « d »**
Contraintes	$s(u,u) = k$ (k : constante) $s(u,v) = s(v,u)$ Symétrie $s(u,v) \leq s(u,u) = k$	$d(u,u) = 0$: Séparation $d(u,v) = d(v,u)$ Symétrie $d(u,v) \leq d(u,v) +$

		$d(v,u)$ inégalité triangulaire
Conversion	$\mathcal{F}(s) = d$ et $\mathcal{F}^{-1}(d) = s$	
Métriques de proximité utiles comme attributs entre 2 nœuds u, v	Nombre de voisins communs ($Neigh$ l'ensemble des voisins d'un nœud) $CN(u,v) = \|Neigh(u)\| \cap \|Neigh(v)\|$ Attachement préférentiel ((Newman 2001a)) $PA(u,v) = \|Neigh(u)\| \times \|Neigh(v)\|$ Coefficient de Adamic/ Adar ((Adamic & Adar 2003)). $AA(u,v) = \frac{1}{\sum_{z \in Neigh(u) \cap neigh(v)} \log(Neigh(z))}$ PropFlow : Probabilité de 'Random Walk' restreint par un seuil entre 2 nœuds ((Lichtenwalter et al 2010)). Autre métriques comme : L'inverse de distance géodésique, PageRank, Indice d'allocation de ressources.	
Distances entre 2 individus ((Cuvelier & Aufaure 2011)) u, v ayant chacun, $At \in \mathbb{N}$, attributs	Distance euclidienne : $d(u,v) = \sqrt{\sum_{i=1}^{At}(u_i - v_i)^2}$ Manhattan : $d(u,v) = \sum_{i=1}^{At} \|u_i - v_i\|$ Tchebychev: $d(u,v) = \max_{i=1\ldots At} \|u_i - v_i\|$ Cosine: $d(u,v) = \frac{\sum_{i=1}^{At} u_i.v_i}{\sqrt{\sum_{i=1}^{At} {u_i}^2}\sqrt{\sum_{i=1}^{At} {v_i}^2}}$	
Exemple de mesures de similarité ((Cuvelier & Aufaure 2011)) entre 2 ensembles : 2 clusters A, B (2 communautés)	Indice de Jaccard ((Cuvelier & Aufaure 2011)) : $J(A,B) = \frac{\|A \cap B\|}{\|A \cup B\|}$, $0 \leq J(A,B) \leq 1$ Coefficient de Tanimoto ((Cuvelier & Aufaure 2011))	

En effet, une mesure de di-similarité peut être convertie en une mesure de similarité suivant certaines fonctions strictement décroissantes ((Cuvelier & Aufaure 2011)). Les métriques de proximité entre 2 nœuds définies sur ce plan topologique, et même avec des variantes en cas de pondération sont utiles comme attributs. En théorie des graphes, l'indice de Jaccard est une mesure populaire pour estimer le taux de similarité entre 2 clusters (2 communautés) dans un graphe social. Cet indice est originalement basé sur la taille de **chevauchement** entre l'ensemble des voisins d'un nœud u et v (taux de proximité entre les 2). Sa valeur s'annule quand il y n a pas des membres partagés (chevauchement vide), ou atteint le max 1, quand il s'agit du même ensemble et c'est le cas d'une équivalence structurelle en théorie des graphes ((Cuvelier & Aufaure 2011)).

3.4.2. Détection des communautés

La communauté est une structure clé dans un SN là où des sous-ensembles d'individus ayant en effet des caractéristiques similaires, des intérêts communs ou même des proximités géographiques (dans le cas des réseaux sociaux en ligne) se regroupent (Nettleton 2013). La découverte des structures communautaires dans les SNs est un sujet intéressant dans le domaine de sociologie et les sciences comportementales (Gilbert et al 2010). C'est l'un des sujets classiques, le plus étudié et discuté en SNA (Zhou et al 2007) (Missaoui 2014) et suscite l'intérêt de plusieurs chercheurs (Freeman, Zachary, Newman, etc.) depuis nombreuses années sur un plan multidisciplinaire (Nettleton 2013). Le but est de découvrir ces groupes (des sous-graphes densément connectés) d'acteurs qui sont proches en intra-groupe et moins liés en intergroupes (Zhou et al 2007). En résumant les interactions, les avantages de la découverte des communautés sont fondamentalement axés sur la compréhension du système social et aussi ses **phénomènes sous-jacents** (Parthasarathy et al 2011). Le premier qui a montré formellement que les groupes sociaux peuvent être révélés est Homans à travers la représentation matricielle du SN, en cherchant un réalignement des lignes et colonnes approximatifs sous forme de blocs en diagonal (Parthasarathy et al 2011). L'idée de Homans constitue encore un outil basique de structuration des communautés et de visualisation (Parthasarathy et al 2011). Le travail de Rice est l'un des anciens travaux connus dans ce contexte en analysant des communautés individuels sur la base des préjugés politiques et tendances de vote (Parthasarathy et al 2011).

Le problème se formule en cherchant comment identifier une organisation en communautés d'un SN qui est déjà représenté par un graphe. ***Chercher comment identifier une communauté fait appel essentiellement au choix du pattern sous-graphe qui va incarner soit sa définition stricte ou plus étendue****: clique, n-clan, etc.,* ***jusqu'à son concept général (cohésion).***

« Le problème de découverte de communauté a été formulé par plusieurs chercheurs comme un problème similaire au problème de partitionnement de graphe en théorie des graphes» (Zhou et al 2007) ((Bello-Orgazet et al 2016)). Plusieurs techniques de partitionnement ont été proposées pour la détection des structures communautés dans le graphe social, qui varient selon plusieurs critères et dimensions: Partant des algorithmes agglomératifs hiérarchiques (inspirées essentiellement du Clustering), de division vers des algorithmes basés sur des heuristiques. Beaucoup d'algorithmes d'extraction de communautés incluent une composante (fonctions objectives) permettant d'évaluer la qualité de partitionnement résultant, selon laquelle le nombre d'itérations de leurs processus peut être souvent déterminé. Ces fonction objectives sont à optimiser afin d'incarner le concept général du groupe en termes de connectivité intra/inter communautés trouvées en sortie ((Leskovec et al 2010a)). Même s'il y a des fonctions populaires (coupures normalisées et **modularité,** etc.), et comme il n y a pas un accord complet sur le concept de communauté, l'applicabilité de ces mesures de qualité n'est pas universelle car elle ne s'adapte pas avec toutes les situations (Parthasarathy et al 2011).

3.4.2.1. Modularité et autres mesures de qualité de décomposition

❖ **Modularité**

La modularité (Newman & Girvan 2004) est une mesure de référence calculée et une fonction à optimiser dans plusieurs algorithmes de détection des communautés afin de caractériser le concept général d'une communauté en terme de connectivité inter/intra-groupes. La figure suivante présente un exemple de sous-ensemble d'acteurs qui représente une communauté 'D' au sein d'un ensemble 'S'.

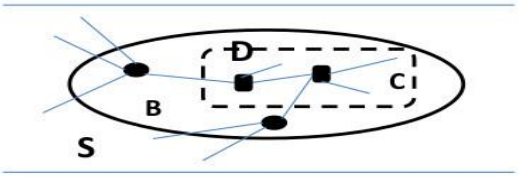


Figure 16. Liens inter/ intracommunautaire – inspiré de ((Chen et al 2009))

Pour une communauté D, la modularité évalue sa qualité comme un rapport en se basant sur le sous-ensemble des liens entre les nœuds de 'C' et entre les nœuds de 'B' et 'S' tel que : $C \subset D, B \subset D, C \cap B = \Phi, D \subset S$ En d'autre terme, 'B' est le sous-ensemble (**la frontière** de 'D') de nœuds qui forment l'une des extrémités de chaque lien intercommunautaire $\{u, v\} \mid u \in D, v \in S$. La version standard d'une mesure de modularité 'Q' évaluant l'exactitude et la qualité de partitionnement d'un graphe divisé en 'K' communautés se formule comme suivant (Nettleton 2013):

$$Q = \sum_{i}^{K} (e_{ii} - {a_i}^2) = Tr_e - ||e^2|| \qquad (27)$$

$e: K \times K$ Est une matrice symétrique, tel que chaque élément e_{ij} est le nombre des liens (intercommunautaires) qui connectent 2 communautés i et j. Donc e_{ii} est le nombre des liens intra-communauté i. La somme des valeurs d'une ligne i est calculée par $a_i = \sum_j^k e_{ij}$ qui représente la fraction des liens qui connectent les nœuds d'une communauté i avec chaque communauté j (Nettleton 2013). La trace de la matrice qui est la somme des valeurs de diagonal, notée $Tr_e = \sum_i^k e_{ii}$ (Nettleton 2013). Formellement, le paramètre de modularité mesure la fraction des liens intra-communautaire au sein des k modules moins le même nombre de liens dans un graphe organisé selon la même partition mais avec des liens aléatoires (Nettleton 2013) (Erétéo 2011) (Parthasarathy et al 2011). En outre, différentes variantes de modularité peuvent être proposées selon différents types et caractéristiques des graphes sociaux (ex. dans les graphes orientés), (Erétéo 2011). La modularité peut être normalisée par le nombre total des liens m (Parthasarathy et al 2011): $Q = \sum_i^K \left(\frac{e_{ii}}{m} - \frac{{a_i}^2}{2m}\right)$

Lorsqu'un partitionnement présente une modularité supérieure, cela veut dire qu'il y a plus de connectivité entre les nœuds au sein de chaque communauté qu'à l'extérieur ((Erétéo 2011)). Donc, ***plus la modularité est supérieure plus la partition est meilleure*** ((Erétéo 2011)).

❖ Kernighan-Lin (KL)

C'est une autre mesure de qualité (une fonction objective) qui cherche à minimiser le nombre des liens (somme des poids des liens) inter-clusters, sous la contrainte d'avoir tous les clusters ayant la même taille (Parthasarathy et al 2011).

$$\text{KLObj} = \sum_{i \neq j}^{k} \text{e}_{\text{ij}} \ , |C_1| = |C_2| = \cdots = |C_k| \qquad (\mathbf{28})$$

❖ Coupure normalisée (Ncut)

Dans un graphe G ayant une matrice d'adjacence A, le Ncut d'un sous-graphe C (groupe) mesure le nombre des liens (la somme des poids) qui relient C avec V\C normalisé par le poids total dans et dans le reste du graphe V\C (Parthasarathy et al 2011).

$$\text{Ncut (C)} = \frac{\sum_{u \in C, v \in V \setminus C} w\{u, v\}}{\sum_{u \in C} degree(u)} + \frac{\sum_{u \in C, v \in V \setminus C} w\{u, v\}}{\sum_{v \in V \setminus C} degree(v)} \qquad (\mathbf{29})$$

La conductance de C (vue précédemment) qui évalue également la qualité de C, est étroitement liée à Ncut. La généralisation de Ncut (conductance) sur un partitionnement de K clusters est donnée par la somme des Ncuts de chaque cluster (Parthasarathy et al 2011).

Tableau 6. Modularité et autres indices populaires pour mesurer la qualité de décomposition

	Modularité	**Ncut (Conductance)**	**Kernighan-Lin (KL)**	**Autres indices**
Qualité meilleure de décomposition	Plus la modularité est supérieure plus la partition est meilleure La modularité est indépendante du nombre de clusters dans une partition	Une valeur inférieure de Ncut reflète des bonnes communautés	Minimiser les liens inter-clusters et maintenir la même taille des clusters	Indice de Silhouette compare la silhouette (des caractéristiques comme le diamètre) des clusters obtenus avec celle de l'ensemble du réseau ((Erétéo 2011)). Indice de Dunn et Indice Davies-Bouldin : comparer la distance intra-cluster et inter-clusters ((Erétéo 2011)).
Complexité	Optimisation de chacune de ces fonctions est NP-difficile (Parthasarathy et al 2011) (Brandes et al 2007).			

La modularité est récemment la mesure populaire : la fonction objective à optimiser, pour estimer la qualité du partitionnement de plusieurs algorithmes de détection de communautés (Parthasarathy et al 2011) (Newman & Girvan 2004).

3.4.2.2. Algorithmes et classification des approches

Les méthodes proposées dans la littérature pour la découverte des communautés jusqu'à nos jours s'améliorent et leur approches varient selon plusieurs critères et dimensions (Parthasarathy et al 2011).

Tableau 7. Approches d'extraction des communautés varient selon des critères & dimensions

Critères \ Dimensions	Catégories & exemples d'algorithmes
Optimisation explicite d'une mesure de qualité	Algorithme (KL) de Kernighan-Lin optimisant KLObj. Algorithmes de division et quelques approches agglomérative, optimisant une modularité. Méthodes spectrales (Ncut)
	Autres algorithmes: Markov Clustering (MCL) & clustering via "shingling" n'optimisent pas une mesure spécifique
Importance accordée à une partition équilibrée	Algorithme (KL) tend vers une partition équilibrée : Clusters de même taille
	Dans le reste des algorithmes l'équilibre est implicitement ou jamais préservé
	Certain algorithme tend vers une partition déséquilibrée : MCL
Contrôle de granularité de décomposition	Clustering agglomératif permet de contrôler la granularité des communautés en sortie. Méthodes spectrales permettent un bi-partitionnement utilisé pour diviser récursivement le réseau ((Parthasarathy et al 2011)).
	Autres algorithmes (ceux qui optimisent une fonction de modularité) ne permettent pas de contrôler le nombre de communautés ((Parthasarathy et al 2011)).
Caractéristiques des performances de scalabilité	Algorithmes récents de clustering proposés pour améliorer les performances de scalabilité du partitionnement des grands réseaux : Partitionnement multi-niveaux : Metis, algorithmes clustering local, MLR-MCL Graclus (Parthasarathy et al 2011) (Dhillon et al 2007).
	les autres approches sont moins performantes (agglomératives et de division)

3.4.2.2.1. Algorithme de Kernighan-Lin(KL)

C'est l'une des premières approches (heuristiques) classiques de partitionnement de graphe L'algorithme est itératif et commence par une bipartition (Parthasarathy et al 2011) (Kernighan & Lin 1970), en cherchant à minimiser les liens inter-cluster à travers une fonction objective (KLObj). A chaque itération, un sous-ensemble de nœuds est cherché dans chaque partie, tel que l'échange de ces nœuds entre les parties réduit les liens inter-cluster (Edge Cut) (Parthasarathy et al 2011). La recherche du sous-ensemble suit une procédure 'greedy', tel que Chaque nœud est associé à un gain qui évalue la réduction de 'Edge Cut' lorsqu'il est déplacé d'une partie à une autre (Parthasarathy et al 2011). L'algorithme selecte le nœud ayant le plus grand gain et met à jour les gains de ses nœuds voisins, avec une complexité de $O(m.\log(m))$ qui a été amélioré après jusqu'à : $O(m)$, (Parthasarathy et al 2011).

3.4.2.2.2. Extraction ascendante ou descendantes des communautés

Généralement, les algorithmes hiérarchiques (Parthasarathy et al 2011) de clustering est un moyen d'extraction de communautés, qui a plus de succès dans la représentation des liens inter/intracommunautaire (Nettleton 2013). Dans ce sens, les auteurs distinguent deux méthodes principales pour extraire des communautés d'un graphe social, en se basant sur la stratégie selon laquelle la hiérarchie ou la granularité (Parthasarathy et al 2011) des clusters (des communautés) se construit depuis et vers la composition atomique (nœuds) du graphe social.

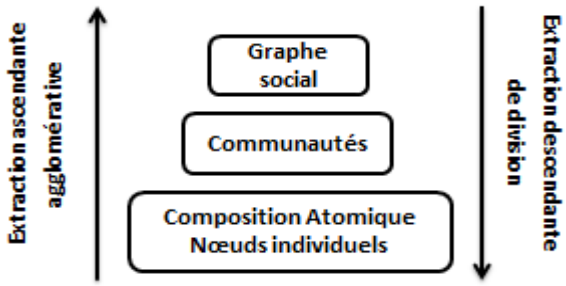


Figure 17. Méthodes d'Extraction des communautés depuis et vers la composition atomique du graphe social

A- Extraction ascendante agglomérative (hiérarchique)

Les algorithmes qui suivent une approche agglomérative regroupent les nœuds dans des communautés de plus en plus grandes et denses suivant un ordre ascendants, en remontant dans la hiérarchie qui semble être un arbre ((Erétéo 2011)). Ce type d'approche commence par les nœuds individuels (les feuilles) comme des **groupes singletons**. Des hiérarchies successives se construisent du bas vers le haut, en utilisant une méthode tel que les dendrogrammes (Nettleton 2013). A chaque étape, des communautés jugées suffisamment similaires sont fusionnées jusqu'à arriver à une dissemblance qui empêche la fusion ou à un nombre souhaité de communautés (Parthasarathy et al 2011) ou encore arriver à la racine qui représente l'ensemble du graphe social. **Comment regrouper ?**

En effet, le principe du clustering agglomératif consiste souvent à regrouper itérativement des individus ayant des attributs les plus similaires (Classification par des mesures de proximité,

similarité et distance). Dans le cas général des SN, 2 nœuds (2 communautés) qui sont proches dans le graphe social sont censés être regroupés dans une même communauté, selon des considérations topologiques, en se basant par exemple sur les chemins qui passent par cette paire de nœuds. Des valeurs sont ainsi attribuées selon des critères sur le nombre de ces chemins ((Erétéo 2011)). En outre, le processus de regroupement de 2 communautés A et B peut se réaliser par l'un des principes suivants (Cuvelier & Aufaure 2011) de distance ou liaison (Linkage) entre les deux.

Tableau 8. Principe de liaison possible entre 2 communautés dans le processus agglomératif

Single linkage	**Complete linkage**	**Average linkage**
$D(A,B) = \min\{d(x,y): x \in A, y \in B\}$	$D(A,B) = \max\{d(x,y): x \in A, y \in B\}$	$D(A,B) = \frac{\sum_{x \in A}\sum_{y \in B} d(x,y)}{\lvert A\rvert \times \lvert B\rvert}$

Tableau 9. Algorithmes agglomératives, exemples et critères de regroupement

Exemples d'Algorithmes Agglomératifs Hiérarchiques	**Plan topologique (nœuds & liens) du graphe social pour mesurer la similarité structurelle**	**Complexité en temps**
Algorithme de Donetti & Munoz ((Erétéo 2011))	Eigenvector des nœuds	$O(n^3)$
Algorithme de Netwalker	Le temps moyen pour atteindre un nœud depuis un autre suivant des marches aléatoires (Random Walk)	$O(n^3)$
Algorithme 'SCAN' pour trouver des communautés qui se chevauchent	Nombre des voisins partagés entre 2 nœuds	-
Algorithme de Newman (Erétéo 2011) ((Parthasarathy et al 2011)) (Newman 2004)	Au moins un lien intercommunautaire pour fusionner 2 communautés. Une fonction de modularité à optimiser	De $O(n^2)$ à $O(n.log^2(n))$ Apres une amélioration $O(m.d.\log(n))$
Clustering hiérarchique (Average-linkage) de ((Newman 2012))	Basé sur la similarité de cousine sur un plan topologique	-

L'exemple du SN du club de karaté de (Zachary 1977), est le graphe social qui a été régulièrement testé comme benchmark de référence par les algorithmes de détection des communautés (Parthasarathy et al 2011). La figure suivante montre le dendrogramme de ce SN, résultant de l'application d'une méthode de clustering hiérarchique (Average-linkage) proposée par (Newman 2012) basée sur la similarité de cousine.

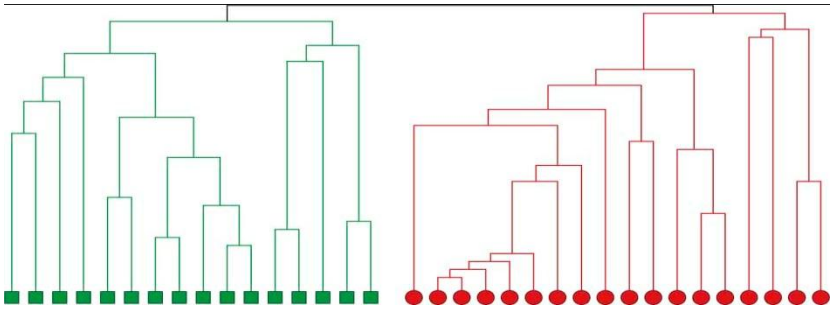


Figure 18. Exemple de décomposition du SN du club de karaté de Zachary selon une méthode agglomérative hiérarchique proposée par ((Newman 2012))

2 communautés principales sont détectées (Newman 2012) qui reflètent exactement la fission ou la rupture de la cohésion sociale du réseau qui s'est produite réellement. Dans un autre algorithme agglomératif optimisant une fonction de modularité comme celui proposé aussi par (Newman 2004), la fusion des groupes dans des communautés plus larges à chaque étape doit marquer une augmentation de modularité du partitionnement résultant (Newman 2004) (Parthasarathy et al 2011). La complexité en temps du processus de fusion est de $O(m)$ car il suffit d'avoir au moins un lien intercommunautaire pour fusionner 2 communautés (Parthasarathy et al 2011) (Newman 2004). En revanche et avec la structure des données (matrice) utilisée pour calculer la modularité, la complexité de l'algorithme peut aller jusqu'à : $O(n^2)$. En introduisant une structure différente (max-heaps), la complexité a été amélioré à $O(m.d.\log(n))$ dans (Clauset et al 2004), tel que d représente la profondeur du dendrogramme (Clauset et al 2004) (Parthasarathy et al 2011).

'NetWalk' est une autre méthode agglomérative hiérarchique, basées sur Random Walks, proposée par ((Zhou & Lipowsky 2004)) en définissant un indice de proximité (similarité) entre 2 nœuds par rapport les autres nœuds. Mais, si la structure communautaire du réseau est fortement connectée, Random Walker prend beaucoup de temps car la densité des liens et le nombre de chemins à suivre est assez important ((Bello-Orgazet et al 2016)).

B- Extraction descendante (de division)

Les approches de division ou de séparation fonctionnent inversement (Parthasarathy et al 2011) en construisant l'arbre suivant un ordre descendant. Une telle extraction commence la division depuis l'ensemble du graphe comme étant une première communauté (Parthasarathy et al 2011) (Nettleton 2013). A chaque étape une communauté est divisée en 2 parties en cherchant à optimiser une fonction objective (généralement la modularité). **Comment diviser ? Quels sont les critères de division ?**

Généralement, les critères de division portent essentiellement sur les liens et l'impact de leur rupture sur le graphe social. En effet, la division peut être précédée par un processus d'évaluation des liens (chaque lien est associé à une valeur). Etant la mesure locale la plus significative et informative sur la cohésion et la résilience du SN, la centralité d'intermédiarité et son alternative appliquée sur les liens 'Edge Betweenness' semble être idéale pour évaluer et distinguer les liens ciblés (à supprimer). C'est le cas dans **l'algorithme de division de** (Newman & Girvan 2004). Par définition, les liens les plus intermédiaires se trouvent souvent sur les chemins géodésiques (Shortest path Betweenness) comme ils constituent probablement des ponts qui relient des communautés différentes. Ainsi, ce sont des liens inter-clusters qui

sont plus intermédiaires que les liens intra-cluster (Parthasarathy et al 2011). Lorsque la division s'effectue au niveau de tels liens (ponts), la connectivité du SN sera sérieusement affectée. C'est l'image d'un retrait ou une **attaque ciblée** dans le réseau. Cependant, Newman propose 2 autres métriques dérivées (Parthasarathy et al 2011) (Nettleton 2013) depuis 'Edge Betweenness' pour évaluer les liens autrement. C'est l'exemple de 'Random Walk Betweenness' (Parthasarathy et al 2011) (Nettleton 2013) qui est en revanche basée sur l'idée que l'intermédiarité d'un lien ne s'évaluent pas toujours à travers les géodésiques. Car selon (Newman 2003), les flux dans un réseau ne suivent pas forcément les plus courts chemins. Donc des chemins aléatoires sont plutôt considérés dans 'Random Walk Betweenness'. En transformant virtuellement le SN en réseau de résistances (Parthasarathy et al 2011), une autre extension 'Current Flow Betweenness' est proposée, basée sur la théorie des circuits (Parthasarathy et al 2011).

Tableau 10. Algorithmes de division, exemples, critères et principes de division

Exemples d'Algorithmes de division	**Critères et principes de division, basés sur l'évaluation des liens**	**Complexité en temps**
Algorithme de ((Newman & Girvan 2004)) : N&G	Edge Betweenness en optimisant la fonction de modularité : Q	$O(m^2.n)$
	Extension dérivée depuis 'Edge Betweenness': (Parthasarathy et al 2011) (Nettleton 2013) 'Shortest path Betweenness' 'Random Walk Betweenness' 'Current-flow Betweenness'	$O(m^2.n.\log(n))$ pour des graphes pondérés Elle peut atteindre $O(n^3)$
Approche de ((Fortunato et al 2004)): variante de l'algorithme de (N & G).	Optimisation de la qualité de la partition résultante depuis l'algorithme (N & G), ((Fortunato et al 2004)) ((Erétéo 2011))	$O(m^3.n)$
Approche de ((Bothorel & Bouklit 2008)).	Adaptation de de l'algorithme de (N & G) sur les hypergraphes ((Erétéo 2011))	-
Approche ((Martínez Arqué & Nettleton 2012)).	-	Réduire partiellement le coût de calcul de ((Nettleton 2013)) ((Martínez Arqué & Nettleton 2012))
Algorithme de ((Blondel et al 2008)) **'Louvain Method'.**	**L'optimisation la plus efficace de l'algorithme de** (N & G)**,** ((Nettleton 2013))	**Complexité linéaire** $\boldsymbol{O(m)}$

Algorithme de Radicchi ((Erétéo 2011)).	Coefficient de clustering de lien ((Erétéo 2011)). Le lien ayant le plus faible coefficient est supprimé à chaque itération ((Erétéo 2011)).	-

Algorithme de 'Newman & Girvan' et de Blondel 'Louvain Method'

Ce sont les 2 algorithmes de division les plus populaires pour extraire des communautés. D'abord l'algorithme de (Newman & Girvan 2004), (Parthasarathy et al 2011), a été proposé le premier et consiste à identifier les groupes par leurs frontières qui sont en effet les liens les plus centraux (les plus intermédiaires) retirés ou divisés ce qui conduit à débrancher ces communautés (Nettleton 2013) (Parthasarathy et al 2011). Le résumé de l'algorithme dans sa forme général fait appel à 2 autres processus l'un pour le retrait des liens (processus de division (Algorithme 1)) et l'autre pour calculer la modularité (Nettleton 2013) (Parthasarathy et al 2011) (Newman & Girvan 2004):

Modularité _ Q :

(1) Générer la matrice e: k × k des communautés: k le nombre courant des composantes obtenues
\\ Calculer la modularité selon l'équation de Q
(2) Calculer la trace Tr_e
(3) Calculer la somme de la matrice $a_i = \sum_j^k e_{ij}$
(4) Calculer Q.

Algorithme 1. Aperçu sur le pseudo-code, la forme général de l'algorithme de ((Newman & Girvan 2004)), ((Nettleton 2013)) ((Parthasarathy et al 2011))

```
Recherche des communautés selon l'algorithme de Newman & Girvan
TANT QUE (Q ≥ 0) ou (Nombre d'itérations < Nombre d'itérations prédéfini ) FAIRE
        (1) Sélectionner une composante_réseau : Initialement, la plus grande composante (le réseau lui-même).
        (2) Calculer Retrait _lien_ Edge _Betweenness (compostante_réseau) jusqu'à avoir une division
        (3) Sauvegarder les 2 nouvelles composantes.
        (4 )Calculer (Modularité _ Q).
Si le processus itératif est terminé alors faire l'affectation des nœuds du réseau aux communautés.

Procédure: Retrait _lien_ Edge _Betweenness (compostante_réseau)
(1) Calculer les scores d'intermédiarité (Shortest path Betweenness) de tous les liens
(2)
TANT QUE (Composante_réseau n'est pas divisée en 2) FAIRE
        (2-1) Chercher le lien edge_h avec le plus grand score (le plus intermédiaire)
        (2-2) Retirer edge_h
        (2-3) Recalculer les scores des liens restants sur composante_réseau affectée.
```

Dans cet algorithme, l'extraction des communautés est un processus itératif, guidé par l'optimisation d'une modularité Q dont l'intervalle empirique habituelle $0.3 \leq Q \leq 0.7$ (Nettleton 2013) (Newman & Girvan 2004). A chaque itération et après le retrait d'un lien, la

modularité sera calculée quand il y a 2 nouvelles composantes sont résultantes depuis la division (Nettleton 2013) (Newman & Girvan 2004). L'algorithme s'arrête quand un optimum ou un nombre prédéfini d'itérations est atteint (Nettleton 2013) (Newman & Girvan 2004). Dans le cas le plus simple, recalculer notamment l'intermédiarité après chaque retrait de lien, coûte lui-même $O(n.m)$ et constitue l'étape la plus critique dans l'algorithme. En plus, le cout de l'algorithme entièrement, peut s'exploser considérablement et atteindre $O(n^3)$ (Parthasarathy et al 2011). Ainsi, le problème majeur (l'inconvénient principal) de cet algorithme est le coût de calcul en augmentation devant le nombre important de liens. Il est donc praticable sur des petits réseaux ((Erétéo 2011)). Cet algorithme a été appliqué sur le graphe social modélisant le réseau de co-apparence des personnages dans le roman 'les Misérables' (Newman & Girvan 2004).

Cependant un réseau analysé peut être clairement divisé après la première itération s'il est déjà clairsemé, dispersés (Nettleton 2013). Un prétraitement d'échantillonnage ou une résiliation prématurée de l'algorithme (comme c'est proposé par Martínez-Arqué Nettleton) peut régler partiellement le problème du coût (Nettleton 2013) (Martínez Arqué & Nettleton 2012).

L'approche '**Louvain Method**' proposée par (Blondel et al 2008) quatre ans plus tard constitue l'optimisation la plus efficace de l'algorithme précédent en termes de cout de calcul (Nettleton 2013). Cette méthode se base sur l'agrégation qui permet de manipuler une communauté comme un nœud en optimisant localement sa modularité (Nettleton 2013). Elle s'exécute itérativement en 2 étapes: Des communautés plus petites sont cherchées et un nouveau réseau est construit en agrégeant les nœuds de chaque communauté en un seul nœud (Blondel et al 2008) (Nettleton 2013). L'optimisation consiste à maximiser la modularité, en évaluant s'il y a de changement de modularité d'une communauté C_x résultant d'un déplacement possible d'un nœud de C_x vers une autre adjacente C_y (Blondel et al 2008) (Nettleton 2013). A cause de ces agrégations, le nombre de teste ainsi que le coût de calcul se réduit rapidement à chaque itération (Nettleton 2013). Par conséquent, ***Louvain Method est maintenant la plus efficace et est devenue une norme industrielle*** (Nettleton 2013).

Tableau 11. Comparatif : Approches agglomératives et de division

	Approches agglomératives	**Approches de division**
Points communs	L'arbre et la relation père-fils : Une communauté représentée par un sommet père est obtenue depuis / divisée en des communautés fils. La modularité peut être la fonction à optimiser dans les 2 catégories	
	Clustering agglomératif donne la main pour contrôler la granularité des communautés en sortie suivant certains paramètres Avec les principes de liaisons, les nœuds peuvent ne pas être	Algorithmes de divisions optimisant une fonction objective (Modularité, KL, etc.) ne permet pas de contrôler le nombre communautés. Deux considérations sont

Avantages & Inconvénients	correctement classés ((Cuvelier & Aufaure 2011)) Tendance de perdre les liens inter-communautés ((Nettleton 2013)). Des membres périphériques les plus isolés de leur communauté sont exclus ((Erétéo 2011)).	discutables: Le choix des nœuds initiaux est une difficulté ((Nettleton 2013)). La résilience du graphe (l'impact de suppression sur le graphe).
	Leur simplicité Mais moins performants et ne supportent pas les réseaux à grande échelle. La pluparts s'appliquent sur des graphes non-pondérés	

3.4.2.2.3. Méthodes spectrales

Souvent proposées pour l'analyse des graphes et matrices, les méthodes de pertinemment spectrales sont utilisées également pour détecter des structures (des groupes) dans les réseaux du monde réel (réseaux de co-auteurs) ainsi que dans les bases de données (Nettleton 2013). Avec ces approches, les nœuds sont affectés aux clusters en se basant sur les 'Eigenvectors' des matrices : Matrice Laplacienne $\mathcal{L}$ matrice d'adjacence A et matrice diagonale contenant le degré des nœuds D (Nettleton 2013) (Parthasarathy et al 2011). L'idée générale consiste à intégrer les nœuds du réseau dans un espace de dimension k et par la suite les assigner aux clusters à travers une technique classique de classification (K-means), (Parthasarathy et al 2011) (Von Luxburg 2007). Par exemple, un vecteur $X = \{X_1, \dots, X_k\}$ est associé à chaque nœud pour indiquer son appartenance à une communauté par 0 ou 1 (Zhou et al 2007). D'autre part, le clustering spectral peut être une solution pour les problèmes de décompositions des graphes pondérés (en optimisant une fonction Ncut), (Parthasarathy et al 2011) (Buffa 2008). Cependant, l'inconvénient majeur est pratiquement la complexité de calcul qui ne permet pas à ces algorithmes de se mettre à l'échelle des grands réseaux (contenant des dizaines de milliers de nœuds), (Parthasarathy et al 2011). Sachant qu'un algorithme comme le k-means est plus rapide par rapport le calcul des Eigenvectors (Parthasarathy et al 2011), des solutions s'orientent récemment, vers des approches parallèles : des **algorithmes multi-niveaux** comme 'Graclus' (Dhillon et al 2007) qui propose une qualité de décomposition comparable à celle du clustering spectral (Parthasarathy et al 2011) (Dhillon et al 2007).

3.4.2.3. Bilan, Heuristiques et Tendances

Malgré l'ambiguïté qui entoure la définition du concept de communauté, différentes techniques sont proposées et utilisées pour détecter des communautés. Quelques soit la méthode, la connectivité et après la densité sont typiquement les points communs dans les clusters résultants. Un sous-ensemble de nœuds forme un bon cluster (bonne communauté) si le sous-graphe induit est dense avec moins de liens intercommunautaires.

La plupart des approches de découverte de communautés et de partitionnement de graphe en général, abordées ci-dessus, sont applicables souvent sur des modèles de graphes sociaux simples: non-orientés, non-étiquetés, non-pondérés. Pourtant, l'orientation ou le typage par exemple des liens peut enrichir et diversifier les résultats. Parfois, le partitionnement résultant n'est pas totalement recouvrant, avec des clusters qui ne couvrent pas tous les acteurs du réseau. D'autre part, parmi les approches les plus intéressantes sont celles qui profitent du bénéfice informationnel du concept d'intermédiarité. Cependant, la complexité de calcul est un inconvénient majeur qui laisse ces algorithmes s'exécutent efficacement sur des réseaux d'ordre 10^4 à 10^6 de nœuds ((Erétéo 2011)), et pose en effet un problème de scalabilité sur les réseaux à grande échelle (**SNs émergents en ligne**). Noter aussi que selon la représentation d'une structure communautaire en sortie, la plupart des approches existantes procèdent par une décomposition du SN en communautés disjointes. Un nœud (un acteur) est souvent attribué à une seule communauté. Alors qu'en effet, l'acteur dans la plupart des réseaux réels, est susceptible d'être affilié à un ou plusieurs groupes sociaux en même temps, en exprimant même des degrés différents d'implication : Des communautés qui se chevauchent. Différents algorithmes qui sont testés sur différents réseaux en entrée, en optimisant différentes fonction objectives ce qui ramène à afficher des communautés de caractéristiques différentes en sortie ((Leskovec et al 2010a)). En conséquence, le choix d'un algorithme doit tenir en compte ces caractéristiques ainsi que les performances et les contraintes devant le réseau analysé ((Leskovec et al 2010a)).

3.4.2.3.1. Heuristiques

Certaines approches de détections des communautés sont classées comme des heuristiques et jouent sur les caractéristiques structurelles d'une communauté ((Erétéo 2011)) afin de résoudre un des problèmes de partitionnement rencontrés ou adopter une des hypothèses supplémentaires :

Des communautés qui se chevauchent : Des algorithmes comme par exemple 'Fast unfolding of communities in large networks' (Blondel et al 2008) ou bien FOCAL : 'Fast Overlapping Clustering ALgorithm' proposé dans ((Magdon-Ismail & Purnell 2011)) prennent en compte le problème de recouvrement ou de chevauchement ((Chuan Shi et al 2013)) ((Gregory 2009)). Dans ce sens, des auteurs proposent des structures notées ((α, β)-communities) pour montrer des communautés qui se chevauchent ((Wang et al 2013)). Aujourd'hui, il y a une tendance pour concevoir des algorithmes efficaces exclusivement nommés par des acronymes comme CODA, CESNA, BIGCLAM, et CONGA ((Gregory 2007)) pour détecter des communautés qui se chevauchent. CONGA [2] 'Cluster Overlap Newman Girvan Algorithm' est une variante de l'algorithme de Newman & Girvan, proposée par ((Gregory 2007)) qui introduit une hypothèse supplémentaire basée sur le concept 'Split

Betweenness'. Il s'agit de diviser imaginairement un nœud en 2, au lieu de retirer un lien (Si 'Split Betweenness' est supérieure du maximum 'Edge Betweenness'). Après, calculer 'Split Betweenness' qui est le 'Edge Betweenness' du lien imaginaire entre les 2 nouveaux nœuds ((Gregory 2007)). Par conséquent, le nœud original peut être attribué à plusieurs communautés. On cite aussi SLPA ((Xie & Szymanski 2012)) dans la détection de communautés qui se chevauchent, une variante de l'algorithme LPA, ou encore le k-means adapté par ((Bello-Orgaz et al 2012), Adaptive k-means algorithm for overlapped graph clustering).

((Xie et al 2013)) ont présenté un état de l'art et une étude comparative entre ces différents algorithmes appliqués sur des réseaux du monde réel. La densité et la diversité des chevauchements, très importantes dans ces réseaux laisse le problème de détection des communautés qui se chevauchent, pas tout à fait résolu ((Xie et al 2013)). Une caractéristique commune a été remarquée également dans le faible taux de chevauchement (moins de 30%) pour les nœuds, En réalité, chacun n'appartient qu'à 2 ou 3 communautés ((Xie et al 2013)).

Détection de communautés basée sur la propagation des étiquettes : L'algorithme de propagation étiquettes RAK ((Raghavan et al 2007)) ou LPA 'Label propagation Algorithm' ((Xie & Szymanski 2011)) est un algorithme itératif efficace pour la détection des communautés qui ne demande pas des informations préalables comme le nombre, taille de communautés, des nœuds centroides, etc. Il est juste basé sur une seule initialisation d'étiquettes pour les faire propager itérativement :

(1) Assigner une étiquette aléatoire unique à chaque nœud.
(2) Propagation qui consiste à remplacera l'étiquette de chaque nœud par l'étiquette la plus utilisée par ses voisins, si sa propre étiquette est différente. Dans le cas où plusieurs étiquettes sont les plus utilisés, le choix se fait aléatoirement.
(3) Tant qu'au moins un nœud change d'étiquette, allez à l'étape (2)
(4) Sinon les nœuds qui partagent la même étiquette vont former une communauté.

Vue sa complexité approximativement en temps linaire ((NGUYEN et al 2013)), LPA peut s'appliquer sur des réseaux à grande échelle. Cependant, l'instabilité du résultat est l'inconvénient majeur de cette stratégie de décomposition basée sur des étiquettes aléatoires, car elle produit à chaque exécution une partition différente. Une instabilité très indésirable en pratique, car elle peut agrandir par exemple l'espace de recherche quand on étudie la dynamique des communautés. ((Xie & Szymanski 2013)) ont proposé LabelRank, une version améliorée, plus stable de LPA et une extension de MCL (voir plus loin), en introduisant des opérateurs pour contrôler et stabiliser la dynamique de propagation à n'importe quelle exécution sur le même réseau.
D'autre part, LPA a été également améliorée devant la multiplicité des usages sur le web social, en remplaçant les étiquettes aléatoires par des tags, ce qui donne plus de signification (une dimension sémantique) aux communautés détectés. Il s'agit de l'algorithme 'SemTagP' ((Erétéo et al 2011)), (voir plus loin).

Etant la mesure de qualité la plus populaire, la modularité suscite également l'intérêt de plusieurs approches variantes comme celle de Chen ((Erétéo 2011)) qui a pour but de maximiser les liens intra-cluster et minimiser les liens inter-clusters, ou même adapter sa

définition par exemple sur des graphes bipartites avec Barber ((Erétéo 2011)). Dans ce sens, ((Yongcheng Xu et al 2013)) ont adapté l'optimisation par colonie de fourmis pour la détection de communautés dans un réseau Biparti comme 'Corporate interlocks in Scotland (1904-5)' ((Batagelj & Mrvar 2006)) ((De Nooy et al 2004)) ((Yongcheng Xu et al 2013)).

Tableau 12. Des approches supplémentaires : Heuristiques

Heuristiques	**Description**	**Complexité**
Algorithme de 'Wu' ((Erétéo 2011)).	Basé sur la similarité entre SNs et la circulation en réseaux d'électricité ((Erétéo 2011)).	En temps linéaire $O(m+n)$
Algorithmes basés sur 'Random Walks'	Markov Clustering (MCL) ((Erétéo 2011)) ((Parthasarathy et al 2011))	Multiplications des matrices est très consolateurs en temps dans les premières itérations
	Algorithme proposé par Pons et al. ((Erétéo 2011))	Le plus efficace en temps $O(n^2.\log(n))$ Le plus couteux en espace $O(n^2)$
Approche de Djidev ((Erétéo 2011))	Réduire le problème d'optimisation de modularité dans un graphe pondéré à un problème de 'Weighted Min-Cut' ((Erétéo 2011))	$O(n.\log(n)+m)$
Algorithme de propagation d'étiquette (RAK), ((Raghavan et al 2007)) Et sa variante pour détecter des groupes qui se chevauchent ((Gregory 2009))	Un processus itératif permet initialement d'attribuer un label aléatoire unique à chaque nœud. La propagation consiste à remplace son étiquette par le label le plus utilisé par ses voisins. Communauté détectée regroupe les nœuds partageant un label unique ((Raghavan et al 2007)).	Presque en temps linéaire
Méthodes multi-niveaux (ML) de partitionnement	ML Spectral	$O(m)$
	ML – KL	$O(m)$
	Metis: qui optimise la fonction objective KLObj ((Parthasarathy et al 2011)). Graclus : qui optimise Ncut (Dhillon et al 2007).	-

	MLR-MCL ((Satuluri et al 2010)) ((Parthasarathy et al 2011))	

MCL est l'approche la plus populaire dans les algorithmes basés sur 'Random Walks'. Le processus MCL décompose le graphe en exécutant itérativement 2 opérations « Expand/ inflate » en alternance (Parthasarathy et al 2011). Ces opérations (de multiplications) s'appliquent sur la matrice stochastique ou de probabilités de transition entre 2 nœuds, désignées également comme flux stochastiques (Parthasarathy et al 2011). Pendant les itérations, le processus permet d'accroitre et renforcer les flux stochastiques avec plus de chemins intra-cluster et affaiblir les flux inter-clusters (Parthasarathy et al 2011). En plus de la complexité en temps de calcul, MCL conduit à des regroupements déséquilibrés : beaucoup de groupes singletons (ou contenant 2 ou 3 nœuds) ou un très grand cluster (Parthasarathy et al 2011). Récemment, **des approches multi-niveaux** plus rapide offrent des performances de haute qualité de partitionnement (Parthasarathy et al 2011). Elles se basent sur 3 stratégies essentielles. Il y a le 'Shrinking/ Coarsening' qui utilise le '**matching'** (Les 2 extrémités d'un lien sont effondrés dans un seul nœud) visant à réduire le graphe afin d'être partitionné plus rapidement (Parthasarathy et al 2011). Ensuite, l'étape de partitionnement qui peut se réaliser par une méthode spectral sur le petit graphe. Enfin, la phase de 'Uncoarsening' qui permet de raffiner la partition jusqu'à arriver au graphe original (Parthasarathy et al 2011). Par exemple, le MLR-MCL (Multi-level Regularized MCL) est une variante de MCL, proposée pour résoudre ses 2 problèmes précédents (scalabilité et déséquilibre), ((Satuluri et al 2010)). L'optimisation dans les heuristiques de raffinement multi-niveaux est plus complexe que le clustering à base de modularité, car le nombre de cluster n'est pas connu à l'avance.

Les catégories précédentes couvrent beaucoup de méthodes de partitionnement de graphe de la littérature, utilisées notamment pour la découverte des communautés dans les SNs: Partant des plus anciennes (algorithme de KL) passant par les méthodes hiérarchiques, spectrales, et le partitionnement multi-niveaux récemment, etc., ces méthodes ont différentes stratégies : maximisation de modularité, Random Walks, etc. ***Les algorithmes populaires les plus performants qui s'exécutent presque en temps linéaire*** sont : 'ML-KL', 'ML- spectral' et parmi les plus connus est celui de (Newman & Girvan 2004) et 'Louvain Method' ((Blondel et al 2008)). La modularité semble être la meilleure mesure de qualité de décomposition, mais son calcule et sa maximisation est un problème NP-complet ((Bello-Orgazet et al 2016)). Mais des méthodes ('greedy techniques') ascendantes agglomératives (hiérarchiques) comme celle de (Newman 2012) ont donnée des bonnes approximations de modularité.

Cependant, les algorithmes d'extraction de communautés (de division ou autre) commencent à se distinguer du clustering classique en se focalisant beaucoup plus sur les caractéristiques et la richesse particulières des SNs (en ligne). Par exemple le problème de partitionnement des graphes complets qui est un problème NP-complet ne se pose pas trop puisque les caractéristiques de la majorité des SNs n'affichent pas cette configuration.

3.4.2.3.2. Autres tendances

L'étude des communautés reste un sujet d'actualité 'hot topics' dans SNA partant des SNs du mode réel, synthétiques, de proximité physique, vers les données sociales sur le web : hyperliens, e-commerce, SNs en ligne, etc., (Cyber communities). C'est une piste potentialisée par la disponibilité croissante des données sociales, et les motivations commerciales (Nettleton 2013). ***Des nouvelles approches adoptent des tendances récentes pour améliorer la détection des communautés sur le web ainsi que sur des modèles de graphes sociaux (en ligne) plus riches*** : 'Local Graph Clustering', 'Flow-Based Post-Processing', une découverte des communautés applicable par exemple sur des graphes orientés, biparties (via 'Shingling'), (Parthasarathy et al 2011), sur des SNs hétérogènes (Zhou et al 2007) ainsi que l'étude de **l'évolution des communautés dans les SNs dynamiques**, etc. La prolifération des média sociaux offre plus de la richesse informationnelle au niveau des relations et du contenu comprenant : documents, images, et même des localisations géographiques, etc., (Nettleton 2013). Ainsi, l'extraction des groupes tend récemment d'aller au-delà de ses définitions topologiques en cherchant une signification sémantiquement cohérente (Parthasarathy et al 2011).

Détections des communautés dans les SNs hétérogènes

L'analyse d'un SN hétérogène est un défi récent qui doit supporter formellement l'hétérogénéité des liens ou des sommets. La découverte des communautés dans un SN hétérogène incluant par exemple des documents est l'une des tendances récentes qui s'incarne par des approches basées soit sur le contenu ou sur le partitionnement de graphe bipartite, tripartite, etc., (Zhou et al 2007). L'exemple dans (Zhou et al 2007) formule un SN qui est composé par des auteurs, des mots, et les lieux de publication, par un graphe tripartite G_{XYZ} ; Le partitionnement se fait suivant une approche spectrale qui minimise une fonction coût généralisée sur un graphe tripartie à travers 2 fonctions coûts de partitionnement de graphes bipartis G_{XY} et G_{YZ} ; Cette optimisation est montrée comme un problème NP-difficile (Zhou et al 2007). Il existe autres approches proposées mais limitées par certaines restrictions. Comme l'algorithme de clustering 'NetClus' ((Sun et al 2009)) applicable sur un réseau en étoile ou l'algorithme 'RankClus' (Parthasarathy et al 2011) qui est uniquement destiné pour traiter des réseaux ayant 2 types de nœuds (Parthasarathy et al 2011). On trouve aussi l'outil C-GROUP appliqué sur un SN hétérogène (auteurs, articles, conférence), (Kang et al 2007) et qui se base sur un mécanisme (une sémantique) de regroupement (Kang et al 2007).

Métriques pour les groupes

Des métriques pour les groupes sont aussi proposées pour évaluer par exemple le degré d'appartenance d'un acteur à un groupe, la centralité d'un membre dans sa communauté et l'influence d'une communauté (Nettleton 2013) en introduisant même le concept de centralité de groupe (Everett & Borgatti 1999) ((Everett & Borgatti 2004)). Pour mesurer la centralité de groupe, certains auteurs comme (Everett & Borgatti 1999) ((Everett & Borgatti 2004)), proposent des méthodes de généralisation pour faire étendre les métriques (de centralité) individuelles connues sur les groupes. Ils ont testé leurs propositions sur des datasets tel que : Des données collectées par Linda Wolfe qui enregistre 3 mois d'interactions dans une troupe de 20 singes (l'interaction est définie par la présence conjointe de 2 singes au bord de la

rivière d'Ocala à Floride), (Everett & Borgatti 1999). Des groupes sont formés par sexe et par âge. L'une des considérations se concentre sur le positionnement d'un groupe pour estimer sa centralité par rapport aux nœuds externes (Everett & Borgatti 1999). En revanche, si toute une configuration de réseau en groupes a été considérée, des tests sur certains nœuds externes auraient pu être évités. Car parmi ces nœuds ceux qui peuvent être affiliés également à un autre groupe. En plus, les éléments qui forment la frontière d'un groupe, ne sont pas vraiment valorisés en tant qu'une source fournissant des nouvelles informations à leur collectivité où le partage sera plus rapide après à l'intérieur. On trouve aussi que ((Gil-Mendieta et al 1997)) ont calculé la centralité des groupes étant des cliques.

D'autre part, la normalisation des centralités de groupes, s'est montrée significative dans (Everett & Borgatti 1999) par rapport au cas individuel. Par exemple le groupe des males était plus central que celui des femelles mais la situation s'est inversée après la normalisation. Les groupes qui sont en effet larges, ayant moins de liens externes peuvent avoir une centralité normalisée plus élevée. Devant toutes ces considérations, les métriques de centralité de groupes semblent être très informatives avec un énorme potentiel pour chercher des groupes centraux et expliquer des phénomènes liés à l'efficacité et le succès des groupes (Everett & Borgatti 1999).

3.5. Conclusion partielle

Même dans ce premier contexte statique topologique, les graphes analysés qui modélisent des SNs se distinguent par des propriétés inhérentes. Les parties précédentes ont abordé 2 pistes principales en SNA classique portant sur une analyse locale (individuelle) et sur l'organisation globale du SN, à travers une diversité d'études analytiques comprenant les principales métriques, algorithmes, etc. Beaucoup de techniques proposées s'appuient sur des fondements de la théorie des graphes et s'appliquent sur des représentations de graphes sociaux.

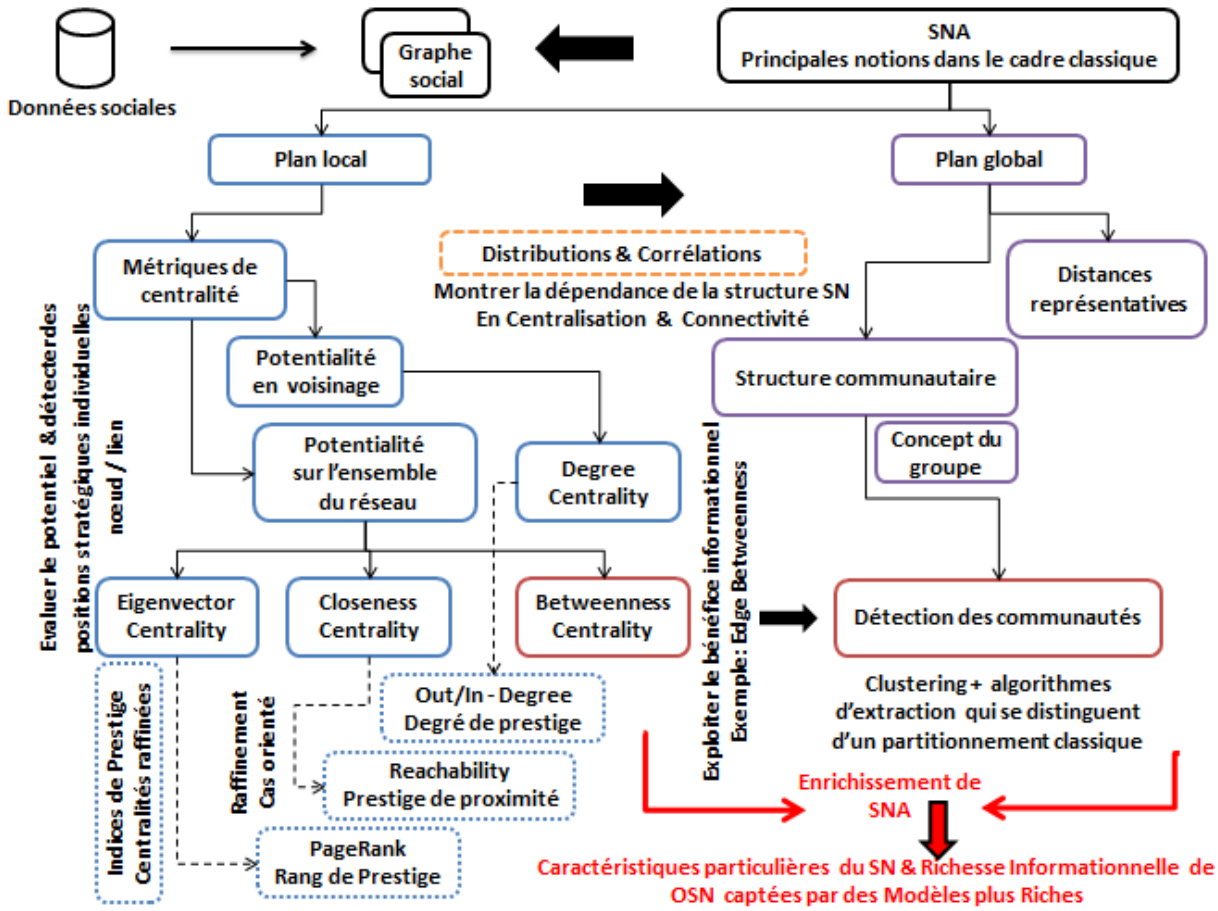


Figure 19. Les deux plans d'analyse des SNs, dans son contexte classique

Devant leur bénéfice informationnel et les motivations récentes, une grande partie a été réservée pour les concepts de centralités notamment d'intermédiarité ainsi que l'organisation du réseau en structure communautaire. La recherche d'une meilleure qualité de performances de ces métriques et algorithmes, tend vers dépasser les premiers SNs testés, en s'appliquant sur des jeux de données plus grands notamment avec l'émergence des SNs sur le web. Cependant, ces techniques d'analyse sont souvent proposées dans un cadre classique, statique et topologique. La pondération, l'orientation, l'étiquetage des liens et même des nœuds avec attributs sont des propriétés à ajouter en perspectives vis-à-vis l'augmentation de la complexité de calcul. Aujourd'hui, les besoins d'analyse sont évolués et ne se contentent pas d'améliorer les performances de calcul, mais en requérant également plus d'enrichissement et développement de modèles, algorithmes et d'outils puissants plus riches.

4. Jeux de données (Données sociales)

Bien que, beaucoup de techniques et méthodes se sont développées pour analyser des SNs, la recherche était limitée dans des petits ensembles de données collectées par les sociologues et en science comportemental (Kazienko et al 2011). Des techniques comme : Texte Mining, des questionnaires, etc. ont été utilisées pour extraire des réseaux bien connus en SNA: Réseaux de collaboration scientifique, réseaux film-acteurs, réseaux d'amitié entre élèves, réseaux de contacts sexuels, proximité physique, marché d'emploi, santé publique, de la psychologie, des données collectées depuis des sondages et questionnaires, etc. (Kazienko et al 2011). L'évaluation des performances de ces algorithmes a été l'un des premiers besoins qui cherchent des réseaux plus

grands. Par exemple, dans une petite société high-tech qui vend, installe et maintien des systèmes informatiques, des liens d'amitié entre employés ont été recueillis par des questionnaires pour construire un réseau d'amitié 'Friendship and unionization in a hi-tech firm' ((De Nooy et al 2004)) ((Krackhardt's 1999)). Qui considérez-vous comme un ami personnel ? La plupart des nominations des amis sont réciproques : un lien est créé si les deux personnes impliquées se connaissent ((De Nooy et al 2004)) ((Krackhardt's 1999)).

4.1. Données synthétiques

La génération des données synthétiques semblait être une solution pour fournir des graphes de grande taille qui ne reflètent pas forcément des SNs du monde réel. Ces données synthétiques sont artificiellement générées par différents modèles selon les contextes et les objectifs spécifiques des chercheurs (Nettleton 2013). Toutefois, l'usage de ces datasets est limité, exploité juste pour comparer et évaluer les performances de certains algorithmes sans pouvoir tester vraiment leur efficacité réelle. En outre, la perte du bénéfice informationnel de ces techniques d'analyse est un autre inconvénient de ces données. Souvent, les propriétés inhérentes d'un SN ne se présentent pas et c'est le cas avec une génération de graphes aléatoires, même si les générateurs se sont améliorés après, en essayant de reproduire ces propriétés. Par exemple le modèle de Barabasi et Albert ((Erétéo 2011)) produit des graphes aléatoires dont les degrés sont distribués selon la loi de puissance en se basant sur la notion d'attachement préférentiel. Un autre exemple de Watts et Strogatz reproduit l'effet du petit monde. La qualité du générateur peut être évaluée par exemple à travers la fréquence des occurrences des structures significatives comme les triades, etc. Il y a aussi des générateurs basés sur des propriétés basiques du mécanisme évolutionnaire du SN.

4.2. Données sociales sur le web

Les premiers services de réseautage, tels que Usenet, ARPANET et BBS 'America Online and CompuServe' (Nettleton 2013) ont affiché des caractéristiques rudimentaires de SNs en ligne. Après le lancement de l'internet qui a été rapidement accompagné par la création du web (WWW) par 'Tim Berners Lee' en 1994 (Nettleton 2013) ((Buffa 2008)), des portes sont ouvertes aux individus et organisations pour se socialiser en dépassant les frontières géographiques et politiques. Par conséquent, le web est maintenant la meilleure source pour extraire facilement des données sociales à analyser, plus grandes et reflètent de plus en plus les interactions du monde réel. Au départ, sa structuration avec les hyperliens était exploitée pour fournir régulièrement des réseaux de grande taille, permettant d'évaluer la qualité des performances. Des graphes de pages web liées par des hyperliens, sont extraits et analysés avec les techniques de SNA : ex. le calcule de PageRank par ((Kamvar 2003)), mais sans pouvoir présenter des vraies entités et interactions sociales. Malgré ça, certaines caractéristiques (distributions en loi de puissance, le phénomène du petit monde, etc.) des SNs et des propriétés de développement dynamique ont été explorées ((Watts & Strogatz 1998)) ((Carlson & Doyle 1999)) ((Albert & Barabasi 2002)) ((Huberman & Adamic 1999)) ((Watts et al 2002)) sur des graphes web : Ex. Les graphes web de Kleinberg[79], les graphes web de Stanford de Sepandar D. Kamvar[80] ((Kamvar 2003)). ***Cependant la pertinence de ces données doit être en mesure de répondre aux objectifs de SNA et ses tendances pour fournir un bénéfice informationnel plus significatif***.

4.2.1. Web Mining pour extraire les premiers SNs sur le web

La masse volumineuse des données en général sur le web a attiré l'attention des chercheurs en faisant appel à des techniques de fouilles de données pour gérer cette masse et extraire des connaissances. Dans ce sens, le Web Mining s'est installé comme une discipline dérivée du Data Mining et destinée pour tirer des informations pertinentes et découvrir des connaissances depuis les données du web. Cette discipline a été bien placée pour répondre aux intérêts des sociologues visant à extraire des SNs implicites sur le web comme un cas d'application du Web Mining. Selon ((Mika 2005a)), les pages (homepages) personnelles ainsi que les cooccurrences des noms dans les pages constituent une source pour inférer par certaines techniques du web Mining, des SNs implicites sur le web.

L'inférence à partir des pages web personnelles

C'est l'exemple d'un réseau d'amis, extrait par ((Adamic & Adar 2003)) depuis les pages personnelles des étudiants à l'université de Stanford et de MIT. Chacun des étudiants met des hyperliens dans sa page personnelle vers les pages de ses amis. Contrairement aux graphes de pages web précédents, les hyperliens dans ce cas, modélisent une interaction sociale (relation d'amitié). Ainsi, ce graphe a été démontré qu'il présente des propriétés d'un SN : l'effet du petit monde, distribution selon la loi de puissance avec un taux de clustering important ((Adamic & Adar 2003)) ((Erétéo 2011)).

L'inférence à partir la cooccurrence des noms sur les pages web

L'évaluation de similarité ou encore la cooccurrence ((Tauveron 2012)) entre les pages personnelles est une autre solution, même si les relations ne se sont pas déclarées par les hyperliens comme dans le cas précédent. Des relations peuvent être déduites depuis la présence simultanée de 2 ou plusieurs mots utilisés dans le même énoncé dans ces pages personnelles ((Adamic & Adar 2003)) ((Tauveron 2012)). En effet, la cooccurrence des éléments textuels est originalement utile pour extraire des informations précieuses sur la connectivité. Par exemple on trouve le réseau '*The Reuters terror news network*' de l'agence de presse britannique Reuters construit à partir toutes les histoires et les versions réalisées par les membres de la maison blanche (Bush team) au sujet des attentats du 11 Septembre aux Etats-Unis ((Batagelj & Mrvar 2006)) ((Corman et al 2002)) ((Batagelj & Mrvar 2003b)) ((Johnson & Krempel 2004)). Les nœuds du réseau sont des mots : 13 332 mots ou termes. Un lien (une arête) est créé entre deux mots s'ils apparaissent dans la phrase : 243 447 liens pondérés par la fréquence). Ces données présentent un défi analytique intéressant, à partir de lequel ((Johnson & Krempel 2004)) essayent par exemple d'analyser et visualiser ce réseau dans le but d'extraire des connaissances sur les conflits (Opinions divergentes de 'Colin Powell') et la dynamique de 'Bush team': Les versions de Bush, Cheney, Rumsfeld, Powell, Rice, Ashcroft, Ridge, Wolfo, Card ((Johnson & Krempel 2004)).

La cooccurrence des éléments textuels, notamment les noms des personnes sur les pages web, est un autre moyen pour extraire des SNs de plus en plus implicites sur le web ((Mika 2005a)) ((Jin et al 2007)), en évaluant la force des relations entre les personnes. Des indices comme le coefficient de Jaccard ou encore le coefficient de recouvrement sont adoptés pour estimer cette force de relation entre 2 individus. Par exemple, ((Mika 2005a)) se base sur le nombre des pages contenant 2 noms X et Y par rapport la fréquence des occurrences de chaque nom

indépendamment ((Erétéo 2011)). Généralement, ces méthodes et mêmes des outils (comme Polyphonet ((Erétéo 2011))) pour extraire des SNs en 'Web Mining' sont limitées par un contexte d'étude spécifique.

4.2.2. Réseaux sociaux en ligne (OSN)

L'avènement des nouvelles technologies de communication du web 2.0 a facilité la création des interactions sociales en ligne qui sont devenues de plus en plus explicites. Selon Davis ((Leblanc 2008)) le succès du web 2.0 provient de l'association attitude\ technologie qui encourage la participation grâce à des applications et services techniquement révolutionnaires mais surtout socialement ouverts ((Leblanc 2008)). Ainsi, ces technologies dites participatives ont rendu les internautes des acteurs participatifs 'Internaute contributeur' ((Leblanc 2008)) ou encore 'entité sociale' sur le web. Même si ces interactions sont médiatisées par ordinateurs dans différents contextes, elles sont créées à partir des interactions sociales et initiées entre des personnes physiques, suivant des liens de causalité du monde réel: Connaissances amicales, des intérêts généraux, intérêts professionnels, des activités, les relations familiales et associatives, etc., (Nettleton 2013). Cet univers social sur le web qui semble être virtuel, amplifie la connectivité des utilisateurs sur web, contribue à l'émergence des SNs médiatisés par ordinateurs et les rapprochent qualitativement aux SNs réels. Il constitue une vraie source pour révéler des structures de SN ayant des propriétés de plus en plus similaires d'un SN réel ((Wellman 2001)) ((Leblanc 2008)).

4.2.2.1. Des SNs extraits depuis des applications-ordinateur de discussion

Les applications et les outils de discussion et de collaboration en mode synchrone ou asynchrone (Mailing chat, forum, wikis, blogs, etc.) permettent suivant des Data logs d'extraire des SNs. Beaucoup de chercheurs sont intéressées aux systèmes des emails qui sont fréquemment utilisés au sein des entreprises et en général dans un contexte institutionnel. Un réseau peut se construire à partir les entêtes des emails: les adresses de l'émetteur et récepteur ou même depuis la messagerie instantanée ((Leskovec & Horvitz 2008)). Les caractéristiques des SNs réels sont démontrées dans tels graphes sociaux ((Leskovec & Horvitz 2008)) ((Erétéo 2011)). Par exemple ((Leskovec & Horvitz 2008)) ont détecté une distance moyenne de 6.6 entre les utilisateurs de 'Microsoft Messenger' (communications par messagerie instantanée) et qui incarne l'effet du petit monde ((Leskovec & Horvitz 2008)). Ces graphes sociaux sont également perçus bien connectés avec une tendance au clustering. En effet, ils peuvent afficher des structures communautaires qui sont même validées dans le monde réel comme dans l'étude de Tyler ((Erétéo 2011)) et regroupent les personnes ayant déjà le même âge, la même langue et emplacement et qui se communiquent fréquemment ((Leskovec & Horvitz 2008)). Ainsi, ***le bénéfice informationnel des études analytiques appliquées sur ces réseaux se multiplie*** selon le contexte (ex. sur le plan économique).

4.2.2.2. Des SNs explicitement émergents sur les services et applications d'OSN

Le phénomène des réseaux sociaux en ligne (OSN) est explicitement amplifié par les plateformes et les applications sociales 'des médias sociaux' (web 2.0), (Erétéo et al 2009).

Tableau 13. Exemples d'applications et plateformes qui ont commencé à amplifier le phénomène des OSNs contemporaine.

1994	Geocities Une première application avec un environnement de Chat-Rooms qui facilite l'interaction entre personnes ((El Akkad & So Long 2009)).
1997	SixDegrees ((Boyd & Ellison 2007)) plus de fonctionnalités pour gérer les profils d'utilisateurs et listes d'amis.
2002-2003	Friendster [6] en 2002, MySpace et LinkedIn, (Nettleton 2013)
2004	Lancement de Facebook
2005-2010	La période de croissance de l'utilisation des applications d'OSN (Nettleton 2013): **Facebook, Twitter, LinkedIn, Google+, MySpace sont les OSNs les plus populaire au niveau mondial**, en offrant plus de fonctionnalités : discussions, albums photo, broadcast, collaborer dans les jeux, etc. En 2009 Facebook est devenu le plus grand site d'OSN (avec plus de 845 millions d'utilisateurs en 2011). Plus de 100 millions sur Twitter en 2010 ((Erétéo 2011)). 61 millions sur MySpace (Nettleton 2013). **Des Applications spécifiques d'OSN populaire à l'échelle nationale** : Chine : RenRen[8] (environ de 160 millions d'utilisateurs), Weibo[9] (application chinoise de 'microblogging' social avec 300 millions d'utilisateurs) (Nettleton 2013). Espagne : Tuenti[10], Amérique centrale et du Sud : Hi5[11], Brésil & Inde : Orkut[7], Allemagne : StudiVZ[12], France : Skyrock[13]. Pays-Bas : Hyves. **Des Applications spécifiques pour certaines activités :** Partage de photos : Flickr, Picasa, Partage de Vidéos & Musique : YouTube ou Spotify etc.

Diverses fonctionnalités et moyens sont intégrées pour pouvoir se socialiser et s'exprimer facilement. Maintenant, les médias sociaux ne sont plus un sous-ensemble du web. Ils sont omniprésents sur quasiment tous les grands sites web avec des contenus recyclés (Malgré la notion de '*Walled garden'* [3]). Il s'agit d'un **Web Social** basé sur plusieurs catégories de médias sociaux qui sont apparues. Frédéric Cavazza propose dans son 'blog post' [3] un panorama qui catégorise les services des médias sociaux, et qui est révisé chaque année et qui marque l'effet d'évolution du marché de ces services et son utilisation.

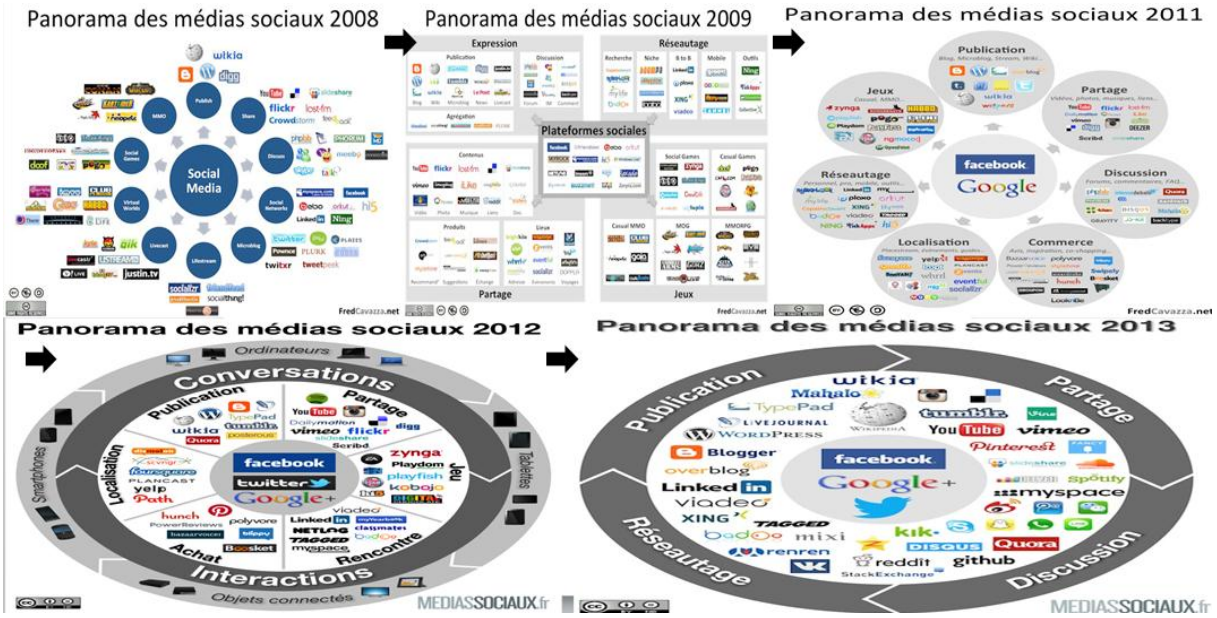


Figure 20. Chronologie des panoramas de Frédéric Cavazza [3] catégorisant les médias sociaux et son usage en évolution ((3) http://www.fredcavazza.net/2014/05/22/social-media-landscape-2014/)

Depuis 2008 jusqu'à 2013, le panorama présente une configuration de plus en plus mûre en classant les services selon leur usage dans des catégories de plus en plus densifiées [3]: Passant de 16 catégories en 2009 vers 4 seulement en 2013:

Tableau 14. Catégories des services médias sociaux.

Services de Publication	Services de Partage	Services de Discussion	Services de Réseautage
les plateformes de blog (WordPress, Blogger, Live Journal, etc.) [3] Les wikis (Wikipédia, Wikia, Mahalo, etc.) [3]	Partage de photos, vidéos, musique, etc. (Flickr, Pinterest, YouTube, Vimeo, Dailymotion, Spotify, Deezer, SoundCloud, MySpace[14], Slideshare, Delicious[49], etc.) [3]	Plateformes conversationnelles (Skype, Quora, Reddit, Github, Tencent Weibo, etc.) [3] Applications mobiles de communication (Facebook Messenger,	SN grand public (Tagged, Nextdoor, etc.) [3] National ou continental (asiatiques et russes): Qzone,VKontakte, RenRen, Mixi[3], etc. SNs BtoB (LinkedIn[15], Viadeo, Xing) [3] Services de rencontre

	Applications mobiles (Instagram, Vine, etc.)	Viber, BlackBerry Messenger, Kik, MessageMe, Pheed, Nimbuzz, etc.) [3]	(Badoo, OKcupid, etc.) Applications de rencontre (Tinder, Skout) [3]

Facebook[16], Twitter[17] et Google+ sont les 3 plateformes sociales 'opérateurs de OSN' ((Maheswaran et al 2010)) dominantes, au centre des panoramas[3] et qui couvrent quasiment tous les usages précédents et ne cessent pas d'évoluer. Facebook est la plateforme de référence et également la plus complexe [3]. Tandis que Twitter est le grand rival, même si la taille de son audience est 4 fois plus petite que Facebook, mais il a une grande résonance puisqu'il est le plus préféré par les personnalités (politiques, sportifs, artistes, etc.)[3]. Google+[40] constitue la couche sociale de Google qui bénéficie de l'audience de ses différents services[3]. A partir de 2014, ce web social montre des changements remarquables. D'abord, la concurrence provient actuellement de cette vague d'applications mobiles populaires de communication: nord-américaines et asiatiques, vue que les Smartphones sont le premier outil de communication dans le quotidien des consommateurs (6 applications mobile dominantes au centre). Par conséquent, le panorama de 2014 [3] donne l'image d'un web social, mobile et mondial qui semble être un iceberg dont la partie visible couvre toutes ces applications et services.

Figure 21. Panorama des médias sociaux de 2014 proposés par Frédéric Cavazza [3] dans son blog post ((3) http://www.fredcavazza.net/2014/05/22/social-media-landscape-2014/)

Les usages sont plutôt linéaires que cycliques, car les utilisateurs publient ou partagent en mode publique ce qui permet de générer des conversations ainsi rencontrer de nouvelles personnes. En outre, des plateformes sociales comme Facebook qui a dépensé par exemple plus de 19 millions $ pour racheter WhatsApp [3], montrent une tendance d'éviter le syndrome du portail en investissant et multipliant les applications mobiles. Elles se préparent pour le

commerce et même la transformation en médias payants. En revanche, les changements résident aussi dans la partie immergée avec des acteurs et des grands groupes plus traditionnels en médias, commerce, technologie et services. Ces acteurs s'intègrent, rachètent et investissent dans les médias sociaux en cherchant des relais de croissance et donc créer 'un marché gris' actif [3].

Tableau 15. Les grands acteurs traditionnels qui s'intègrent, rachètent et investissent dans les médias sociaux (Le marché gris du web social et mobile).

Médias	Commerce	Services & Technologies
Yahoo!, AOL, GlamMedia Le français Webedia qui a racheté Overblog. le russe DST (propriétaire de Mail.ru) [3] a investi dans Facebook, Twitter, Zynga, Spotify, etc.	Amazon, Alibaba, Le japonais Rakuten qui participe dans Pinterest et Kik Messenger [3]	Google, Tencent

Selon Cavazza [4], les médias sociaux forment un vaste écosystème[5] de services et applications mobiles. Par conséquent, toutes ces technologies (web social) et usages offrent aux sociologues ainsi qu'aux chercheurs en SNA, une source des données sociales ayant des caractéristiques des sociétés réelles (en termes de tailles ou de propriétés d'un SN) avec plus de richesse informationnelle. En effet, les relations sont explicitement déclarées entre utilisateurs. Une étude comme dans ((Bonneau et al 2009)) montre comment un graphe social des étudiants de Stanford et Harvard peut être reconstitué ((Bonneau et al 2009)) à partir de leur profil sur Facebook depuis '*public search listings*' ((Maheswaran et al 2010)). 'The public listings' donne un aperçu sur la liste des amis d'une personne (jusqu'à 8 amis) sans rejoindre son réseau ((Maheswaran et al 2010)). Cette approximation du graphe social via 'public listings' ((Bonneau et al 2009)) ((Maheswaran et al 2010)) était suffisante pour que certaines mesures et techniques appliquées, donnent des informations précieuses et significatives: en centralité d'intermédiarité et le découpage en communautés.

D'autre part, au lieu de faire des extractions couteuses (par exemple : web Mining), des APIs (*Application Programming Interfaces*) appropriées sont fournies par les opérateurs d'OSN comme Twitter[22] et LinkedIn[23] afin d'exporter des structures de SNs et les représenter sous formes de graphes sociaux. Ces APIs permettent (au programmeur) de faire son propre grattage 'scraping' de ces OSNs, sous certaines restrictions de confidentialité (Nettleton 2013). Cependant, des traitements d'inférences sont souvent nécessaires pour résoudre des problèmes comme l'interopérabilité, qui surgissent depuis la diversité des services, APIs et ainsi les formats de ces données. L'utilisation d'une seule API commune est une initiative de la part de « Open Social[18] (interface qui supporte plusieurs opérateurs OSN), ((Maheswaran et al 2010)), de Google qui semble être une solution mais n'est pas encore adoptée par ces grands services sociaux ((Erétéo 2011)). Par ailleurs, la multiplicité des profils d'une même personne peut être un élément trompeur dans l'analyse des OSNs. Toutefois, les OSNs

peuvent exprimer plus de richesse informationnelle, ce qui est un grand point avantageux devant des problèmes précédents (l'interopérabilité, multiplicité, etc.). Par exemple, l'orientation, typage, pondération, **time ordering**, etc. des liens permettent d'**enrichir SNA**. Par exemple, les réseaux sur Facebook ont des relations symétriques, avec la possibilité d'étiquetage pour distinguer leurs types et filtrer les contacts (famille, amis) d'un utilisateur. Tandis que, les relations sont asymétriques 'follows' sur Twitter (graphe orienté) ((Erétéo 2011)). Les utilisateurs de Twitter 'tweet' sur n'importe quel sujet et suivent d'autres en recevant leurs tweets ((Kwak et al 2010)). En outre, la pondération des liens sur ces graphes sociaux, permet d'évaluer la fréquence d'interactions en termes de messages, commentaires ((Erétéo 2011)) ou par exemple l'intensité d'interaction collaborative entre les auteurs sur les pages Wikipédia ((Brandes et al 2009)). Les réseaux sur les services de partages sont plus compliqués. Il y a des interactions qui peuvent impliquer 2 ou 3 types de nœuds ((Erétéo 2011)): Utilisateur, Ressource (photos, vidéos, bookmarks, site web, documents, etc.) et même Tag, qui nécessitent des représentations comme: graphes biparties, des hypergraphes (hyperliens), etc. ***Cependant, la richesse de ces SNs n'est pas évidente à exporter et capturer sur des représentations (graphes) topologiques et statiques***.

4.2.2.3. Inférence des OSNs implicites (depuis le 'Social Tagging': Folksonomies)

L'évolution de ces services et applications sociales laissent les personnes exposer de plus en plus leur vie sociale, activités en ligne ainsi que leurs intérêts à travers par exemple des tags. D'abord, le tag[19] est un mot clé utilisé pour annoter (décrire) une ressource partagée en ligne. Les tags sont aujourd'hui des métadonnées qui aident par exemple à gérer et chercher des liens vers des sites et contenus favoris et pertinents. Par exemple pour tagger un blog d'un club de football américain 'Green Bay Packers'[19], des tags comme: Blog, Green Bay[19], Packers[19] et football sont envisagés. Récemment, les tags sont interprétés automatiquement dans les grandes plateformes sociales sous forme de 'hashtag' (#Tag) : Twitter depuis 2009, Facebook depuis 2013, etc. Bibsonomy[21] est aussi un exemple de système de partage, publication et 'social-bookmarking' qui contient un grand nombre de publications (et leurs auteurs) en informatique, et qui permet aux visiteurs d'annoter ces ressources par des tags (Nettleton 2013). L'ensemble des tags comme : 'Theory', 'Software', 'Ontology', etc., associés à chaque papier montrent également les centres d'intérêts des auteurs correspondants. 'Flickr personal taxonomies' ((Plangprasopchok et al 2010)) est un autre exemple plus clair qui reflète l'intérêt des utilisateurs (groupes publiques des fans de la photographie de la nature) dans le monde naturel. Les utilisateurs organisent le contenu (les photos du monde de la nature) hiérarchiquement avec des tags descriptifs. Cela engendre des collections comprenant des ensembles annotés par des tags : Les tags sont propagés à partir des ensembles vers des collections parents ((Plangprasopchok et al 2010)).

Le 'Social Tagging' est ainsi une classification collaborative ((Erétéo 2011)) de ces ressources annotées avec des tags par les internautes. Il s'agit aussi de « collaborative tagging systems[31] » qui constituent des vrais **systèmes sémiotiques[31]** en ligne. Les données de ces usages peuvent être une base d'une ***véritable investigation scientifique sur le comportement des agents humains sur le Web et la dynamique de l'information dans les communautés en ligne***[31]. Le projet de recherche ***TAGora***[31] est une illustration qui exploite cette opportunité en

cherchant dans ces données selon différentes points de vues informatique, systèmes complexes, science cognitive, psycholinguistique et architecture de l'information. Le 'Social Tagging' se réfère à une sorte de taxonomie sociale ou encore une folksonomie, un terme introduit par l'architecte de l'information Thomas Vander Wal[20]. Formellement, la structure de folksonomie $F = \{U, T, R, Y\}$ est définie par des triplets $Y \subseteq U \times T \times R$ qui sont des relations ternaires ('Tagging instances' ou 'assignment') définies sur les ensembles finis $U, T\ et\ R$ des 'users', tags, et ressources respectivement ((Limpens 2010)). T, représente une liste de tags ou bien un nuage de tags 'tag cloud' dont les tags populaires sont communément notée par une police plus grande, plus foncée par rapport les moins populaires. Une telle structure a été formalisée par ((Mika 2005b)) qui modélise ces instances de 'Social Tagging' par un hypergraphe: un graphe tripartite :

$H(F) = (V, E),\ V = U \cup T \cup R\ et\ E = \{u, t, r | (u, t, r) \in F\}$, ((Limpens 2010)). Chaque lien représente l'assignation ternaire d'un tag à une ressource par un acteur.

Différents traitements sur les dataset de 'Social Tagging', depuis lesquels des folksonomies sont extraites sous forme d'hypergraphes. Par exemple dans (Bothorel & Bouklit 2008), un hypergraphe de folksonomie a été extrait depuis Flickr, et sur lequel l'algorithme de (Newman & Girvan 2004) a été généralisé (sur un graphe triparti). ((Mika 2005b)) a construit à partir le 'Social Tagging' d'un service de 'bookmarking': delicious[49] ((Erétéo 2011)), 2 graphes bipartis (Two-mode networks). Le premier H_1 relie les acteurs en U à leurs tags en T alors que le deuxième relie T avec les ressources annotées R. De ce fait, cette formalisation de 'Social Tagging' offre une source précieuse pour découvrir **des acteurs sémantiquement liés** à travers leurs intérêts communs et ainsi des **SNs implicites**. H_1 permet d'inférer un SN d'affiliation qui relie chaque paire d'acteurs partageant les mêmes tags, à travers un lien pondéré par le nombre des tags partagés entre les deux. L'analyse de tels réseaux donnent des résultats significatifs. Par exemple, l'application d'une détection de communautés, permet d'extraire **des communautés d'intérêts** (groupes d'acteurs qui partagent les mêmes intérêts). D'autre part, deux réseaux de tags peuvent être également inférés depuis H_1 et H_2, pondérés respectivement par le nombre d'acteurs qui utilisent chacun deux tags ou le nombre de ressources annotée par 2 tags ((Erétéo 2011)). Dans ((Brandes et al 2009)), un autre exemple d'extraction et typage (étiquetage) de liens implicites à partir les interactions collaboratives de Coédition et révision des pages Wikipédia, entre auteurs. Différents graphes sociaux selon les types de liens définis dans les wikis, ont été construits et permettent d'obtenir une meilleure interprétation des acteurs intermédiaires et des structures communautaires dans Wikipédia ((Brandes et al 2009)) ((Erétéo 2011)).

4.3. Données de SNs\OSNs : Collections, échantillonnage et illustrations

Tableau 16. Classification des datasets selon différents critères visant des interprétations de SNA plus significatives et informatives

Données	Taille	Implicite	Pertinence des données	Signification des Interprétations de SNA
Données collectées par	Ne sont pas		Des interactions et le	Techniques de SNA

des sociologues, de la science comportementale, etc.	de grandes tailles	-	contexte du monde réel	ont été originalement appliquées sur les données du monde réel
Données synthétiques (Générateurs artificiels)	Graphes de grandes tailles	-	Certaines propriétés de SN sont reproduites.	Evaluation des performances sans efficacité et bénéfice informationnel
Graphes de pages web	Graphes de grandes tailles	-	Pas de propriétés de SN : Pas de vraies entités et interactions sociales.	Evaluer juste la qualité des performances
Données extraites par Web Mining	-	Des relations sont implicites inférées (la cooccurrence des éléments textuels) et parfois sont explicites	Les graphes extraits montrent des propriétés de SN (graphes sociaux). Le contexte est parfois intéressant	Tout dépend du contexte d'étude spécifique.
OSN (Outils de discussions emails, etc.)	Tout dépend du contexte	Les relations et leur typage nécessitent des extractions	Interactions collaboratives fréquentes dans le contexte institutionnel et qui affichent les caractéristiques de SN (plus de richesse informationnelle)	Les bénéfices informationnels se multiplient selon le contexte et la pertinence des données (Pour la fouille des organisations et environnements).
OSN (plateformes sociales)	Graphes sociaux à grande échelle	Relations peuvent être explicitement déclarées et captées et ce n'est pas le cas souvent pour le typage	OSNs affichent les propriétés des SNs réels (plus de richesse informationnelle). Des relations plus ou moins pertinentes	La recherche des interprétations plus significatives sachant que les représentations sont souvent topologiques statiques
Données extraites depuis les usages du web sociales : Social Tagging, wikis, etc.	-	interactions implicites entre acteurs (parfois sémantiquement liés)	Des liens entre des acteurs ayants des orientations, intérêts, etc., communs	Des interprétations informatives (Détecter des communautés d'intérêt)

Les résultats et les interprétions de SNA sont plus ou moins informatifs, tout dépend de la qualité des données traitées. Cette qualité dépend premièrement de l'étude analytique elle-même, ainsi que de plusieurs autres facteurs\ considérations selon des degrés différents. Par exemple la taille des réseaux à grande échelle aide à fournir des réponses en performances de

calcul mais l'étude risque d'avoir une tendance statistique. Les liens implicites extraits entre les acteurs peuvent être plus pertinents par rapport à leurs usages explicitement déclarés et capté de ce web social. ***Par ailleurs, les interprétations seront plus significatives sur des modèles de représentation plus réalistes qui supportent plus de richesse informationnelle. Dans notre contexte d'étude, il s'agit d'une richesse en termes de dynamique temporelle et richesse sémantique.***

4.3.1. Collections

Beaucoup de collections de datasets sont accessibles en ligne librement, ou sous permission\autorisation des auteurs originaux :

- Mark Newman (Université de Michigan) expose dans une page personnelle[25] une collection de datasets de réseaux (notamment de SNs) les plus utilisées dans ses études au fil des années.
- Plus de 300 datasets sont proposés sur la page de Linton C. Freeman[75] qui présente un ensemble de SNs (classiques) collectées par des sociologues, des sciences comportementales, les analystes des collaborations scientifiques et autres.
- Le site web de SNAP[24]: SNAP 'Stanford Large Network Dataset Collection' propose des collections de SNs (graphes assez larges, pour 'Benchmarking') dont certains sont très connus et extensivement utilisés dans la littérature, avec également des réseaux de communications et de transport.
- Des données collectées par Lingfei Wu et Cheng-jun Wang[30] sur la page personnelle de ce dernier qui fait des références vers SNAP ainsi que d'autres ensembles de données qui captent le comportement humain en ligne (sa socialisation) et qui sont aussi accessibles au public.
- Eric D. Kolaczyk propose une liste de datasets (différents types de réseaux) [28] utilisés dans son livre 'Statistical Analysis of Network Data' modèles et méthodes ((Kolaczyk 2009)). Les données qui captent le comportement en ligne des humains (SNs) attirent plus d'attention.
- Tore Opsahl, recueille sur sa page[29] un certain nombre de datasets des réseaux simples (OSN et autres) en modes 'Two-mode Networks', selon des formats standards convenables pour certains softwares.
- ((Cheng et al 2008)) proposent un ensemble de datasets[76] 'Statistiques et SNs des vidéos YouTube' destinés pour l'usage académique.
- L'entreprise web Quora[85] permet aux users de collaborer par questions réponses sur différents sujets. Au sujet de 'Where can I find large datasets open to the public?' Il y a des listes de répertoires et des collections de données interdisciplinaires (linked data), parmi lesquelles, on trouve des données sociales.
- Sur le site web crée par Kevin Chai[86] (chercheur scientifique dans les données et Datamining à l'université 'Curtin' en Australie), certaines données pour l'analyse des liens et des SNs sont partagées

Il y a certaines collections de datasets proposés par des softwares et applications pour l'analyse des réseaux complexes et notamment les SNs. Par exemple :

- La collection des datasets de Pajek ((Batagelj & Mrvar 2006)): Un programme pour l'analyse et la visualisation des grands réseaux ((Batagelj & Mrvar 2012)) ((Batagelj & Mrvar 2003a)) ((Batagelj & Mrvar 1998)) ((Beauguitte 2011)). Certains entre eux présentent des vrais SNs.
- 'UCINET IV Datasets'[78] propose des datasets standards du logiciel UCINET ((Borgatti et al 2002)): un software pour l'analyse des données des réseaux sociaux.
- Des outils de visualisation et de SNA récents comme SoNIA[90] peuvent supporter les réseaux dynamiques (longitudinaux). Dans ce sens, le site web de SoNIA[90] présente une liste de sources de quelques données (dynamique temporelle) de SNs.

En outre, ils existent aussi des datasets collectés dans le cadre de quelques projets de recherche et par des groupes de recherche qui s'intéressent aux réseaux complexes, 'Social computing', etc., par exemple:

- La collection de datasets de 'social tagging' (Des assignations de tags) qui sont utilisés dans le projet TAGora[31] avec certains services ('Tag Filtering', etc.) pour étudier 'Semiotic Dynamics in Online Social Communities'.
- Kristina Lerman[32] (chef de projet à l'institut des sciences de l'information à l'USC) met sur sa page quelques datasets qui sont étudiés dans des projets en 'Social Dynamics and Networks', etc.
- 'The University of Florida Sparse Matrix Collection' [35] propose une galerie de représentation matricielles et des réseaux larges, notamment les collections des graphes-SNs, compilées par Mark Newman et celles de SNAP et autres, etc. Ces graphes sont visualisés par Yifan Hu (AT&T Labs Visualization Group) [44, 45].
- Les membres du groupe Alex Arenas[36], (université Rovira i Virgili, Tarragona, Espagne) compilent un ensemble de datasets: Réseau de collaboration entre musiciens de jazz, réseau d'échange des emails, etc.
- Le groupe de recherche de l'université de Ben Gurion (BGU) [37], spécialisé dans la recherche sur les SNs, propose des collections de datasets d'OSNs orientés, non-orientés, étiquetés (versions anonymes extraites via Crawling).
- L'entrepôt des données sociales de 'Arizona State University' [46] propose une collection datasets de différents médias sociaux ((Zafarani & Liu 2009)).
- Le groupe 'Social computing' [54] de 'Max Planck Institute for Software Systems' (mpi-sws) citent une collection de données sociales qui ont été étudiés dans leurs œuvres et extraites depuis : Flickr, LiveJournal, Orkut, YouTube, et Facebook ((Zafarani & Liu 2009)) ((Cha et al 2010)) ((Cha et al 2009)) ((Viswanath et al 2009)), Twitter : Twitter topology data ((Cha et al 2010)) ((An et al 2011)), Twitter spammer data ((Ghosh et al 2012)).
- Un ensemble de données collectées dans le cadre de la thèse de ((Mislove 2009)) est publiquement disponible, à la demande par courriel : 'amislove@mpi-sws.org'
- KONECT 'The Koblenz Network Collection' [55] est un projet qui vise à recueillir une large gamme de données de réseaux de différents types, collectés par 'Institute of Web Science and Technologies' à l'université allemande de 'Koblenz–Landau'. En 20 juillet 2015, KONECT contient 203 réseaux (orientés, bipartis, pondérés, non-orientés,

etc.) collectés notamment à partir des SNs\ OSNs, des Folksonomies, des réseaux de communication et collaborations, etc., en affichant les caractéristiques de chaque réseau : distribution de degré, coefficient de clustering, etc.

- Le programme 'Yahoo! Webscope'[77] est une bibliothèque intéressante de datasets de différents types, particulièrement les graphes et les données sociales. Ces données sont destinées vers une utilisation non-commerciale par les universitaires et les chercheurs scientifiques selon un accord de partage de données. À partir des millions de communautés et groupes, Yahoo! Webscope offre un échantillon de graphe anonyme d'appartenance biparti users-groups ((Kevin 2005)) ((Vigfusson 2010)). C'est une illustration d'un large graphe du monde réel en présentant la propriété de loi puissance. Un autre exemple concerne un échantillon d'un graphe anonyme d'amitié de Yahoo! Messenger.

- Les datasets du centre CASOS[81] de 'l'analyse computationnelle des systèmes organisationnels et sociaux qui sont généralement recueillis par CASOS à partir des interactions du monde réel. Des travaux considérables de CASOS sont intéressés par des sujets qui concernent la sécurité interne, le terrorisme et la lutte antiterroriste ((Frankenstein et al 2015)) ((Fellman et al 2011)) ((Il-Chul et al 2008)) ((Il-Chul & Kathleen 2007)) ((Max & Kathleen 2005)). Une grande partie de travaux (idées et outils développés) sont actuellement exportables vers d'autres domaines et chercheurs (experts antiterroristes).

- L'association professionnelle des chercheurs intéressés par SNA INSNA[82] (le réseau international de l'analyse des SNs) présente un ensemble de données sociales populaires, collectées à partir différentes sources (les collections ci-dessus).

- Le laboratoire LWA 'Laboratory for Web Algorithmics'[83] du département d'informatique de l'université de Milan, qui cherche dans les aspects algorithmiques et Crawling du web et des SNs, propose des Datasets ((Boldi et al 2004)) ((Boldi et al 2011)): des petits et gigantesques crawls de graphes web et SNs\OSNs produites par les interactions humaines ex. Graphe social des acteurs dans les films Hollywood de 2011, Des sous-graphes Facebook géographiquement (régionaux) et temporellement limités ((Backstrom et al 2012)).

- La communauté CRAWDAD[84] présente une collection de données dont certaines faisant référence à des nouveaux SNs.

- La plate-forme de collaboration entre chercheurs et développeurs pour la télédétection SocioPatterns[89] fournit quelques jeux de données. Par exemple 'High school contact and friendship networks' représente un ensemble de contact et des liens d'amitié direct ou sur Facebook entre les étudiants d'un lycée à Marseille en décembre 2013

((Mastrandrea et al 2015)). Certains réseaux dynamiques temporels sont également proposés.

- En utilisant ses outils développés, ainsi que des données de localisation (appareils et téléphones mobiles et Bluetooth, etc.), le laboratoire de la dynamique humaine de la MIT[91] collecte des Datasets de SNs dans des contextes organisationnels et dynamiques : 'Badge Dataset' ((Olguin et al 2009)), 'Social Evolution Dataset' ((Madan et al 2012)), etc.

- Certains projets comme 'Sampling Online Social Networks' conduit par 'Athina Markopoulou' avec autres chercheurs et collaborateurs [34] travaillent sur l'échantillonnage (via Crawling) des OSNs comme Facebook, et proposent une collection de datasets.

Il existe plusieurs autres ressources et outils comme Databip[88] permettant d'aider les chercheurs à trouver et localiser dans le monde entier des dépôts en ligne des données dans différents domaines notamment en science sociales.

4.3.2. Echantillonnage (par 'Crawling')

L'échantillonnage est un aspect clé pour traiter des grands datasets (Big data), (Nettleton 2013). Il permet de mapper d'un espace de dimension supérieure à une dimension inférieure tout en maintenant les propriétés des données originales (distributions statistiques, etc.). Par exemple, si 10% des nœuds dans le SN ont un degré = 1, un bon échantillon aura la même proportion. Donc, l'échantillonnage semble être une solution pour résoudre le problème de scalabilité des techniques et algorithmes de SNA (ex. l'intermédiarité) sur les OSNs à grande échelle. Certaines méthodes d'échantillonnage sont proposées pour atteindre une meilleure approximation ((Erétéo 2011)) ((Geisberg et al 2008)). Cela permet par exemple d'estimer efficacement la centralité d'intermédiarité en réduisant la complexité de calcul. Ainsi, des algorithmes de détections de communautés comme celui de (Newman & Girvan 2004) peuvent être également optimisés.

Des techniques d'échantillonnage qui sont spécifiquement conçues pour les graphes (Nettleton 2013) sont variantes selon la façon choisie pour suivre\ extraire des liens, nœuds, etc. À partir d'une population d'intérêts, le problème classique consiste à retirer un échantillon tel que la probabilité pour qu'un individu soit inclus est connue ((Gjoka et al 2011)). Il suffit d'estimer la propriété de l'intérêt d'un échantillon de nœuds. Cependant la population dans les OSN ne peut être énumérée ainsi que l'échantillonnage est moins efficace. Donc les auteurs s'orientent vers des méthodes alternatives basées sur les réseaux. Par exemple, les liens sociaux peuvent être exploités pour trouver une probabilité d'échantillonnage de telle population cachée ou bien échantillonner par 'crawling' ou 'link-trace sampling' ((Gjoka et al 2011)). Selon ((Gjoka et al 2011)), il existe 2 catégories de méthodes moins utilisées (moins efficaces): indépendantes et 'traceroute' et une catégorie des méthodes d'exploration (Crawling). Les méthodes de 'Crawling' sont typiquement les plus applicables et adaptés pour échantillonner les graphes des OSNs.

Deux classes de 'Crawling' sont distinguées. La première couvre des approches dites 'Traversals' ou bien un échantillonnage sans remplacement basées sur BFS, DFS ou encore Snowball, etc. La méthode 'Snowball' est un exemple dont l'algorithme générique original est proposé par Goodman (Nettleton 2013). Sur un graphe d'OSN, le principe consiste à choisir les voisins immédiats d'un nœud donné et ainsi de suite. C'est une résiliation anticipé de BFS. L'échantillon résultant dépend de la sélection des nœuds initiaux qui est un aspect critique dans ce type d'échantillonnage. Par exemple le choix aléatoire parait statistiquement correct mais la distribution de degré du réseau échantillonné peut s'incliner (Nettleton 2013). Si le 'Snowballing' commence à construire le graphe à partir des nœuds bien connectés, il finira par ignorer les individus isolés ou à faible degré en favorisant un nombre disproportionné de nœuds ayant des degrés élevés ((Chakrabartiet al 2004)). Plusieurs auteurs comme ((Chakrabartiet al 2004)) ainsi que (Mislove et al 2007) confirment tel problème lors de l'échantillonnage ou 'Crawling' des datasets de Flickr, LiveJournal, YouTube, etc. En outre, les auteurs soulignent également que l'échantillon Snowball peut ignorer tout une composante faiblement connectée (2-star, triangles isolés, etc.) en dehors d'une composante géante (Nettleton 2013) ((Snijders 2010)).

Certaines solutions sont proposées pour améliorer la méthode 'Snowbal' de telle sorte que la sélection des nœuds soit influencée en affectant des poids. Par exemple Snijders (Nettleton 2013) suggère des approches de pondération suivant des notions comme les relations symétriques, transitivité, etc. Par ailleurs, ((Shafie 2010)) propose un schéma de pondération qui calcule la probabilité d'inclusion d'un nœud associé avec la probabilité d'inclusion de ses voisins comme suivant ((Shafie 2010)): $\left[\frac{1}{n} + (n-1)\frac{d(i)}{\sum_i^n \mathrm{d(i)}}\right]$

Donc, la deuxième classe de Crawling inclus des alternatives basées sur 'Random Walks (RW)' visant à améliorer l'échantillonnage avec remplacement : RW, RWRW, MHRW, ((Gjoka et al 2011)). Les auteurs montrent la qualité des performances d'un Crawling via RW testé sur le graphe social Facebook par rapport à des méthodes comme BFS mais avec certaines limites. Dans ce cas, 'Star sampling' ou 'Induced Subgraph Sampling' peut être plus favorisé.

Parmi les questions clés du Crawling : La sélection des nœuds initiaux, préciser quelle information à collecter, quand s'arrêter et atteindre la convergence, comment évaluer la qualité de l'échantillon dérivé, etc. À titre d'exemple, ((Bartz et al 2009)) testent 2 méthodes d'échantillonnage 'multiple bridge' et 'Snowball' suivant des modèles de générateur de graphes 'un modèle de graphe aléatoire exponentiel' et 'estimation du maximum de vraisemblance' (MLE) ((Bartz et al 2009)). La qualité est mesurée en termes de fréquence des structures : les triangles et 2-stars. Des idées s'orientent également vers obtenir une connaissance préalable sur la façon dont un graphe social se produit et l'appliquer en échantillonnage (avec Snowball), (Nettleton 2013). Dans ce sens, les petites structures (2-stars, triangles, etc.) sont informatives, car elles sont liées aux paramètres qui déterminent la connectivité et comment les grands structures sont formées (Nettleton 2013).

La stratégie d'échantillonnage des données sociales a un impact sur la découverte de l'information, comment elle se diffuse ((De Choudhury et al 2010)) et l'analyse des SNs en général. Le but de toutes ces améliorations est de dériver un échantillon le plus représentatif possible et fidèle à ces sous-graphes. Une approche d'échantillonnage est jugée innovative ou efficace selon la qualité de l'échantillon produit. Par exemple un algorithme qui s'applique sur un échantillon de dataset, dit représentatif, génèrera une partition en communautés représentative comme sur le réseau original ((Maiya & Berger-Wolf 2010)).

4.3.3. Illustrations

Nous illustrons en 3 catégories selon le point de vue de (Nettleton 2013), les datasets couramment **populaires** chez les analystes des SNs et **pertinentes** comme dans le tableau suivant:

Tableau 17. Exemples de datasets utilisés selon 3 catégories

Exemples de datasets & descriptions	Nombre de nœuds	Nombre de liens
SNs et des réseaux de collaborations ayants des propriétés des SNs, collectés ou intégrés dans les outils et softwares de SNA (pour des tests) Données collectées par les sociologues, des sciences comportementales, analystes des collaborations scientifiques et autres		
Dataset du Club de karaté de Zachary[35, 25,75] (Zachary 1977)	34	156 (liens symétriques)
Dataset de 'bottlenose dolphins\ social network of dolphins, Doubtful Sound, New Zealand' [25, 35] étudié par (Lusseau et al 2003) ((Lusseau 2006))	62	159 (318 en intégralité) liens symétriques
Réseau de co-apparence des personnages dans le roman 'Les Misérables' [35,75] ((Knuth 1993))	77	508 (liens symétriques pondérés)
Réseau de collaboration des scientifiques résidents dans le Santa Fe Institute sur une période de deux ans (Nettleton 2013) testé par Girvan & Newman dans (Girvan & Newman 2002)	271	
'Seventh graders' [65,75] Réseaux des élèves de septième année d'une école à Victoria. Des questions permettent à chaque élève de nommer et évaluer les camarades de classes préférés ((Watts & Strogatz 1998)) ((Vickers & Chan 1981)) ((Wasserman & Pattison 1996)) ((Robins et al 1999)).	29 Elèves	376 liens asymétriques pondérés de 1 à 3 : Un élève favorise un autre pour l'activité : 1, 2 ou 3
PADGETT FLORENTINE FAMILIES[27, 75] ((Kent 1978)) ((Breiger & Pattison 1986)).	16	
Réseau de 'American football games between Division IA colleges during regular season Fall 2000' (Girvan & Newman 2002) de Newman[25]	115	1 226 (liens pondérés symétriques)

'Football.net' : Réseau des équipes nationales participant à la coupe du monde en France 1998 qui ont des joueurs qui jouent à l'étranger ((Batagelj & Mrvar 2006)) ((Brandes et al 2001)).	35 nations dont 22 ont participé au championnat du monde (à paris)	118 liens asymétriques pondérés par le nombre des joueurs exportés vers d'autres nations.
'The Dutch Soccer Team network' : DST SN ((Kooij et al 2009))	691 joueurs sélectionnés jusqu'à juin 2008	10 450 liens (entre joueurs sélectionnés dans le même match)
WOLFE PRIMATES[27] (Everett & Borgatti 1999)	20	
'SOUTHERN CLUB WOMEN' [29,75] collecté par ((Davis et al 1941)) ((Breiger 1974))	18 femmes participantes à 14 évènements sociaux	278 en 'one-mode ties' 89 en 'Two-mode ties'
'Norwegian Interlocking Directorate' (Aout 2009) [29] ((Seierstad & Opsahl 2010)) : Réseau bipartite	367 'Companies' 1 495 'directors'.	1834 en 'two-mode ties' 4065 en 'One-mode ties'
Dataset de 'Corporate interlocks in Scotland (1904-5)' ((Batagelj & Mrvar 2006)) ((De Nooy et al 2004)) ((Yongcheng Xu et al 2013)): Réseau bipartie	244 nœuds:136 'multiple directors' et 108 entreprises	356 liens (Direction),
'Madrid Train bombing' [63,75] Réseau des contacts entre les terroristes présumés impliqués dans l'attentat du train de Madrid le 11 Mars 2004 ((Hayes 2006))	64 terroristes	243 contacts : Liens symétriques, pondérés. poids inclut l'amitié, coparticipation dans un camp d'entrainement ou attaque précédente
'Infectious' [64] Réseau des contatcs face-à-face des personnes visiteurs de l'exposition INFECTIOUS: STAY AWAY en 2009 à la Galerie des sciences à Dublin ((Isella et al 2011))	410 Visiteurs	17,298 Contacts en face-à-face pendant au moins 20 secondes. liens symétriques multiples avec horodatage: **Timestamp**
'Jazz muscicians network' [36] ((Gleiser & Danon 2003)) ((Guimerà et al 2003)): Réseau de collaboration entre les musiciens Jazz, collecté en 2003.	198 (musiciens)	2 742 (5 484): un lien entre 2 musiciens ayant joué dans un groupe
Réseau d'affiliation (bipartite) des athlètes à leurs équipes [57] ((Auer et al 2008)), extrait de DBpedia	935 627 athlètes et équipes	1 366 466 liens d'affiliation
Réseau de collaborations entre co-auteurs scientifiques travaillant sur la théorie et expérimentations dans les réseaux [25, 35] ((Newman 2006))	1 589	5 484 liens pondérés symétriques
Réseau de collaboration scientifique[29, 25] de ((Newman 2001b)) ((Newman 2001c)) ((Newman 2001d)): Coauteurs et publications dans la section 'Condensed Matter' de arXiv E-Print Archive 1995-1999	16 726 auteurs 22 016 publications	47 594 en 'two-mode ties' entre auteurs.

Réseau de collaborations entre les auteurs des articles de :'High Energy Physics – Phenomenology', 'e-print arXiv'[24] ((Leskovec et al 2007)), (Janvier 1993-Avril 2003), testé dans ((Wang et al 2013))	12 008	11 8521
'Students' Cooperation Social Network (Multi-Graph)' [37] ((Fire et al 2012b)) des données sont collectés en analysant les coopérations implicites et explicites entre des étudiants participant au cours de 'sécurité informatique et réseau' à l'université BGU [37]. la coopération se fait sur les devoirs.	185 étudiants de 2 départements	360 liens en 3 types
Datatset de 'PhD students in computer science': Réseau des liens entres les doctorants et leurs superviseurs en informatique théorique ((Batagelj & Mrvar 2006)) ((De Nooy et al 2004)) ((Johnson 1984-1985)).	1 025 chercheurs scientifiques en informatique	1 043 liens asymétriques pointant à partir d'un superviseur vers un doctorant
'Amazon Political Books about US politics' de V. Krebs[26, 25, 35,75] ((Krebs 2012)): Réseau de livres sur politiciens, vendus au moment des élections présidentielles de 2004 en US, sur Amazon.com : un lien représente le co-achat fréquent des livres par les mêmes acheteurs Comme indiqué avec la caractéristique 'customers who bought this book also bought these other books' sur Amazon.	105	882 liens symétriques
Réseau de 'Western States Power Grid' [25, 35] de Watts & Strogatz ((Watts & Strogatz 1998))	4 941	13 188 liens symétriques
'hollywood-2011'[83] ((Boldi et al 2004)) ((Boldi et al 2011)) Graphe social des acteurs dans les films Hollywood	2 180 759	228 985 632 Liens symétriques entre les pairs d'acteur qui apparaissent dans le même film
Plusieurs autres datasets sont des standards dans des softwares : UCINET (Steve Borgatti), Pajek (Vladimir Batagelj and Andrej Mrvar), etc.		
SNs depuis les Data logs des applications qui ne sont pas strictement d'OSN Données sociales depuis les outils et plateformes de collaboration et de discussion (Mailing chat, forum, wikis, blogs, etc.)		
'Enron dataset' [24] ((Leskovec et al 2009)) ((Klimmt & Yang 2004)) composé depuis le log des emails échangés entre les employés de la société 'Enron Corporation' pendant une période donnée (Tang et al 2010b) (Nettleton 2013).	36 692, Comprenant des adresses mail qui ne sont pas d'Enron.	18 383, Comprenant des échanges avec des adresses mail qui ne sont pas d'Enron
wiki-Vote ((Leskovec et al 2010b)) ((Leskovec et al 2010c)), Données du vote (2 794 élections) en Wikipédia depuis la création de Wikipédia jusqu'à Janvier 2008	7 115 Wikipédia users	103 689 (Arc: user i vote sur user j) : asymétriques

email-EuAll : Réseau extrait depuis 'email data' d'une grande institution de recherche européenne depuis Octobre 2003 jusqu'au mai 2005 (18 mois), ((Leskovec et al 2007))	265 214	420 045 liens asymétriques
'Political blogosphere Feb. 2005' [35, 25], Réseau d'hyperliens entre les weblogs des politiciens d'US en 2005, compilé par ((Adamic & Glance 2005)) via Crawling	1 490	19 025 lins pondérés asymétriques (arcs)
'Digg 2009' historique de votes & liens d'amitiés, des users sur la page d'accueil de Digg[33] pendant un mois en 2009	71 367 users distincts 139 409 users électeurs	1 731 658 liens d'amitié 3 018 197 votes sur 3553 histoires et contenus populaires
'Academia Dataset' [37, 39] ((Fire et al 2011)) ((Fire et al 2013a)) à partir de Academia.edu[39] une plateforme de partage des documents de recherche	200 169	1 398 063 liens asymétriques
'Last.fm song' [66] Réseau bipartite 'user-song' en liant les users à leurs habitudes d'écoute ((Celma 2010)).	992 users 1 084 620 chansons	19 150 868 liens (écoutes) : Avec des liens multiples et horodatage: **Timestamp**
'Filmtipset' [67] Réseau bipartite entre les users-movies tel que : users commentent des films sur le site web suédois Filmtipset.se ((Said et al 2010))	29 530 users 45 830 films	1 266 753 liens (commentaires) : Avec des liens multiples et horodatage: **Timestamp**
'BibSonomy user–tag [61] ((Benz et al 2010)) (folksonomie) Réseau bipartite des users et leurs tags utilisés pour les documents. Le réseau est extrait à partir le dépôt officiel de Bibsonomy[21]	5 794 users 204 673 tags	2 555 080 assignations : comprenant des liens multiples Avec horodatage: **Timestamp**
'Trust network of Advogato' [72, 87] ((Massa et al 2009)) ((Massa et al 2008)) Réseau de confiance entre les users de: Advogato[71]	6 541 users : Apprenti Journeyer, Master ou observateur, selon les poids de confiances	51 127 Liens pondérés :Trust Relationship 'certification' à 3 niveaux (3 type de poids)
'Manufacturing emails' [58] Réseau interne de communication par emails entre les employés d'une entreprise de fabrication ((Michalski et al 2011))	167 employés	82,927 (emails) Liens asymétriques avec l'horodatage: **Timestamp** L'émetteur et récepteurs des emails sont distingués
SNs depuis les Data logs des OSNs		
'Facebook (NIPS)' [68] Réseaux d'amis entre un ensemble de users sur Facebook ((McAuley & Leskovec 2012))	2 888 users	2,981 Liens symétriques d'amitié

'Facebook-like Social Network' [29] ou 'UC Irvine messages' [59]: Réseau des messages envoyés sur Facebook entre users d'une communauté en ligne : des étudiants de l'Université de Californie, Irvin ((Opsahl & Panzarasa 2009)) ((Panzarasa & Opsahl 2009))	1 899	20 296 ou 59 835 liens asymétriques (messages) avec horodatage: **Timestamp**
'Facebook-like Forum Network' [29] étudié dans ((Opsahl 2013)) : le réseau (two-mode network) est extrait à partir la même communauté en ligne précédente mais en se basant sur les messages postés par users pour un 'topic' sur le forum	899 users 522 topics	71 380 Liens pondérés par le nombre de messages\caractères postés par 'user for a topic'
'Facebook wall posts' [60] Réseau d'un petit ensemble de messages postés par des users de Facebook sur les murs des autres users ((Viswanath et al 2009))	46 952 users	46 952 messages sur les murs: Liens asymétriques comprenant des boucles (poster sur son propre mur) et des liens multiples entre une paire et avec horodatage: **Timestamp**
Epinions OSN 'who-trust-whom' ((Richardson et al 2004)) (étudié dans (Jamali et al 2011))	75 879	508 837 liens asymétriques
LiveJournal OSN ((Backstrom et al 2006))	4 847 571	68 993 773 liens asymétriques
'Twitter (MPI)'[74] Réseau des relations 'follow' entre users sur Twitter pendant un snapshot en 2009 ((Cha et al 2010))	52 579 682 users	1 963 263 821 liens asymétriques (follows)
Twitter OSN : tweets collectés entre Juin et Dec 2009 ((Yang & Leskovec 2011))	17 069 982 users	476 553 560 tweets
Slashdot SN (février 2009), ((Leskovec et al 2009)), étudié dans ((Wang et al 2013))	82 168	948 464 liens asymétriques
'BlogCatalog Dataset' [46] ((Agarwal et al 2009)) ((Zafarani & Liu 2009)) : Réseau d'amis crawled en Juillet 2009 depuis BlogCatalog [47]	88 784	4 186 390
'Buzznet DataSet' [46] ((Zafarani & Liu 2009)): Réseaux entre bloggers sur Buzznet[48]	101 168 bloggers	4 284 534 pairs d'amitié
'Flixster Dataset' [46] Réseau d'amitié crawled en Décembre 2010 par Javier Parra ((Zafarani & Liu 2009)) depuis Flixster[50]	2 523 386 utilisateurs	9 197 338 liens d'amitié
'Foursquare DataSet' [46] Réseau d'amitié crawled en Décembre 2010 par Fred Morstatter ((Zafarani & Liu 2009)) depuis Foursquare [51]	106 218	3 473 834 liens d'amitié
'Hyves Dataset' [46] Réseau d'amitié crawled en Décembre 2010 par Mahsa Mojtahedi ((Zafarani & Liu	1 402 611	2 777 419 liens d'amitié

2009))		
Réseau d'affiliation (bipartite) des utilisateurs d'Orkut à leurs groupes [55, 56] (Mislove et al 2007)	11 514 053 nœuds entre users et groupes	327 037 487 liens d'affiliation
'Livemocha Dataset' [46] Réseau d'amitié crawled en Décembre 2010 par Xia Hu (Ben) ((Zafarani & Liu 2009)) depuis Livemocha[52]	104 438	2 196 188 liens d'amitié
'Twitter user– hashtag' [62] ((De Choudhury et al 2010)) (folksonomie) Réseau bipartite des users de Twitter et les tags mentionnés dans leurs tweets (pour tagger des URLs)	175 214 users 530 418 hashtags	4 664 605 liens (usages) avec horodatage: **Timestamp**
Dataset de folksonomie 'crawled' depuis Flickr entre 2006 et 2007(étudié dans le projet TAGora[31])	319 686 users 1 607 879 tags 28 153 045 ressources	112 900 000 assignations
Dataset de folksonomie 'crawled' depuis Delicious entre 2006 et 2007 (étudié dans le projet TAGora[31])	532 924 users 2 481 698 tags 17 262 480 ressources (URLs)	140 126 586 assignations
'Flickr personal taxonomies' [32] : des collections de photos crées par 'user' et 'taggées' hiérarchiquement ((Plangprasopchok et al 2010)), (folksonomie)	7121 Flickr users 7 656 031 tags	
'Any Beat Dataset' [37] ((Fire et al 2012a)) extrait depuis Anybeat.com [38] : communauté d'interaction en ligne	12 645	67 053 liens asymétriques
'Google Plus Dataset' [37] ((Fire et al 2013a)) : Données sont extraites via un 'crawler' depuis Google+ [40]	211 187	1 506 896 liens asymétriques
'YouTube2 Dataset' [46] : Réseaux de contact entre users ainsi que les groupes ((Zafarani & Liu 2009)) ((Mislove 2009)) crawled par ((Tang & Liu 2009a)) ((Tang & Liu 2009b)) depuis YouTube[53]	1 138 499 users 47 groupes	2 990 443 pairs d'amitié
'The Marker Cafe Dataset' [37, 41] ((Fire et al 2011)) ((Fire et al 2013a))	69 411	1 644 848 liens symétriques
'WikiTree Dataset (Multi-graph)' [37] ((Fire & Elovici 2013)) extrait depuis wikitree.com[42] : profils des individus contribués par users formant les généalogies des familles	1 382 751	9 192 212 liens asymétriques
'Gowalla'[24,70] ((Cho et al 2011)) Réseau d'amitié à partir du SN Gowalla[69] basé sur la localisation	196 591 users	950 327 liens symétriques d'amitié
'Brightkite' [24,73] ((Cho et al 2011)) Réseau d'amitié à	58 228 users	214 078 liens

partir du SN Brightkite basé sur la localisation		symétriques d'amitié
'fb_it-2011'[83] ((Backstrom et al 2012)) sous graphe Facebook régional (italien)	17 131 210	3 395 374 924
'Datasets de 6 SNs de 6 organisations sur Facebook' [37] collectés à partir les pages Facebook des employés sur 3 échelles différents ((Fire et al 2013b))	De 500 à 2000 employées De 4 000 à 20 000 Plus de 50 000	Liens symétriques

La première catégorie peut couvrir aussi des réseaux de citations : Ex. le réseau des citations entre 34,546 papiers d'Arxiv HEP-PH (high energy physics phenomenology) et 421,578 arcs entre janvier 1993 et avril 2003[24] (124 mois), ((Leskovec et al 2007)). Ce dataset est étudié par exemple par ((Wang et al 2013)).

Dans 'The Dutch Soccer Team as a Social Network' de ((Kooij et al 2009)), on trouve l'un des SNs exceptionnels qu'on classe dans cette catégorie. Malgré la popularité mondiale du football et son contexte social très intéressant, n'a pas suscité vraiment l'attention des études du point de vu réseaux complexes (SNs). ((Kooij et al 2009)) ont construit un SN de la sélection néerlandaise à partir des données disponibles sur **www.voetbalstats.nl**, qui contient toutes les informations officielles sur les matchs joués par la sélection nationale néerlandaise, ainsi que les matches européens joués par les clubs de la ligue néerlandaise. Chaque nœud correspond à un joueur qui a joué un match officiel dans la sélection. Deux joueurs sont liés si les deux sont alignés dans le même match. Les auteurs ont pris en comptes tous les matchs jusqu'à Juin 2008 ((Kooij et al 2009)). Le 'DST SN' est connecté et affiche l'effet du petit monde avec un coefficient de clustering très élevé. Harry Dénis par exemple a été trouvé le joueur qui a le degré le plus élevé, mais avec coefficient de clustering le plus faible, car ses co-joueurs sont les moins reliés mutuellement.

Dans la deuxième catégorie, certains auteurs sont intéressés aussi par les data logs des appels téléphoniques mobiles qui sont analysés dans certains cas d'étude. Par exemple les appels téléphoniques autrichies proposés par Eric D. Kolaczyk[28]. ((Seshadri et al 2008)) ont analysé les données d'un million d'utilisateurs et dix millions d'appels, en examinant certains facteurs comme la distribution de ces appels par client, etc. ((Seshadri et al 2008)) (Nettleton 2013).

En troisième catégorie, les journaux des messages (tweets) ainsi que les users sur Twitter sont actuellement collectés et analysées par nombreux auteurs (Nettleton 2013) ((Yang & Leskovec 2011)) ((Zafarani & Liu 2009)) ((Gjoka et al 2011)) ((An et al 2011)) ((Ghosh et al 2012)). L'une des premières études quantitatives de 'Twittersphere' dans son intégralité a été proposée par ((Kwak et al 2010)). Les auteurs ont étudié les caractéristiques topologiques de Twitter et son pouvoir en tant qu'un nouveau moyen de partage et diffusion d'information (service de réseautage social et de microblogging), ((Kwak et al 2010)). Récemment, les auteurs sont partis même vers étudier l'efficacité\ influence et les intérêts des utilisateurs sur des SNs de Twitter ((Cha et al 2010)) ((Bhattacharya et al 2014)) ((Babaei et

al 2015)). En outre, les outils de localisation géographique et de proximité physique aident à extraire et analyser des OSNs avec des données supplémentaires intéressantes. Par exemple, ((Caputo et al 2015)) ((Socievole et al 2014)) ont collecté des données 'The unical/ socialblueconn dataset' sur les proximités des appareils (Smartphones) Bluetooth, les profiles sociaux Facebook et certains intérêts de 35 users (étudiants et autres). Les données sont obtenues via une application ad hoc 'SocialBlueConn' installée sur les Smartphones des étudiants à l'Université de Calabre (Italie), ((Caputo et al 2015)) ((Socievole et al 2014)).

Le dernier exemple de dataset a été proposé par ((Fire et al 2013b)) et contient 6 SNs de 6 organisations ou entreprises. En utilisant un algorithme 'crawler' de SN d'organisation ((Fire et al 2013b)), les auteurs ont identifié chaque réseau par un ensemble de profiles utilisateurs sur Facebook, déclarées comme employés (et des liens informels entre eux) d'une organisation donnée.

Tableau 18. SNs de 6 organisations collectés par ((Fire et al 2013b)) et décomposés en communautés d'employés

Entreprise (organisation)	**Description**	**Son SN 'crawled' sur Facebook**		Communautés
		Nœuds (Facebook users)	Liens informels	
Small Hardware Company	Entreprise spécialisée dans le développement de matériel de réseau	165 (en Amérique du Nord & Asie)	726 liens symétriques	Bleu: groupe d'administration et R&D en Asie Rouge : ingénieurs de vérification du matériel et concepteurs de puce en Asie Jaune : R&D de matériels, etc. ((Fire et al 2013b))
Small Software Company	Société internationale spécialisée dans le développement des logiciels ((Fire et al 2013b))	320 (en Amérique du Nord, Europe, Asie, Australie et Moyen-Orient)	2 369 liens symétriques	Bleu: groupe IT au Moyen-Orient. Rouge : groupe R&D au Moyen-Orient Mauve : groupe en Amérique du Nord, etc. ((Fire et al 2013b))
Medium Telecommunication Service Company	Société internationale de technologie spécialisée dans les services de télécommunication ((Fire et al 2013b))	1 429 (En Amérique du Nord)	32 876 liens symétriques	Rouge : la haute direction Jaune: Les consultants internationaux et les ingénieurs d'assistance Vert : Siège social nord-Américain, etc. ((Fire et al 2013b))
Medium Software Provider and Outsourcing Company	Fournisseur international de logiciel et d'externalisation spécialisé en services	3 862 (axé sur la zone : l'Asie du sud).	87 324 liens symétriques	Rouge : R&D et SDE liés à l'Australie, Europe et en Amérique du Nord Jaune : R & D et la SDE liés à l'Afrique, l'Amérique du Nord,

	de télécommunication et serve une base de clientèle mondiale ((Fire et al 2013b))			et l'Asie Bleu : R & D et la SDE liés aux employés nord-américains et asiatiques…etc. ((Fire et al 2013b))
Large Information Technology Corporation	Société de technologie de l'information fournissant des produits et services aux clients autour du monde ((Fire et al 2013b))	5 793 (Dans l'Amérique du nord et du Sud, en Asie et en Europe de l'Est)	45 266 liens symétriques	Mauve: Ingénieurs d'assistance et consultants (marketing, des ventes et prix) de l'Europe de l'Est Noir : Succursale nord-Américaine et R&D de l'Est asiatique, etc. ((Fire et al 2013b))
Large Technology Corporation	société fournissant des produits en matériel, logiciels et infrastructures et d'autres services de technologie à des clients internationaux ((Fire et al 2013b))	5 524 (Dans l'Amérique du nord et du Sud, en Asie et en Europe de l'Est)	94 219 liens symétriques	Rouge : La haute direction internationale (cadres supérieurs, chercheurs seniors) Vert : Équipe de sport amateur de l'entreprisen etc. ((Fire et al 2013b))

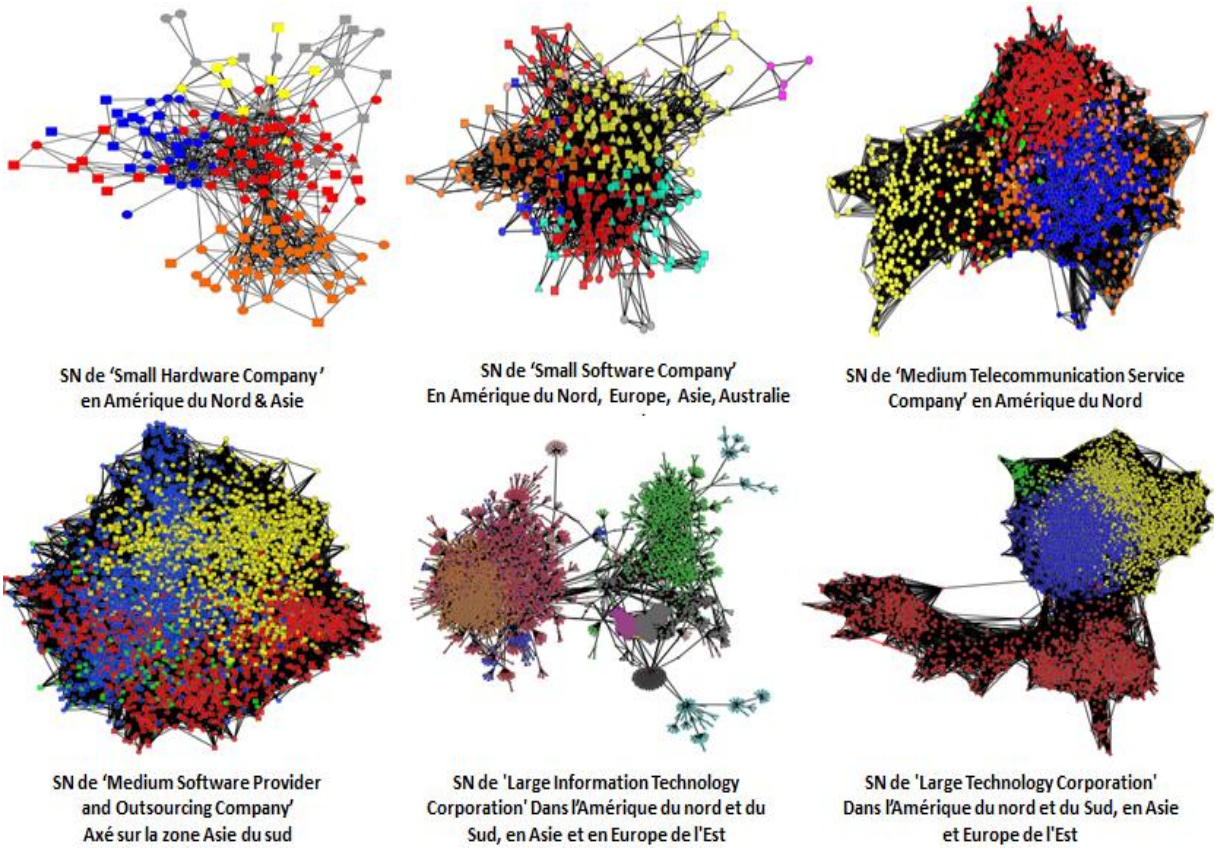


Figure 22. SNs & communautés (visualisés par Cytoscape[43]) de 6 organisations 'crawled' par ((Fire et al 2013b)) via les profils utilisateurs des employés et leurs liens informels sur Facebook

Des communautés sur chaque réseau sont découvertes par l'algorithme de (Newman & Girvan 2004) implémenté sur Cytoscape [43] ((Shannon et al 2003)). Chaque communauté est marquée par une couleur différente. La forme des nœuds est également significative, car elle indique si ou non l'employé correspondant occupe une position de gestion: Les nœuds triangles représentent ceux qui occupent des postes de gestion. Les carrées représentent ceux qui n'occupent pas tels postes cependant les cercles représentent des employés dont la position est inconnu dans l'organisation. Selon ((Fire et al 2013b)), la plupart de ces users (employés) ne divulguent pas leurs positions dans une organisation sur leurs pages de profils. De ce fait, les auteurs identifient les rôles leadership dans l'organisation via différentes mesures de centralités ((Fire et al 2013b)) appliquées sur le SN correspondant. Par ce moyen, les nœuds (les profiles utilisateurs) sont classés, tel que le top 20 permet d'inférer les employés occupants les positons de gestion (direction) dans l'organisation.

4.4. Pertinence et richesse des données sociales (des SNs\ OSNs)

Tous ces services et plateformes permettent aux utilisateurs de maintenir leurs relations établies déjà en monde réel ou développer des nouvelles relations et affiliations en ligne. L'émergence des OSNs et la disponibilité croissante des données sociales en ligne ont donné l'impulsion au domaine de recherche de SNA via des nouveaux chercheurs comme Kleinberg, Kumar, Mika, etc., ajoutés aux anciens. Outre, les données sociales (en ligne) constituent maintenant un service pour l'apprentissage automatique, Data Mining, et les communautés des sciences sociales, etc. Même si l'anonymat des données est importante pour

protéger la vie privée des utilisateurs sur les OSNs, les datasets de la 2ème et 3ème catégorie (Tableau 17) sont maintenant de plus en plus variés: Nombreuses plateformes et outils, technologies, fonctionnalités, différents types d'utilisateurs, mode de fonctionnement et usage, des idéologies, etc. De ce fait, les études ne doivent pas se contenter seulement de fouiller dans des comportements sociaux d'utilisateurs en ligne sans des motifs sérieux et pertinents. ((Mika 2005a)) est allé plus loin et souligne parfois le caractère incomplet de certains de ces SNs en raison de l'absence de certaines composantes en ligne par rapport aux réseaux 'offline'. Par exemple, ((Viswanath et al 2009)) ont trouvé que la représentation structurelle des OSNs comme le graphe d'amis sur Facebook de New Orleans' (USA) n'est pas toujours la vraie image des amis réels pour un individu. Selon eux ((Viswanath et al 2009)), beaucoup d'utilisateurs ne sont pas très discriminatifs quand ils créent des liens avec des personnes comme des amis. De ce fait, ((Viswanath et al 2009)) ont voulu extraire un réseau qui donne une meilleur image de qui communique avec qui en proposant une mesure de « Activity ». Elle mesure l'intensité de "activité" qui est proportionnelle à la force de la relation. L'activité qui a été choisie est 'Writes to Wall' (Facebook Wall posts), ((Viswanath et al 2009)). Les conclusions sont prometteuses. Cependant, les données ont été géographiquement limitées. En outre de nombreux autres canaux de communication peuvent être également utilisés pour estimer la force (la pertinence) réelle de la relation sur Facebook.

En conséquence, les notions de pertinence et de richesse des données sociales sont profondes et ambitieuses. ***La pertinence est liée au contexte (l'environnement) qui aide à approfondir la définition de l'objet ou sujet du SN, quels acteurs et pour quels usages ? Les liens sont-ils pertinents ? Il s'agit d'un point de départ crucial pour que les études analytiques et les interprétations soient plus bénéfiques.*** 80% des données collectées aujourd'hui sont spatialement référencées d'une manière ou d'une autre (Reda et al 2009). Plusieurs types de comportement social sont directement influencés par la position spatiale des individus et de leur circulation dans l'environnement (Reda et al 2009).

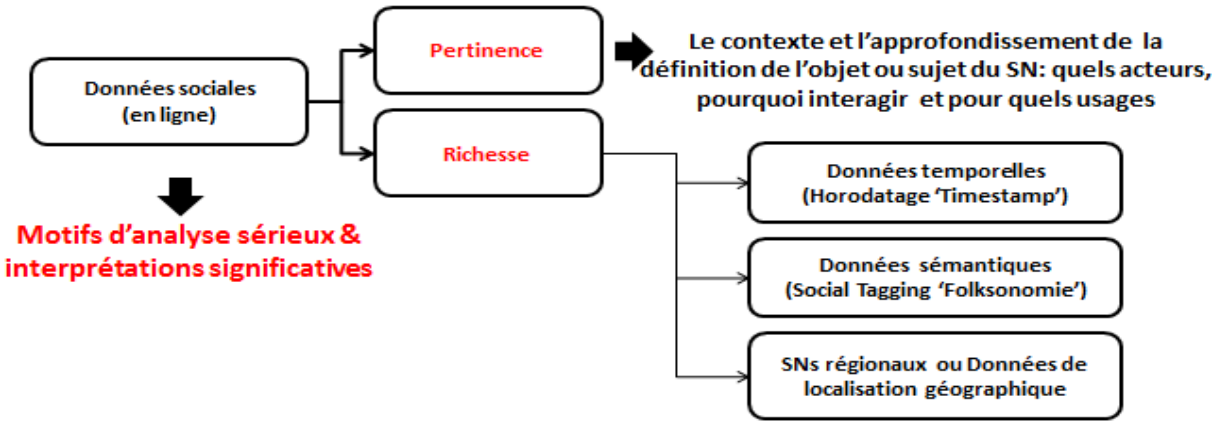


Figure 23. Pertinence et Richesse des données sociales (en ligne) visant des motifs d'analyse sérieux et des interprétations significatives

Les données sociales (en ligne) au sein des contextes organisationnels et des environnements spécifiques: Organisations économiques, politiques, sport, environnements d'apprentissage, réseaux d'innovations, etc., sont très précieuses pour fouiller les organisations et explorer les

interactions humaines, les rôles leaderships, la sécurité interne, la lutte antiterroriste (les données de CASOS[81]), etc. Dans ces contextes, les interactions sociales peuvent être plus orientées et contraintes, car les obligations sociales et institutionnelles sont fusionnées.

Cependant, les interactions sur les plateformes sociales d'OSNs sont plus ou moins pertinentes et il y a beaucoup de liens de connaissances éphémères. Mais le dernier exemple dans le Tableau 17 illustre des OSNs plus pertinents malgré ils sont 'crawled' à partir d'une plateforme sociale publique comme Facebook. Dans ce cas, le Crawling effectué est lui-même programmé pour cibler et extraire des OSNs entre des profils utilisateurs qui sont des employés dans des organisations. En plus, ces données pourraient être également pertinentes si elles couvrent des périodes durant certains évènements du monde réel impliqués derrières les OSNs. Par exemple, Spinn3r.com fournis un dataset de plus de 386 millions de blog posts, articles de presse, annonces, forum posts et des contenus de médias sociaux qui couvrent la période entre 13 janvier et 14 février (2011) : Il s'agit d'une période connue par ses évènements politiques comme la révolution tunisienne et les protestations en Egypte. Par ailleurs, il y a même des sites web qui ne sont pas strictement d'OSNs mais permettent aux utilisateurs d'exprimer leurs intérêts et d'évaluer des films, musiques, documents (Des réseaux biparties). En conséquence, on ne peut pas affirmer toujours que les heuristiques ne seront pas nécessaires pour inférer des SNs même si toutes ces interactions sont explicitement déclarées. Les heuristiques et l'inférence restent importantes pour extraire des SNs plus pertinents.

Récemment l'horodatage des interactions, leurs orientations, pondérations ou sémantique, les intérêts des acteurs, etc., sont devenus des données (données temporelles ou sémantiques) importantes pour enrichir les études analytiques et avoir des résultats plus réalistes et significatifs. **Cette richesse permet de rendre les données des SNs\OSNs de plus en plus pertinents.** En outre le contexte géographique et les données de localisation géographiques sont également des données récentes à ajouter et à exploiter: Réseau Orkut (inde) (Mislove et al 2007), Pokec (slovaque), ((Takac & Zabovsky 2012)), Hyves (Pays-Bas) ((Zafarani & Liu 2009)), Libimseti.cz (Tchèque), ((Kunegis et al 2012)) ((Brožovský & Petříček 2007)), Des sous-graphes régionaux (italien, suédois, USA, etc.) de Facebook ((Backstrom et al 2012)), Gowalla'[24,70] ((Cho et al 2011)), Brightkite' [24,73] ((Cho et al 2011)), etc.

Ces notions de pertinence et de richesse seront incarnées dans nos contributions au niveau des données sociales et les motifs d'analyse.

Des problématiques

Il est important d'établir des datasets standards, sur lesquelles des algorithmes et des propositions sont évaluées. Cela permet de comparer, contrôler et rapprocher les travaux. Cependant, les données sociales qui sont plus riches en termes de données temporelles ou sémantiques ne sont pas accessibles facilement. Ce n'est pas le cas de beaucoup de datasets ordinaires qui représentent juste des ensembles de nœuds (acteurs) et de liens (interactions). Les données des opérateurs et plateformes d'OSNs qui sont accessibles via des APIs sont souvent des graphes sociaux classiques. Peu de données temporelles (l'horodatage) comme le

temps de création d'un lien entre deux utilisateurs, le temps de suivi en Twitter : user u commence à suivre v (Meeder et al 2011), etc., sont publiquement accessibles ou bien elles sont soumises à des restrictions de confidentialité (Nettleton 2013). Bien que beaucoup de bases de données (Facebook ou autres) contiennent des nombreux attributs nodaux (profils des utilisateurs), ces informations ne sont pas disponibles.

A l'heure actuelle, il n'est pas évident d'obtenir des données originales décrivant la dynamique temporelle ou la sémantique des SNs qui sont étudiés par des expériences et travaux précédents. La plupart de ces données se trouvent sous différents formats, parfois liées à des outils d'analyse. Leurs auteurs sont eux même limités par des conditions et des accords de partage et d'utilisation des données pour des raisons de confidentialité, etc., qui sont imposées par les sources. Par exemple ((Backstrom et al 2012)) ont étudiés des datasets de sous-graphes régionaux (limité géographiquement et temporellement) [83] qui sont extraits de Facebook mais ils ne les partagent pas. Selon eux, ces données ne peuvent pas être distribuées pour des raisons évidentes. Seulement des données et des caractéristiques agrégées sont accessibles et téléchargeables (Il y a de peu de sources de données pour des réseaux dynamiques).

Par ailleurs, si les jeux de données sociales portent plus de richesse, ils seront probablement modifiés par des groupes de chercheurs intéressés par des attributs différents. Il n'y a aucune garantie que ces données sont identiques à leur description dans un document, citation ou sur le net. Par exemple, le réseau 'email-EuAll' cité sur SNAP[24] n'inclue pas la composante temporelle bien qu'il a été étudié par ((Leskovec et al 2007)) en abordant l'évolution des graphes.

5. Applications et software et format des données pour SNA

La fouille et l'analyse des données sociales peuvent se faire par des outils d'analyse des réseaux complexes, particulièrement via des applications et modules spécifiques pour les SNs (leur dynamique). Pratiquement, il existe une panoplie de Softwares, packages et applications qui sont utilisés pour analyser et visualiser les SNs sous différents formats de données. Parmi les différentes versions d'applications, logiciels, les plus populaires, gratuits ou commerciaux normalisés, on trouve :

- **Gephi[114]** : Gephi ((Max & Kathleen 2005)) permet de visualiser des graphes et calculer leurs différentes caractéristiques statistiques avec des métriques standards : centralité (Closeness, betweenness) des nœuds, longueur de chemins, diamètre, densité, coefficient de clustering, HITS, modularité, étiquetage des communautés, etc. (Nettleton 2013). Il est compatible avec des réseaux sous plusieurs formats de données : dl, net, gml, csv, etc. Il supporte également des réseaux dynamiques.

- **Pajek[100]** : Pajek ((Batagelj & Mrvar 2012)) ((Batagelj & Mrvar 2003a)) ((Batagelj & Mrvar 1998)) ((Beauguitte 2011)) est un programme d'analyse et de visualisation des réseaux (Figure 24). Il facilite la réduction et l'analyse des grands réseaux (plus d'un million de nœuds) en implémentant des méthodes sophistiquées (Shrinking) et des algorithmes efficaces. Pajek offre également des outils puissants de visualisation. La

structure de Pajek est entièrement basée sur six structures de données et les transitions entre ces structures ((Huisman et al 2011)): Réseaux (nœuds, arcs/ arrêtes), partitions (classification des nœuds, chacun est assigné à une seule classe), permutations (pour réordonner les nœuds), clusters, les hiérarchies (clusters et nœuds hiérarchiquement ordonnés), les vecteurs (propriétés de nœuds). Le format de données (.Net) est l'un des formats supportés par Pajek. Le réseau se présente par une liste des nœuds et des arcs\ arètes, ce qui permet d'introduire des grands réseaux à étudier et manipuler (Figure 24).

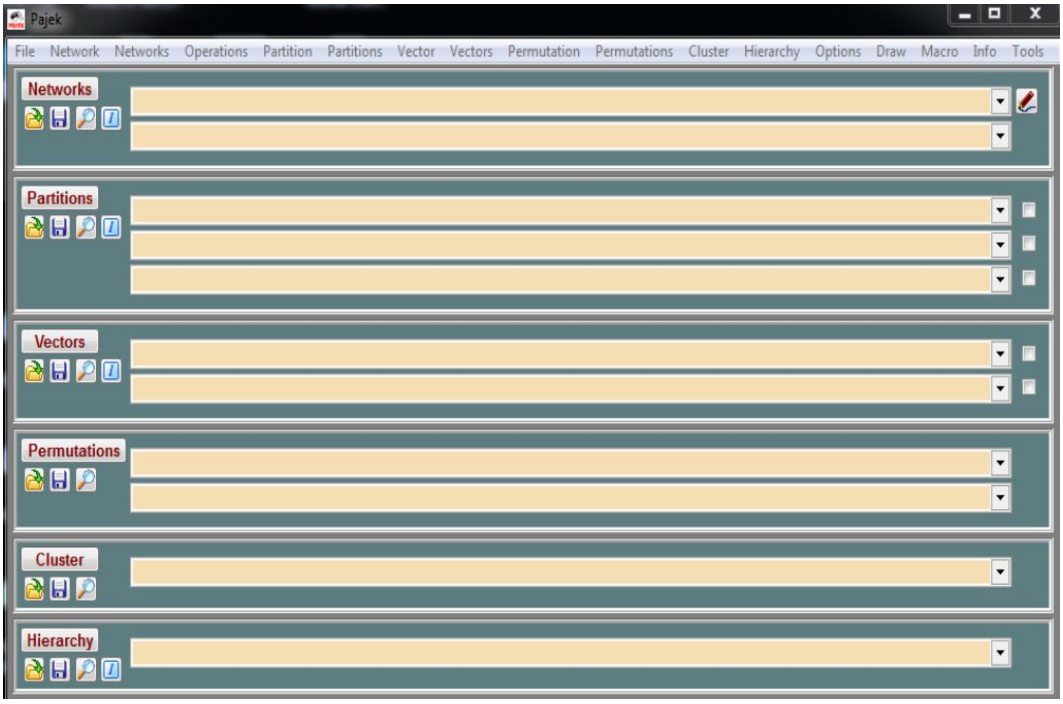


Figure 24. Interface principale de Pajek ((Batagelj & Mrvar 2012)) ((Batagelj & Mrvar 1998)) ((Beauguitte 2011)) ((Batagelj & Mrvar 2006)), (Version 3.08)

Pajek offre plusieurs options et opérations (métriques et algorithmes) pour manipuler le réseau selon ses différentes structures de données : Centralités (Closeness, betweenness), détection des composantes (cliques, clusters, etc.), chemins, opérations binaires sur deux réseaux, etc. Le réseau peut être aussi transposé (cas orienté), réduit par des classes ('Shrinking'), etc., la création des objets comme les clusters, etc. Pajek facilite également l'analyse des réseaux dynamiques dans le temps. Les propriétés de visualisation graphiques en Pajek sont avancées et offrent de nombreuses options pour visualiser et manipuler les données réseaux : A partir des simples 'layouts' (circle, Random) jusqu'à des algorithmes comme 'Kamada-Kawai' ((Freeman 2004)) ((Everton 2002)), 'Fruchterman-Reingold' ((Huisman et al 2011)). Cependant, Pajek contient seulement quelques procédures statistiques de base sur les attributs des nœuds (partitions \vecteurs) permettant de calculer des corrélations, des régressions linaires, etc.

- **UCINET**: C'est un programme créé pour comprendre et analyser les SNs et autres types de données ((Huisman et al 2011)) ((Borgatti et al 2002)). Selon ses

développeurs: Borgatti, Everett et Freeman ((Borgatti et al 2002)), *UCINET 'is built for speed, not for confort'*. Les datasets sur UCINET se présentent généralement sous forme de matrices. Il permet d'importer les données des réseaux sous des formats comme DL, Excel ainsi qu'autres formats de certains programmes comme Pajek[100], NEGOPY[111], etc. ((Huisman et al 2011)). UCINET inclus une collection de datasets standard (UCINET IV Datasets[78]) accompagnés aussi par les données Linton C. Freeman[75]. Il fournit un grand nombre d'outils de gestion et de transformation de données. Il peut traiter des données de type 'Two-mode' (réseaux d'affiliation) et dériver des données 'one-mode'.

UCINET contient des outils graphiques permettant de dessiner des nuages de points, des dendrogrammes et des diagrammes d'arbres. Cependant, le programme lui-même ne propose pas des procédures pour visualiser graphiquement les réseaux. Il fait appel plutôt à d'autres programmes de visualisation comme NetDraw[116] qui possède des propriétés graphiques avancées pour visualiser les réseaux ((Borgatti 2002)). UCINET contient un grand nombre de routines d'analyse de réseau, par exemple la détection des sous-groupes cohésifs (cliques, clans, plex, etc.), l'analyse des centralités, des clusters et les chevauchements de cliques. Des routines statistiques sont également disponibles comme la corrélation de Spearman ((De Nooy et al 2004)), le coefficient Jaccard, etc. Dans les versions récentes, l'option de calcul de centralité d'un groupe donné ((Everett & Borgatti 2004)) est ajoutée. En fixant la taille des groupes, le programme peut trouver le sous-groupe le plus central ((Huisman et al 2011)). En outre, le montage **des modèles de noyau / périphérie** ((Borgatti & Everett 2000)) sont possibles (voir plus loin).

- **NetMiner[92]**: C'est une application logicielle commerciale (pour Windows) basée sur java. Elle permet d'analyser et visualiser des réseaux en incluant des modules spécifiques pour analyser des données de Twitter (Nettleton 2013). C'est un outil qui combine le SNA et des techniques d'exploration visuelle interactives ((Huisman et al 2011)). Il permet de détecter des patterns et des structures sous-jacents dans le réseau. Un modèle optimisé des données de réseau est adopté par NetMiner en combinant 3 types de variables ((Huisman et al 2011)) : Matrice d'adjacence (layers), variable d'affiliation, attributs d'acteur. Les formats de datasets que NetMiner peuvent supporter (enregistrer\ exporter) sont : NTF- files, CSV (Excel), UCINET DL.

  NetMiner possède également des propriétés de visualisation graphiques avancées. La plupart des résultats se présentent textuellement (rapport d'analyse) et graphiquement au même temps contrairement à plusieurs autres programmes qui obligent l'utilisateur de demander la visualisation de certaines analyses ((Huisman et al 2011)). Cependant et comme plusieurs applications, des options et des algorithmes de visualisation connus, ex. l'algorithme Kamada-Kawai ((Everton 2002)) ((Freeman 2004)), sont implémentés dans NetMiner. Des visualisations 3D sont aussi supportées. NetMiner propose différentes méthodes (des routines) pour analyser la connectivité du réseau, les configurations des sous-graphes (cliques, clans, etc.) et pour calculer les mesures de centralités. Des options sont proposés pour appliquer ces centralités sur des graphes

orientés (Ex : 'In / Out -Closeness centrality') ((Huisman et al 2011)). On trouve également des routines statistiques standards : des statistiques descriptives, corrélations, régressions, etc. En outre, NetMiner fournit des tests de simulation avec des méthodes Monte Carlo et chaînes de Markov pour plusieurs mesures de réseau ((Huisman et al 2011)).

D'autre part il y a des bibliothèques (Packages) et des interfaces (APIs Java) pour le développement des applications et logiciels à la disposition des programmeurs comme :

- JUNG[95] (Java Universal Network/Graph Framework): C'est une bibliothèque (Open source) en Java qui offre un cadre de modélisation, analyse et visualisation.
- Neo4j [93] : C'est un système de base de données de graphes pour le traitement à haute performance qui dispose d'un API Java pour les besoins de «Big Data» (Nettleton 2013).
- Gephi dispose également d'une interface API Java.
- NetworkX [94]: C'est un package en Python utilisé pour l'exploration des structures de réseaux complexes et leur dynamique, etc., ((Hagberg et al 2008)). Il est muni de plusieurs mesures, algorithmes standards d'analyse et des générateurs de graphes aléatoires et réseaux synthétiques, etc.
- Igraph[96] : C'est une bibliothèque API en langage C, une collection 'open source' gratuite d'outils d'analyse de réseaux. Elle peut être programmée même en R, Python et C / C ++.
- SNAP[24] (The Stanford Network Analysis Platform\ Procedures): C'est un système de haute performance, écrit en C++, pour l'analyse et la manipulation des grands réseaux (Nettleton 2013).

Outre, KONECT [55] est l'un des projets qui recueille une large gamme de réseaux et développe des outils d'analyse, utilisés pour calculer les caractéristiques/ statistiques des réseaux. ((Huisman et al 2011)), ainsi que plusieurs d'autres auteurs ((Xu et al 2010)) ((Loscalzo & Yu 2008)) ((Kirschner 2008)) ((Handcock et al 2008)) donnent un aperçu sur les logiciels et packages utilisés pour SNA. Ils sont classifiés selon différents points de vue. Dans le tableau suivant la catégorie des applications qui supportent l'analyse et la visualisation des réseaux dynamiques sera distinguée.

Tableau 19. Quelques catégories de logiciels et applications dédiés à l'analyse et la visualisation des SNs

	Académiques \ free	**Commercial \ Non-free**
Logiciels & packages pour analyser et visualiser les réseaux (SNs)	**Agna**[97]: 'Applied Graph & Network Analysis' **Dynet** [97]: (SE & LS): visualisations guidés par les données **GUESS**[97]: 'Graph Exploration System' **NetVis**[98]: Visualisation dynamique	**Blue Spider**[97]: Analyse des réseaux. **InFlow**[26]: 'Network mapping' **Mdlogix Solutions**[97]: VisuaLyzer,

	des SNs **Network Workbench**[99] **:** Analyse, modélisation et visualisation **Sentinel Visualizer**[101]: Analyse & visualisation des liens **SocNetV**[102]: 'Social Networks Visualiser'. **Pajek**[100], **UCINET 6** (version d'évaluation), **Gephi**[114], **Visone**, etc. **libsna**[103] (Python) : Bibliothèque open-source pour SNA **NodeXL** (Excel): Visualisation & analyse des graphes réseaux. **Igraph**[96], **JUNG**[95], **NetworkX**[94], **SNA**[97] **(R)**	LinkAlyzer, EgoNet' **UCINET, NetMiner3** [92], **SNAP**[24], etc.
Logiciels & packages plus spécialisés	**CFinder**[105] : Recherche et visualisation des groupes denses (Clusters & communautés qui se chevauchent). **KeyPlayer**[107]: Identification des nœuds **SIENA**[97], **StOCNET**[108], **statnet**[109] **:** Analyse statistique **Tnet**[97] (R): Analyse des réseaux pondérés et longitudinaux **NEGOPY**[111]: 'Cohesive subgroups' Un des logiciels les plus anciens d'analyse de réseau, etc.	**ONA surveys**[110]: 'Organizational Network Analysis survey tool' **MatMan**[97] (Excel): Analyse structurale, etc.
Logiciels spécialement pour la visualisation	**aiSee**[112], **Graphviz**[97]: Visualisation des graphes **Apache Agora**[113]: Visualisation des communautés virtuelles **Cytoscape**[43]: Visualisation des réseaux d'interactions moléculaires **KrackPlot**[115]: programme de	**yFiles**[104] (Java): Visualisation des réseaux **KeyHubs, TouchGraph**, etc.

	visualisation de réseau social **NetDraw**[116]: Programme associé à UCINET **OGDF**: 'Open Graph Drawing Framework' **Tulip, uDraw(Graph), Zoomgraph**, etc.	
Logiciels pour les réseaux dynamiques	**ORA**[97]: Analyse des réseaux dynamiques **Blanche**[97] : Dynamique des réseaux **Commetrix**[106]: Analyse & visualisation des réseaux dynamiques **SoNIA** : Analyse & visualisation des réseaux longitudinaux (Social Network Image Animator)	

La figure suivante montre un nuage de logiciels/ packages : 'free'/ payants (de la gauche vers la droite du haut vers le bas respectivement) qui sont les plus utilisés dans l'analyse et la visualisation des SNs.

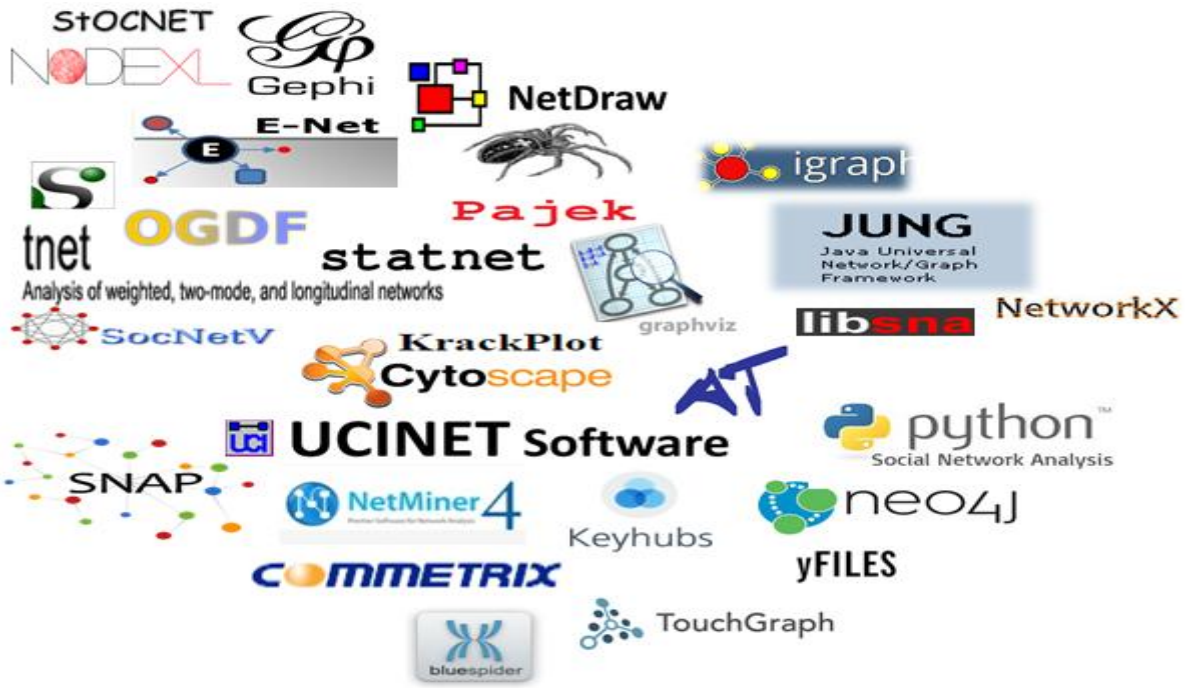


Figure 25. Nuage de logiciels et packages les plus utilisés dans l'analyse et la visualisation des SNs

6. Applicabilité, autres concepts, nouveaux aspects et tendances de SNA

La modélisation et l'analyse des interactions sociales possèdent un large éventail d'applications non seulement en sciences sociales et informatiques. L'applicabilité pluridisciplinaire, l'efficacité des travaux élaborés et les conclusions obtenues via le SNA ont montré leurs bénéfices dans diférents domaines et fourni des réponses devant plusieurs problèmes.

((Srivastava & DeLong 2013)) parlent du ***succès de l'application de SNA/SNAM dans l'innovation et l'impact dans les organisations*** : La découverte d'expertise dans les organisations, les équipes d'intervention rapide dans la gestion des urgences, et en commerce (business) en général, etc. En réalité, les entreprises ne comprennent pas le graphe social de leurs clients. Pour, l'innovation, il ne s'agit pas juste de définir la façon dont les entreprises se rapportent à leurs clients. Le SNA peut déverrouiller un immense bénéfice en comprenant les SNs entre eux, les influenceurs clés, la force des relations. Ainsi, l'impact se voit à travers les pistes de réflexions qui sont dérivés: l'acquisition et la conservation de la clientèle, la recommandation sociale, marketing basée sur l'influence, l'identification des trends setters ((Srivastava & DeLong 2013)). La marque Levis' par exemple essaye de bénéficier des technologies d'analyse sociale (prédictives) pour comprendre la valeur du SN du client ((Srivastava & DeLong 2013)).

Les idées autour des centralités en SN peuvent servir, à titre d'exemple, à explorer l'influence et la puissance des banques dans le monde des entreprises américaines. Plusieurs autres applications ont étendu le SNA sur : Les réseaux politiques, les mouvements sociaux, réseaux religieux, la modélisation des maladies et leur propagation (l'épidémiologie), la transmission d'informations culturelles, la gestion des affaires, l'écologie de comportement (Berger-Wolf & Saia 2006), la criminalité, le terrorisme, réseaux P2P, etc., ((Scott 2011), Social network analysis: developments, advances, and prospects). Les services P2P sont devenus aussi des nouvelles plateformes sociales, permettant par exemple de prêter l'argent directement depuis un autre individu ((Ceyhan et al 2011)). Comme ça, l'intermédiaire d'une institution financière est évité et les conséquences de perte résultantes d'un défaut sur un prêt seront supportées directement par le préteur lui-même ((Ceyhan et al 2011)). C'est l'exemple du site des enchères comme Prosper.com, là où les interactions sociales sont possibles, permettant aux individus d'accéder aux informations (le passé, les scores de crédits, taux d'intérêts, etc.) sur les emprunteurs et influent sur leur décision. Ici, une analyse a été effectuée par ((Ceyhan et al 2011)) qui ont étudié les facteurs influant sur le processus d'un appel d'offre (la dynamique des enchères dans un service de prêt P2P) pour prédire le succès de prêt ((Ceyhan et al 2011)). Par ailleurs, l'analyse des SNs des blogs politiques montre l'efficacité du SNA dans la recherche dans les blogs (politiques) et les aspects d'interactions ((Rosen et al 2011), Social networks and online environments: when science and practice co-evolve). Outre, les travaux d'analyse des systèmes organisationnels et sociaux effectués par le centre CASOS[81] sont actuellement exportables vers des chercheurs et experts de la lutte antiterroristes ((Frankenstein et al 2015)) ((Fellman et al 2011)) ((Il-Chul et al 2008)) ((Il-Chul & Kathleen 2007)) ((Max & Kathleen 2005)).

Les besoins d'analyse sont également évolués, requérant le développement des méthodologies, techniques, algorithmes et des outils plus puissants pouvant algorithmiquement passer à l'échelle, tant en masse de données qu'en débit. Maintenant, les interactions sont animées également autour des contenus partagés dans les OSNs. Ce sont des informations supplémentaires qui aident à comprendre mieux des aspects importants de la socialisation moderne de l'homme. Des études ont montré que des SNs comme YouTube ou Twitter présentent des caractéristiques ('Reciprocative linking', homophilie, etc.) qui les distinguent des SNs (OSNs) traditionnels ((Wattenhofer et al 2012)). Par exemple, la tendance des individus à s'associer avec d'autres individus similaires est remarquablement plus claire. D'autre part, un OSN axé sur le contenu comme Twitter montre que la popularité d'un utilisateur n'est pas acquise seulement par les relations mais également à travers le contenu. L'étude de certains aspects sociaux et le contenu d'un user trouve qu'il y a une forte corrélation entre la popularité sociale d'un user et leur contenus populaires ((Wattenhofer et al 2012)) (Nettleton 2013).

La richesse informationnelle des données sociales (OSNs) dans les relations, les contenus : Localisation géographique, horodatage, ordre chronologique, type sémantique, tags, etc., ***stimule l'enthousiasme pour enrichir les motifs d'analyse bien motivés, en donnant des nouvelles dimensions d'analyse*** ('Hot Topics' de SNA). Face aux enjeux scientifiques et sociétaux, ces nouveaux aspects, dimensions et tendances en SNA vont fournir également des réponses à des nouveaux rôles ou encore des nouveaux 'end users' : 'Community manger', 'SN manager', etc. La gestion, la surveillance et la supervision des SNs (voir même la supervision automatique) sont des nouveaux sujets qui peuvent bénéficier aussi du développement de SNA. Dans une recherche récente, le groupe 'Social computing' [54] de 'Max Planck Institute for Software Systems' (mpi-sws) a développé des méthodes de prévention de spam dans les SNs (Twitter) et propose un cadre pour le suivi de la relation entre l'utilisateur et sa communauté, etc.

Hiérarchie sociale

La hiérarchie sociale (Stratification Sociale) est à l'origine une notion étudiée en sociologie et en commence et évoquée dans les OSNs. Elle se réfère à l'arrangement hiérarchique des individus dans une société, même si elle est implicite (Gupte et al 2011). Il s'agit d'un nombre de 'strats' qui offre une meilleure possibilité pour déduire le rang d'un nœud (une chose qui n'est pas observable directement) (Gupte et al 2011). Les gens les plus hauts ont un statut social plus élevé que les autres les plus bas dans la hiérarchie. Cette division est basée sur des facteurs comme la puissance, la richesse, la connaissance ou encore l'importance (Gupte et al 2011): Qui joue un rôle important dans une société ? Gestionnaires, décideurs (Gilbert et al 2010). Les orientations des liens entre les entités sociales peuvent servir à déduire cette hiérarchie (Gupte et al 2011).

(Gupte et al 2011) ont défini une mesure A(G) ('Agonie') pour découvrir la hiérarchie H(G) dans un OSN orienté, modélisé par un graphe G (V, E) (Gupte et al 2011). Ils ont proposé un algorithme qui s'exécute en temps polynomial pour évaluer cette mesure (Gupte et al 2011). La hiérarchie se représente par une structure d'arbre (Graphe Acyclique 'Directed Acyclic Graph' : DAG). La mesure de la hiérarchie est entre 0 et 1 de telle sorte que si H(G)

= 1, G est un 'DAG' (Gupte et al 2011). Par exemple sur un réseau de journalistes sur Twitter (961 users) la mesure de la hiérarchie calculée donne 0,38. En conséquence, les auteurs affirment qu'il existe une hiérarchie moyenne dans ce graphe (en 7 niveaux 'strates').

Les auteurs ont montré qu'il y a une forte corrélation entre cette mesure et des indices populaires comme le PageRank (Gupte et al 2011). Les personnes qui ont un PageRank élevée ont une tendance d'avoir un niveau plus élevés la hiérarchie sociale calculé par cette mesure (Gupte et al 2011). L'importance de l'orientation des liens dans la hiérarchie du réseau est montrée par (Gupte et al 2011) en se basant sur le réseau de 'US College football' (Girvan & Newman 2002) dans sa version orientée: (u, v) signifie que 'v' a joué et a battu 'u'. Dans ce cas, la mesure de la hiérarchie permet de déduire le nombre de victoires record (Gupte et al 2011). D'autre part, le développement des algorithmes de classification peut bénéficier d'une telle mesure.

Par ailleurs, des conclusions intéressantes sont obtenues, montrent l'effet de la taille du réseau sur la variance de la hiérarchie et le nombre de ses 'strats'. Ces deux paramètres sont étudiés sur des graphes générés aléatoirement par rapport à des OSNs : Delicious, LiveJournal YouTube, Flickr (Gupte et al 2011). La hiérarchie est influencée par la taille des OSNs qui évoluent avec le temps contrairement aux graphes aléatoires. D'autre part, le degré de stratification (Nombre de 'strats') n'affiche pas une augmentation significative devant l'évolution de la taille d'un OSN (Gupte et al 2011).

L'information géographique

La prolifération des média sociaux accompagnée par les outils développés (Appareils et téléphones mobiles, Bluetooth, GPS, etc.) offrent récemment des SNs basés sur des régions géographiques : Orkut (inde) : (Mislove et al 2007), Pokec (slovaque) : ((Takac & Zabovsky 2012)), Hyves (Pays-Bas) : ((Zafarani & Liu 2009)), Libimseti.cz (Tchèque) : ((Kunegis et al 2012)) ((Brožovský & Petříček 2007)), etc., ou encore des SNs basés sur la localisation 'LBSNs' (location-based SNs) (Nettleton 2013). Les LBSNs offrent des services de localisation permettant aux users de se retrouver à des emplacements géographiques et partager leurs expériences avec leurs amis. Les données de localisation géographique (Via des applications de Smartphones) sont maintenant en vogue. C'est un type de données intéressant à intégrer dans les SNs et à exploiter dans le SNA (des SNs plus pertinents). L'incorporation de l'information géographique (géolocalisation) dans les modélisations, les approches et métriques d'analyse aide à fournir des interprétations plus significatives. On trouve par exemple, des sous-graphes Facebook régionaux (italien, suédois, USA, etc.) qui sont étudiés par ((Backstrom et al 2012)). ((Viswanath et al 2009)) ont collecté et analysé un 'Data-log' spécifique de Facebook qui correspond à la région géographique de 'New Orleans' (USA). Ils ont étudié l'évolution d'interactions entre users ((Viswanath et al 2009)). Par ailleurs, ((Gao et al 2012)) présentent l'illustration d'une analyse basée sur la localisation, l'analyse des LBSNs.

Des nouvelles métriques sont définies montrant comment la distance géographique affecte la structure sociale ((Scellato et al 2010)). En se basant sur des OSNs à grande échelle avec la localisation géographique comme Gowalla'[24,70] ((Cho et al 2011)), Brightkite' [24, 73]

((Cho et al 2011)), FourSquare, etc., des conclusions intéressantes ont été tirées. Des études de ((Scellato et al 2010)) prouvent qu'une grande proportion d'users qui se trouvent sur des distances courtes et qui forment des clusters dans le réseau sont en effet des amis géographiquement proches (Nettleton 2013). La figure suivante montre que la distance physique ou bien géographique est un facteur majeur au niveau de la création et le maintien des relations entre personnes (en vert) notamment dans les OSNs. Tandis qu'elle est plus ou moins influente sur les liens entre des instituts ou organisations (en rouge).

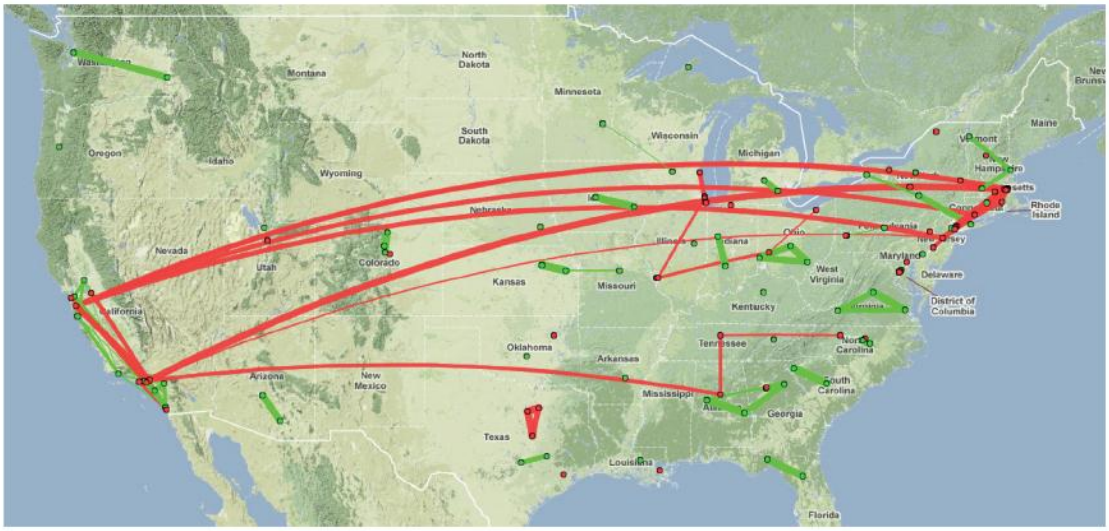


Figure 26. L'impact de la distance physique sur la création et le maintien des liens entre les personnes et organisations (Top US Colleges) ((Gjoka et al 2011))

Les OSNs tendent d'avoir plus de triangles géographiquement confinés par rapport à des SNs basés sur le contenu et le partage comme LiveJournal, Twitter, etc. ((Scellato et al 2010)). Des mesures comme 'la localité du nœud' et 'le coefficient de clustering géographique' qui incorporent le positionnement et les distances géographiques des nœuds ont été proposées ((Scellato et al 2010)) (Nettleton 2013).

Prédiction des liens

Au cours des dernières années, le problème de la prédiction des liens (LP) a attiré une attention accrue vue la variété de techniques fondées sur la théorie des graphes, apprentissage relationnel statistique, factorisation de la matrice, des modèles probabilistes graphiques, etc., (Lichtenwalter et al 2010) ((Xiang 2008)) ((Liben-Nowell & Kleinberg 2007)). L'un des objectifs récents de la prédiction des liens dans les médias sociaux est de détecter l'existence des liens non reconnus afin de faciliter aux users de peupler leurs réseaux personnels. En outre, la prédiction est évoquée aussi dans la dynamique temporelle des SNs. Pour une paire de nœuds disjoints (x, y) le problème consiste à prédire si une relation entre les deux peut avoir lieu dans l'avenir ((Xiang 2008)) ((Loiacono 2011)). Un score de probabilité est attribué à (x, y), plus qu'il est élevé plus qu'il est probable d'avoir ce lien.

Dans la littérature, il se trouve différentes techniques pour calculer ces scores de prédiction des liens dans les réseaux. Les prédicteurs populaires (approches traditionnelles) sont souvent basés sur l'effet des liens potentiels au voisinage local ou les chemins entre nœuds, en

mesurant la probabilité de connexion (de proximité) dans le réseau (similarité : nombre de voisins communs). C'est-à-dire, les méthodes de prédiction sont basées sur des scores de prédiction: des scores (structurels) de similarités entre une paire de nœuds données x et y. Ces scores sont basés sur des métriques de proximité et similarité dont certains sont abordés déjà dans un tableau précédent (Tableau 5).

Prédicteurs basés sur des informations structurelles locales

L'indice : nombre de voisins communs (CN) est le plus utilisé en perdition tel que : Il est plus probable de lier deux nœuds ayant le plus grand nombre de voisins (directs) communs. Avec le coefficient de Jaccard (JC), les deux nœuds ayant la proportion la plus élevée des voisins communs par rapport au nombre total de leurs voisins sont plus susceptibles d'être liés. Le coefficient de Adamic/ Adar (AA) ((Adamic & Adar 2003)) est une version liée au (JC) qui met l'accent sur l'importance des voisins communs. En outre, ***l'attachement préférentiel (PA)*** ((Newman 2001a)) ***est un autre indice qui se base sur le processus du développement du SN*** : La probabilité de création d'un nouveau lien à partir d'un nœud est proportionnelle à son degré ('***Rich get Richer***'). Il existe plusieurs autres indices et extensions comme 'Resource Allocation Index' (RA) ((Zhou et al 2009)), etc. Ces méthodes de prédictions sont basées sur l'information structurelle locale (similarité au voisinage), en ignorant des informations intéressantes sur les nœuds intermédiaires entre deux nœuds.

Prédicteurs basés sur les chemins et le processus de diffusion

Au-delà du voisinage, les scores de prédiction des liens peuvent être calculés sur des chemins de proximité. Par exemple, 'Inverse Path Distance' (IPD) est un score obtenu en inversant la distance géodésique entre deux nœuds x et y. Ce score signifie que les deux nœuds les plus proches sont susceptibles d'être liés. IPD = 1, lorsque x et y partagent au moins un voisin en commun. Cependant, IPD tend vers 0 si la distance augmente. Dans ce sens, PropFlow (Lichtenwalter et al 2010) est une nouvelle méthode qui calcule efficacement la probabilité qu'une marche aléatoire restreinte à partir de x se termine dans y en L étapes (seuil). C'est une mesure qui apparait insensible au bruit dans la topologie du réseau (Lichtenwalter et al 2010). En se basant sur la même hypothèse de l'attachement préférentiel, le PageRank (PR) peut être utilisé aussi pour évaluer les scores (des probabilités) de prédiction. Dans ce cas, la création des liens entre les nœuds x et y, ou la probabilité d'arriver à un nœud cible y (via une marche aléatoire) à partir de x, est guidée par l'importance du nœud. En effet, le PageRank est considéré comme un modèle de marche aléatoire (processus de diffusion (DP)) modifié qui est très utilisé pour résoudre des problèmes de recherche d'information dans plusieurs domaines ((Donoser & Bischof 2013)). Le processus de diffusion est généralement basé sur des modèles de marches aléatoires (Random Walks (RW)) suivant une matrice de transition qui définit les probabilités pour passer d'un nœud donné à un autre voisin. Les chercheurs commencent récemment à appliquer des modèles de RW pour résoudre le problème de prédiction. Tant que les données sociales sont structurellement présentées par des graphes, une prédiction des liens, basée sur le processus de diffusion (LPDP), (Techniques de chaîne de Markov) est ainsi bien adaptée. Souvent, le (DP) se trouve devant le problème de réduction de dimensionnalité. Dans ce sens, la méthode de 'Diffusion maps' (DM) ((Coifman & Lafon

2006)) est une méthode non-linéaire qui a marqué des succès dans des problèmes en dehors de SNA et qui permet de passer vers un espace (euclidien) de dimension inférieure, tout en préservant la structure géométrique locale intrinsèque de données ((Coifman & Lafon 2006)). Avec (DM), les scores de prédictions (LPDM) sont définis par une distance entre les pairs des nœuds qui est approximativement la distance de diffusion.

Dans le cas des réseaux de collaboration scientifique, des expériences montrent que PR, LPDP et LPDM tendent à surperformer les autres techniques. Ces techniques s'adaptent bien avec la façon dont ces réseaux évoluent. Les auteurs sont plus susceptibles de collaborer avec autres auteurs ayant un nombre de publications hautement supérieur. Autres techniques moins connus sont proposées par exemple par le projet KONECT [55] qui met en œuvre divers algorithmes de prédiction des liens, etc.

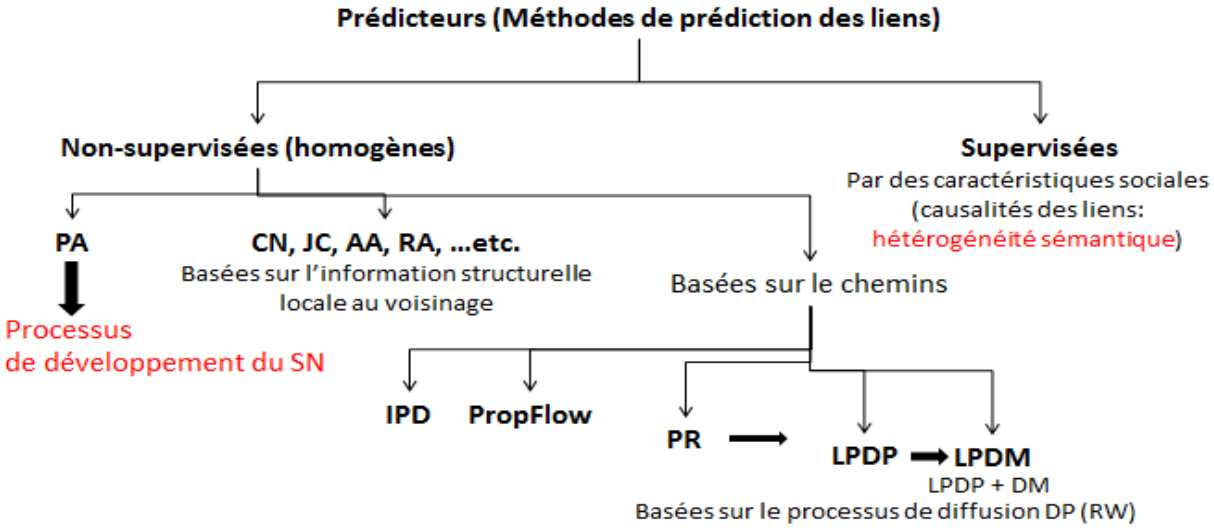


Figure 27. Méthodes pour la prédiction des liens

Inconvénients, des nouveaux aspects pour superviser la prédiction.

La performance et la précision d'un prédicteur sont dépendantes du choix des métriques de similarité (de proximité entre les pairs des nœuds). Les scores de prédiction peuvent être adaptés sur des réseaux pondérés. Cependant, même si ces méthodes de prédiction restent populaires dans la littérature en raison de leur simplicité, elles présentent aussi des inconvénients. Elles peuvent être efficaces seulement si la topologie du réseau est conforme préalablement à la fonction qui calcule les scores. Le classement des paires des nœuds est effectué par une seule métrique malgré qu'il y ait différents patterns structurels contenus dans le réseau. En outre, ***ces prédicteurs ont été montrée très sensibles aux propriétés sous-jacentes du réseau*** ce qui en résulte un déséquilibre dans la taille des communautés, des difficultés pour interpréter les interdépendances dynamiques du réseau (Lichtenwalter et al 2010). ***Même les techniques basées sur la diffusion, les plus performantes, se basent seulement sur l'information structurelle (topologique) du réseau sans considérer les caractères implicites des nœuds.*** Par conséquent, ces prédicteurs sont souvent décrits comme non-supervisés et aussi homogènes car ils traitent souvent tous les liens de manière homogène. En effet, même si le SN est composé par un seul type de relations, les liens du

réseau peuvent être engendrés par des facteurs de causalités hétérogènes (hétérogénéité sémantique). Les connexions dans les réseaux humains présentent souvent le résultat d'un processus social guidé par des orientations et des affiliations, etc. Par exemple, dans un SN de collaborations scientifiques, un auteur A et B peuvent collaborer pour publier un papier en 'machine learning'. A peut également collaborer avec C sur 'parallel computation' et avec D sur 'Data Mining'. Dans un cas pareil, un prédicteur traditionnel n'est pas en mesure de découvrir qu'il est plus probable de prédire une connexion entre B et D qu'entre B et C, puisque le 'Data Mining' est une discipline liée au 'machine learning'. Dans certains cas, les informations disponibles sur les users et leurs orientations ou intérêts comme les tags peuvent servir à identifier les causalités de connexions derrière les liens. Mais, souvent ces facteur sont difficiles à interpréter, particulièrement, quand il s'agit des liens crées dans les communautés qui se chevauchent 'overlapping communities'.

Les scores de prédiction obtenus à partir les indicateurs précédents peuvent être utilisés comme attributs pour des modèles de prédiction supervisés. Généralement, les approches supervisées ont accès à des données étiquetées. Mais, la supervision d'une prédiction de liens qui attire actuellement l'attention des chercheurs peut être guidée par sources d'information externes ou bien à partir les liens existants dans le réseau. D'où, il faut bien estimer la proximité réelle des personnes connectées sur le SN afin de prédire plus précisément l'existence d'autres connexions. Par exemple, dans le cas des réseaux des coauteurs (collaboration scientifique), ((Hasan et al 2006)) ont proposé d'utiliser l'apprentissage supervisé pour la prédiction des liens. Un ensemble de caractéristiques de liens est identifié comme étant la clé d'une performance effective de la prédiction. Les caractéristiques de proximité ne sont pas seulement topologiques. L'ensemble inclus aussi des caractéristiques de proximité extraites à partir les mots clés des manuscrits ou agrées à partir d'autres opérateurs. Donc, au-delà de la proximité topologique, il existe des approches alternatives pour extraire autres caractéristiques à partir des formats relationnels des données (Base des données et requêtes), des modèles statistiques, etc. D'autre part, ((Kamei et al 2012)) ont abordé aussi une méthode d'apprentissage pour prédire les liens manquants dans un réseau de média social ('Japanese word-of-mouth communication website') en se basant sur les données d'activité d'un user. La méthode appliquée montre comment l'apprentissage peut se faire à partir les liens observés et aussi les données d'activités et peut estimer avec précision les probabilités de création des liens absents ((Kamei et al 2012)) (Nettleton 2013). Dans telles études, l'information de supervision vient des propriétés de certaines ressources comme les documents, les activités, etc., selon la disponibilité des données. Cependant, certains auteurs veulent que ces informations soient tirées depuis le réseau lui-même. Il s'agit d'extraire des caractéristiques sociales tenant en compte les causalités des liens et ainsi mettre en place des techniques de prédiction de liens hétérogènes ('Link Prediction using Social Features' (LPSF)). ***Ce qui permet d'améliorer les performances de la prédiction*** notamment dans des réseaux de collaboration scientifique.

Il convient de noter que les réseaux de collaboration peuvent être multi-relationnels en incluant aussi différent type de nœuds (documents, conférences, etc.). D'où, tels réseaux hétérogènes posent autres questions dans la prédiction des liens. À cet égard, il y a quelques

propositions. ((Davis et al 2012)) proposent 'Multi-Relational Link Prediction' (MRLP) ((Davis et al 2012)) qui s'effectue en deux phases. La première est non-supervisée (avec la méthode Adamic/ Adar (AA)) et la deuxième est supervisée. Sur un plan plus large, ((Sun et al 2011)) proposent un modèle de prédiction de relations basé sur les chemins 'PathPredict' pour prédire les liens coauteurs dans un réseau hétérogène bibliographique. En outre, on trouve aussi un algorithme de prédiction basé sur 'Random Walks' proposé par ((Lee & Adorna 2012)) sur un réseau hétérogène bibliographique affecté. Tel que tous liens entre les objets hétérogènes sont pondérés par une combinaison de différentes mesures.

Par conséquent, le problème prédiction des liens (futurs) peut faire appel non seulement à l'aspect dynamique temporelle mais aussi à l'aspect sémantique afin d'améliorer, superviser une méthode de prédiction et inférer plus précisement des relations

Nouvelles dimensions

Aujourd'hui, les tendances de SNA (Un cadre de recherche de ce travail) sont beaucoup plus inspirées de la richesse des SNs qui constitue un autre élément important dans notre travail après avoir défini un motif d'analyse sérieux et accordé une importance à la pertinence des données. Les données sociales sont plus variées et permettent de plus en plus de saisir la richesse des SNs : Sa dynamique temporelle, sa richesse sémantique, cultuelle, etc. De ce fait, les travaux de recherche en SNA sont engagés sur des nouvelles pistes. On cherche à développer des modèles de SNs de plus en plus réalistes afin de donner plus de fidélité aux études analytiques, même si la taille croissante des SNs à grande échelle représente également un autre défi. On parle récemment des nouvelles dimensions d'analyse : Dimensions temporelles, sémantiques, etc. Les études analytiques qui fournissent maintenant des informations précieuses et des interprétations significatives sont celles qui s'intéressent aux aspects temporels ou sémantiques au-delà de la version statique et topologique de SNA.

Sémantique des SNs

Le développement du SN, et le comportement dynamique de ses individus et ses groupes dans le temps sont régnés par une certaine sémantique implicite qui les contrôle. Loin et en parallèle des aspects temporels dynamiques, les chercheurs essayent d'améliorer le SNA en exploitant la richesse sémantique des SNs, la sémantique des relations, des intérêts, etc. Mais le premier pas consiste à représenter cette sémantique de manière de plus en plus explicite.

Aujourd'hui, en utilisant les technologies de communication et services web partout, les OSNs se prolifèrent dans différents environnements, une multiplicité de contextes, de rôles et d'identités ((Erétéo 2011)). Donc les SNs sont devenus déjà complexes à représenter en ajoutant en plus leur sémantique. Si les nœuds représentent numériquement des personnes, ces derniers peuvent être liés par un ou plusieurs types de relations: amitié, parenté, intérêt commun, échanges financiers, goûts, etc. (Kazienko et al 2011). Ils utilisent par ailleurs, divers services : courriers électroniques, systèmes de télécommunication, sites de réseautage social, systèmes de partage multimédia, etc. (Kazienko et al 2011) et selon différentes activités. Chaque utilisateur d'une application représente une personne, dans un contexte donné, pour un rôle particulier, c'est un fragment de son identité dans le réseau ((Erétéo 2011)). Il développe ainsi différents liens sociaux, à travers une ou plusieurs applications ou services ((Erétéo 2011)) (Kazienko et al 2011).

Selon (Kazienko et al 2011), la sémantique (types) des relations entre les acteurs du SN est plus accessible, en la dérivant depuis les différentes activités humaines dans ces systèmes informatiques (Kazienko et al 2011). Mais, les chercheurs manipulent souvent dans leurs études des structures avec un seul type de relation (graphes non-typés) (Kazienko et al 2011). Maintenant, des approches sont proposées pour traiter ce problème de relations multiples en se basant sur différents notations graphiques (algébriques, sociométriques, etc.) (Kazienko et al 2011). Ces notations constituent l'image de l'un des concepts théoriques les plus difficiles à mettre en œuvre en SN (Kazienko et al 2011). Certaines tendances s'orientent vers les technologies de web sémantiques ((Yessad 2009)) ((Takes 2011)), un moyen pour aller plus loin dans la sémantique des SNs (Erétéo et al 2008) vers la sémantique des communautés/ des groupes, etc. ((Erétéo 2011)) ((Erétéo et al 2011)).

- **Représentations sémantiquement plus riches**

La sémantique des SNs se présentent dans la diversité des contextes, les types de relations, les orientations, les intérêts, etc., et tous les aspects implicites derrière le comportement et l'évolution des entités sociales. C'est un aspect important en SNA. Les données sociales sémantiquement structurées nous permettent de connaitre la signification, le type d'une relation, la sémantique du comportement individuel, d'un groupe/communauté, etc., mais permettent aussi à la machine de traiter cette sémantique. Construire des modèles plus riches est le premier pas pour enrichir le SNA. Les langages de web sémantique est une solution et un outil puissant pour améliorer sémantiquement les représentations des SNs (Erétéo et al 2009) (Erétéo et al 2008) ((Erétéo et al 2011)). En s'appuyant sur des modélisations ontologiques (FOAF, RELATIONSHIP, SKOS, etc.), des graphes sociaux RDF (des folksonomies structurées) peuvent être instanciés. Mais les défis se manifestent quand il s'agit d'analyser et interroger ces modèles. En outre, si le 'social tagging' était une source de SN la structuration sémantique de social tagging est aussi une source pour les SNs sémantiques pertinents.

- **Analyse des SNs sémantiques**

Le problème ne s'arrête pas devant la modélisation des SNs sémantique. Comment peut-on exploiter aussi cette richesse pendant l'analyse ? ((Paolillo & Wright 2006)) ((Goldbeck & Rothstein 2008)). Souvent, l'expressivité des représentations se réduit et des connaissances sont perdues. D'autre part, chercheurs ((Erétéo 2011)) pensent à fusionner le web sémantique et la théorie des graphes classique pour enrichir SNA. Mais, si on veut s'en servir toujours des langages (de requêtes) de web sémantique, il faudra mettre en évidence que ces outils ne supportent pas systématiquement la complexité des indices et techniques de SNA. Il faudra penser à des extensions, des contraintes logicielles, et d'environnement, des techniques de résolutions, etc., dans des contextes particuliers ((Anyanwu et al 2007)) ((Kochut & Janik 2007)), ((Corby 2008)). En s'appliquant sur des réseaux de taille moyenne, il y en a certains qui ont obtenu des résultats en temps raisonnables ((Erétéo 2011)) (Erétéo et al 2008). Toutefois, l'analyse des graphes RDF (Centralités, coefficient de clustering, détection de communautés, etc.) en utilisant particulièrement des requêtes est plus complexe et coûteuse en temps et en espace notamment avec les grands réseaux. C'est pour cette raison, certains auteurs ont cherché des approximations en calculant par exemple la centralité d'intermédiarité d'une personne sur un sous réseau. On pense aussi à des techniques (heuristiques,

échantillonnage, etc.) itératives adaptables au parallélisme pour avoir des bonnes approximations. ***L'objectif est d'aller vers une sémantisation d'analyse ou une analyse sémantique des SNs mais le problème est quand on se limite souvent par ce contexte statique.***

- **Regroupement et communautés sémantiques**

Au-delà des caractéristiques topologiques, la formation des communautés ou des groupes implique aussi une sémantique plus énigmatique derrière. Pratiquement, la structuration sémantique des relations et des folksonomies est un moyen pour comprendre la sémantique des communautés. En partageant par exemple le même intérêt est suffisant pour donner un sens sémantique à une communauté. Mais pour beaucoup de chercheur, la connectivité des acteurs reste une condition préalable pour l'identité de la communauté. L'algorithme sémantique 'SemTagP' ((Erétéo et al 2011)) ((Erétéo 2011)) incarne ce principe en se basant sur l'algorithme de propagation (P) des étiquettes LPA ou RAK ((NGUYEN et al 2013)) ((Xie & Szymanski 2011)) ((Raghavan et al 2007)). ***Le mécanisme sémantique de regroupement est un nouveau paradigme, évoqué aussi dans des contextes dynamiques.***

Dynamique temporelle, Visualisation et analyse multidimensionnelle des SNs

L'incorporation de la composante de temps en SNA nécessite de comprendre d'abord les effets influant sur la dynamique temporelle du SN et démontrer comment il évolue. En effet, le premier défi est de développer des modèles qui captent et représentent l'évolution d'un SN dans le temps adin de dépasser les problèmes qui se posent dans les modèles statiques. L'analyse d'un modèle de SN évoluant dans le temps dépend de son degré de complexité et à quel degré la composante temporelle est explicite (Empreintes statiques dans le temps, des graphes temporels, etc.). Dans tous les cas, des résultats informatifs seront obtenus expliquant le processus d'évolution et de diffusion, l'évolution des métriques (vers même la définition des centralités temporelles), la dynamique des communautés et des phénomènes liés, etc.

La visualisation des SNs (graphiques ou descriptive) ou l'analyse visuelle est aussi l'une des pistes de SNA notamment devant la progression quantitative (la taille) et qualitative des SNs. Par exemple, l'exploration des nouvelles méthodes (en incorporant l'aspect temporel) pour visualiser la dynamique temporelle (dynamique des groupes) dans les SNs.

En ce qui concerne la considération de plus d'une dimension d'analyse au même temps, les chercheurs animent récemment un débat sur la possibilité d'une modélisation et analyse multidimensionnelle ((Scott 2011), Social network analysis: developments, advances, and prospects). Le temps, les types de relations, les groupes, etc., peuvent faire l'objet (dimensions) d'une structuration multidirectionnelle d'un SN.

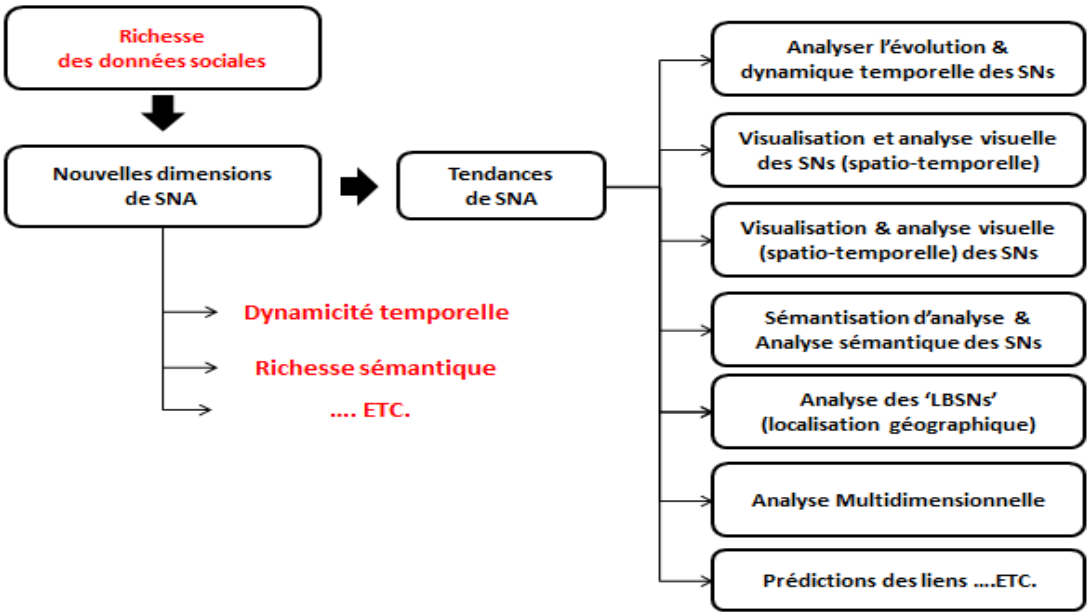


Figure 28. Des nouveaux aspect et tendances du SNA, inspirés de la richesse des données sociales (OSNs)

Chapitre 2 : Modélisation & analyse de la dynamique temporelle des réseaux sociaux

1. Introduction

La dynamicité temporelle des OSNs/SNs, étant l'un des 'Hot topics' laisse toujours le SNA/SNAM un domaine passionnant (Nettleton 2013). On ne cesse pas de développer des nouvelles tendances qui s'intéressent au comportement dynamique individuel et collectif. Sous l'influence de plusieurs effets, mécanismes sous-jacents, l'environnement, évènements externes, etc., les acteurs du SN agissent, développent/retirent des relations. Dans l'ensemble ces modifications/ comportements atomiques sont soumis sous des lois (des caractéristiques) comme coévolution réseau-comportement, phénomène 'rich get richer', etc., et génèrent un contexte dynamique endogène. Il permet d'observer l'évolution de tout le SN comme un processus de développement en temps continu.

Aujourd'hui, les structures dynamiques des SNs encouragent la recherche dans les réseaux dynamiques en général et en SNAM en particulier, avec plus de dimensionnalité. Les chercheurs sont convaincus que l'aspect temporel constitue une dimension cruciale pour formaliser et analyser de manière plus réaliste des SNs et des phénomènes connexes, et au même temps un challenge. Sachant que les représentations statiques ne répondent pas maintenant à ces exigences, le défi commence déjà par construire des modèles mathématiques qui sont capables d'intégrer/ exprimer la composante temporelle et reproduire les propriétés observées des réseaux réels. Deux types d'approches de modélisation seront distingués, selon lesquelles, on verra que les auteurs ont soit une tendance statistique d'analyse sur une séquence d'empreintes dans le temps ou bien ils étudient des aspects, des processus et des patterns d'évolution plus complexes. La première méthode de modélisation aide à étudier la croissance/densification du SN, l'évolution des centralités/ rôles individuelles, etc. Mais avec la deuxième méthode, la composante temporelle est plus explicite et permet de définir des extensions de métriques temporelles basées sur des formalismes plus complexes de graphe variant dans le temps. C'est moyen pour expliquer et simuler des phénomènes et des processus dynamiques comme la diffusion, la médiation d'information. L'analyse de la dynamique temporelle des SNs donne par ailleurs des résultats plus précis et montre que la version statique surestime les géodésiques et sous-estime leur longueur, sans omettre la surestimation de la connectivité du réseau, ses groupes. La dynamique des communautés/ des groupes est un niveau de dynamique plus élevé, encadrée par différentes propositions, dont la majorité s'appuient sur une découverte de communautés enfilée sur des snapshots d'un SN dans le temps. Les travaux récents étudient des propriétés d'évolution (évolution spatiotemporelle), de stabilité de groupes (comprenant des structures sous-jacentes), affiliation chronologiques selon des extensions et des modèles appropriés. L'étude de la dynamique temporelle du SN est en général largement influencée par des paramètres comme la résolution des fenêtres de temps. Elle pousse par ailleurs la recherche à améliorer la

visualisation des SNs par des techniques plus sophistiquées et constitue d'autre part une passerelle vers d'autres paradigmes (sémantique, etc.).

2. Dynamicité temporelle du SN (comment évolue-t-il ?)

Une caractéristique clé des interactions sociales est leur changement de manière continue (Berger-Wolf & Saia 2006) (Kazienko et al 2011) même avant la médiatisation en ligne (OSNs). En effet, les SNs présentent des structures sociales qui ne sont pas statiques et qui changent au fil de temps. Par exemple, les communautés ne sont pas en réalité statiques. Il s'agit des organismes vivants, qui évoluent suivant des facteurs culturels, environnementaux, économiques, des tendances politiques, des interventions externes, des événements imprévus, développements technologique (Ahn et al 2011), etc.

2.1.Des effets influant sur la dynamique des SNs

Les changements qui se produisent dans un SN/OSN au fils du temps ne sont pas liés seulement à des évènements (externes). Les entités sociales (les personnes) qui forment le réseau peuvent eux-mêmes changer de comportement sous l'influence de plusieurs effets. Selon ((Snijders 2005)), le changement du SN peut être vu comme une coévolution du réseau et comportement (attributs) de l'acteur qui désigne souvent sa performance, son attitude, etc. Cette coévolution évoque une dynamique mutuellement dépendante du réseau et des attributs individuelles de ses acteurs. L'influence et la sélection sont des mécanismes sous-jacents (***endogènes***) qui peuvent déterminer l'étendue de cette évolution/ dynamique temporelle. Dans ce sens, ((Steglich et al 2010)) ont étudié ces deux mécanismes dans une école secondaire en Ecosse entre 1995 et 1997 (réseau) par rapport à des comportements comme la consommation du tabac et de l'alcool. Les auteurs soulignent que les études empiriques antérieures sur cette dynamique ont échoué à résoudre des problèmes statistiques et méthodologiques fondamentaux ((Steglich et al 2010)). Leurs travaux qui rentrent dans le cadre du projet européen de recherche collaborative s'intéressent à la dynamique des acteurs et des réseaux sur différents niveaux: individus, groupes, organisations et contextes sociaux.

Influence sociale

L'influence sociale désigne le faite que le comportement d'un acteur peut être influencé par les autres acteurs dans le SN. Le comportement (attitude, mode de vie, etc.) de l'acteur peut être guidé non seulement par ses propres attributs, mais également par sa position dans le réseau, le comportement et les attributs de ceux qui sont directement ou indirectement liés avec lui. Par exemple, l'effet d'amitié incarne remarquablement l'influence sociale de telle sorte que les gens ont tendance à se comporter comme leurs amis (Jamali et al 2011). Si on prend les comportements de tabagisme, la consommation d'alcool, drogue, etc., chez les adolescents, les réseaux ou les groupes d'amitiés (homogènes) sont utiles pour étudier la dynamique des réseaux et la coévolution de tels comportements. Par exemple, quand une personne crée un lien d'amitié avec un fumeur, ce dernier peut lui apprendre à fumer ((Steglich et al 2010)). L'ami d'un fumeur est probablement un fumeur. L'initiation au tabagisme 'Smoking' et sa coévolution avec les réseaux d'amitié a été étudié dans le travail de ((Steglich et al 2010)). Dans ce cas, l'amitié semble être un facteur de nuisance à contrôler. Cependant, les auteurs le considérèrent comme un facteur intéressant à modéliser avec le

comportement de 'Smoking' ((Snijders 2005)) pour expliquer la dynamique de ces réseaux sous des effets comme l'influence sociale. Selon ((Snijders 2005)), une mauvaise spécification d'un modèle donnée de l'amitié peut biaiser les conclusions sur l'influence sociale ((Snijders 2005)), et autres conclusions sur des effets variables qui sont associés à l'amitié ((Snijders 2005)).

Par conséquent, les individus sont influencés par leur SN (Santoro et al 2011). Cette influence se montre également sur le comportement en ligne. Sur les OSNs, (Jamali et al 2011), parlent aussi d'une influence corrélationnelle qui s'illustre beaucoup plus dans un réseau de notation social (Social rating network 'SRN') (Jamali et al 2011). Dans ce cas, le comportement des users se présente par la notation (l'évaluation) des contenus partagés et des sites web (Epinions, Flickr, etc.). Ici, l'influence veut dire que les gens peuvent adopter un comportement ou des motifs d'évaluation similaires à ceux des autres utilisateurs (Jamali et al 2011). Donc, (Jamali et al 2011) modélise la dynamique temporelle de tels SNs en se basant sur des effets bidirectionnels du comportement d'évaluations et des relations sociales.

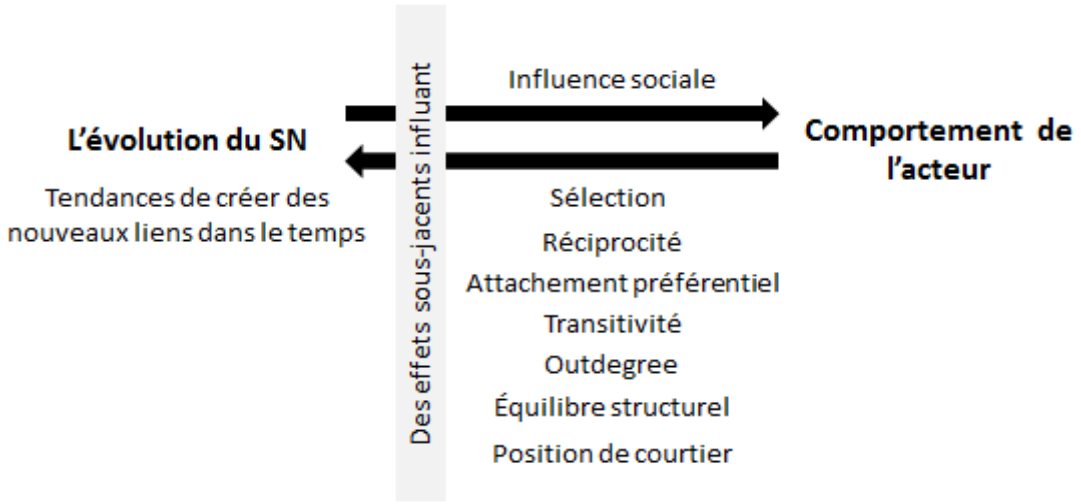


Figure 29. Des effets influant sur l'évolution du SN dans le temps

Sélection

Beaucoup de chercheurs évoquent le mécanisme de sélection sous le concept d'homophilie ((Steglich et al 2010)) qui caractérise souvent la création des liens dans les SNs et notamment les OSNs. Les gens ont tendance à suivre une sélection homophile. Ils créent des relations avec autres personnes qui sont déjà similaires à eux (Jamali et al 2011) au niveau de certains attributs individuels (sexe, âge, argent, profession, intérêt, etc.). Quand il s'agit des entreprises, organisations et les alliances, la sélection consiste à choisir de collaborer ou de coopérer selon les performances en cherchant une meilleure réussite de l'organisation, et des possibilités de récompenses futures, etc. Des entreprises à haute réputation ne collaborent pas avec des entreprises à faibles réputation ((Snijders 2005)). Dans l'exemple précédent (le comportement de tabagisme) étudié par ((Steglich et al 2010)), un fumeur tend à choisir des amis fumeurs. Il ou elle est susceptible de rencontrer d'autres fumeurs dans les zones de fumeurs là où il peut former des liens d'amitié avec eux (une sélection). Par conséquent, la

création des nouveaux liens peut être le résultat d'une sélection homophile (Similitude covariable), ((Steglich et al 2010)).

Les exemples du comportement d'évaluation sur des OSNs (SRN) ou des comportements risqués comme le tabagisme, la prise d'alcool ou de drogues dans le monde réel sont contagieux entre amis (influence), mais aussi opératifs dans la formation (sélection) des liens d'amitié (effet bidirectionnel). Autres hypothèses s'intéressent à l'influence et la sélection sociale en dehors de cette relation réseau-comportement. Les acteurs du SN peuvent imiter structurellement des autres qui sont similaires mais pas nécessairement liés avec eux. Les individus imitent de certains acteurs populaires ou intégrés dans des sous-groupes cohésifs. L'imitation unilatérale est parfois plus forte que les liens d'amitié réciproques. Noter que la sélection et l'influence ne sont pas les seuls effets. Au-delà de la tendance générale de l'individu de vouloir faire des nouveaux liens (***Outdegree***), les auteurs distinguent des effets supplémentaires qui caractérisent une évolution endogène du réseau comme par exemple :

Attachement préférentiel
C'est un effet très important dans le développement du SN, qui a été abordé avec plusieurs concepts précédents (proximité des nœuds, prédiction des liens, etc.). Il explique la tendance à rejoindre des individus populaires dans le réseau.

Transitivité
Les observations ont conduit à découvrir que l'attachement du nœud est également influencée par le phénomène 'friends of friends' (Nettleton 2013), (l'ami de mon ami est un ami) qui désigne l'effet de transitivité. Il se présente sous forme de patterns très fréquents dans le SN : Cycle ou fermeture triadique au voisinage. La tendance à former des cycles relationnels est considérée comme un effet qui s'oppose au concept de la hiérarchie ((Steglich et al 2010)).

Autres effets
Réciprocité : représente une tendance à avoir des liens réciproques

*Équilibre structurel***:** C'est la tendance de faire des liens avec d'autres structurellement similaires ((Steglich et al 2010)).

Position de courtier : c'est un autre facteur qui permet à un acteur, occupant une position intermédiaire dans le réseau, de créer des liens entre des nœuds indépendants.

Il faut noter que ces effets influant sont difficiles à contrôler dans une modélisation qui tente de représenter le comportement dynamique du réseau ((Snijders 2005)). Pour différents types de SNs, il n'y a pas trop de modèles proposés qui sont capables de capturer certains de ces effets au même temps afin d'expliquer cette dynamique temporelle. Par exemple, on trouve (Jamali et al 2011) qui proposent un modèle capable de représenter les 4 effets essentiels : L'influence sociale, la transitivité, la sélection et l'influence corrélationnelle. Il y a même des travaux qui considèrent ces effets comme statiques ou constants durant l'évolution d'un réseau comme OSN (Jamali et al 2011). Cependant, les observations et les points de vue les plus réalistes confirment que ***tous ces effets sont dynamiques et contribuent à la***

formation et l'évolution endogène du SN dans le temps. Mais ils ne sont pas les seuls facteurs influant sur cette dynamique.

Influence de l'environnement et les évènements

Des chercheurs ont fait preuve que l'environnement et même les évènements du monde réel ont une influence remarquable sur l'évolution des SNs. Par exemple, (Meeder et al 2011) ont donné une démonstration sur un évènement qui a changé Twitter en 2009. Il y avait une augmentation du nombre des nouveaux comptes Twitter (Meeder et al 2011) lors des élections en Iran à la fin de Juin 2009. En outre, (Meeder et al 2011) ont montré aussi que des OSNs (Celebrity Follower Subgraph sur Twitter) peuvent afficher des changements qui ne sont pas dues à l'interface de Twitter mais plutôt sous l'influence des évènements du monde réel. D'autre part, (Reda et al 2009) ont étudié un cas du monde réel : un troupeau de zèbres de Grévy en Kenya, où la dynamique d'un SN, est montrée influencée par le temps et aussi l'environnement. En plus de l'état de reproduction des individus, les besoins de ressources (l'eau et de l'herbe) laissent ces individus choisissent des associations qui maximisent l'accès à ces ressources (Reda et al 2009). C'est une évolution spatiotemporelle qui illustre le rôle de l'environnement dans la formation et l'évolution de la structure sociale (notamment groupes) avec le temps.

2.2.Comment le SN évolue-t-il au fil du temps

Le SN est défini également comme un ensemble de relations, des Pattern des liens entre un groupe d'acteurs et un comportement(s) qui est n'importe quel attribut individuel modifiable ((Snijders 2005)) dans le temps. Nous devons comprendre et admettre préalablement que la dynamicité temporelle d'un SN est observée comme un processus de développement du réseau en temps continu.

2.2.1. L'évolution du SN au niveau atomique (dynamique endogène)

D'après les effets précédents, il existe une relation de rétroaction dans la dynamique des liens et les comportements/ performances des acteurs du SN. Le processus de rétroaction (feedback) sociale alimente la dynamique des acteurs qui est ainsi le moteur principal de l'évolution d'un SN. Mais, la recherche sur ce type de processus est difficile ((Snijders 2005)). Chaque acteur contrôle ses relations sortantes et son comportement implicite ((Snijders 2005)). Ce sont des variables endogènes qui se développent dans une dynamique simultanée et s'influencent mutuellement ((Snijders 2005)). Sa tendance de vouloir changer ces variables est liée à son appréciation qui est au cœur de cette évolution. ((Snijders 2005)) modélise l'appréciation d'un acteur i de sa position dans le réseau x par une fonction objective fi (x), ((Snijders 2005)). Tant que la situation actuelle est loin de ce qui est optimal pour eux ((Snijders 2005)), les acteurs changent leurs objectifs (au niveau du comportement/ relations). Autrement dit, ils cherchent à obtenir une valeur élevée de cette fonction objective suivant un taux de modification ((Snijders 2005)). Sinon, ils maintiennent une structure d'équilibre dynamique quand la situation est proche de l'optimum ((Snijders 2005)). ((Snijders 2005)) a défini une telle fonction objective en exprimant (modéliser) certains effets (OutDegree, réciprocité, etc.) influant sur ces changements: Il s'agit par exemple d'une somme pondérée du nombre de liens sortants (Outdegree) et le nombre de liens réciproques (les poids sont des paramètres estimés), ((Snijders 2005)).

En appliquant des changements, à un moment donné, les acteurs agissent d'une manière indépendante, sans coordination ((Snijders 2005)). Mais au fil du temps, ces modifications génèrent un contexte dynamique endogène ((Snijders 2005)) et les acteurs du réseau semblent être dépendants ((Snijders 2005)). Par exemple suivant l'effet de réciprocité ou la fermeture transitive, etc., un lien peut produire un autre, souvent sans une décision conjointe comme dans la théorie des jeux ((Snijders 2005)), mais plutôt suivant ***l'ordre chronologique*** de leurs décisions.

2.2.2. Vue d'ensemble sur l'évolution du SN (Caractéristiques)

L'ensemble des changements microscopiques appliqués par les acteurs (les users) au niveau du comportement et des relations affichent des caractéristiques de haut niveau décrivant la dynamique temporelle des SNs. Même s'il s'agit des réseaux à grande échelle, les SNs et notamment les OSNs affichent souvent des changements perceptibles au fil du temps. Des nombreux travaux ont expliqué les changements des SNs et leur dynamique dans le temps ((Rowe et al 2014)). Les conclusions prouvent que ***les graphes sociaux tendent généralement à se densifier au fil du temps***. Il s'agit d'une évolution incrémentale du graphe social (Nettleton 2013). Des études comme celles de (Santoro et al 2011) ainsi que (McGlohon & Faloutsos 2008) sur l'évolution de certains indicateurs sont des bons exemples qui montent/ illustrent l'évolution du SN au fil du temps dans le cas général. À cet égard, des données sur une collection de papiers scientifiques et leurs citations (comme sur arXiv) ont fait l'objet de ces études pendant une période d'observation (Santoro et al 2011) (McGlohon & Faloutsos 2008).

Lorsqu'un SN devient de plus en plus dense/grand, une des premières conclusions est que le diamètre tend à se rétrécir lentement dans le temps vers une valeur constante (entre 6 et 5) (McGlohon & Faloutsos 2008). ((Leskovec et al 2007)) ont étudié la densification et le rétrécissement du diamètre sur différents SNs. Ils ont proposé un modèle de générateur de graphe « Forest Fire » pour expliquer la façon dont la création des liens se propage à travers le réseau et les changements conséquents par rapport les modèles statiques.

D'autre part, ce changement de connectivité a un impact direct sur la structure modulaire du SN (les groupes). La (Figure 30) montre l'évolution du coefficient de clustering global (une moyenne) et la modularité du graphe précèdent au moment où il tend à se densifier.

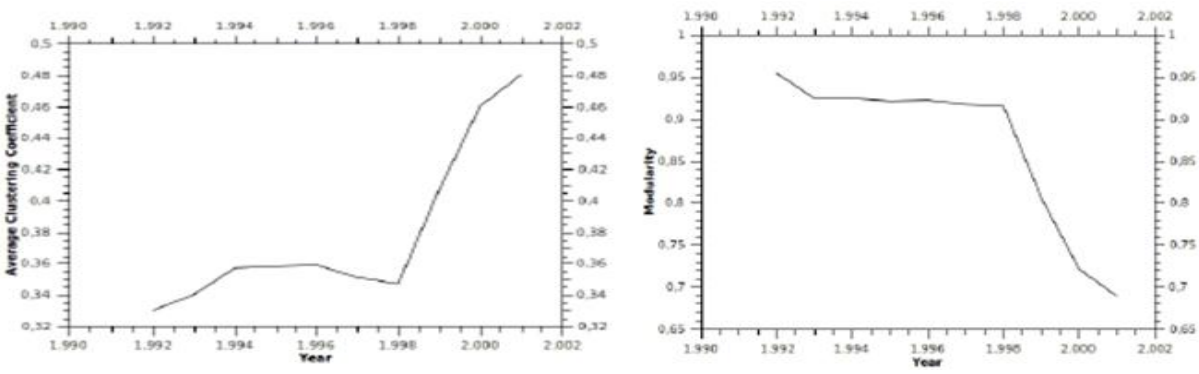


Figure 30. L'évolution du coefficient de clustering et la modularité d'un graphe de citations sur arXiv dans le temps ((McGlohon & Faloutsos 2008))

Au début, Les nœuds ont une faible tendance à se regrouper (faible coefficient de clustering). Ils forment des groupes séparés (une modularité saillante). A partir de 1999, l'évolution du coefficient de clustering indique une augmentation de la structure du réseau en cluster. Les nœuds (les petits groupes) commencent à se rassembler en sous-groupes vers des groupes de plus en plus denses (Santoro et al 2011) cela explique pourquoi la modularité du réseau diminue (Figure 30). L'étude de la modularité au fil du temps permet de fournir des indices très intéressants pour l'analyse de la dynamique des réseaux (la formation des groupes), (Santoro et al 2011): Voir si les communautés ont une tendance à se spécialiser ou homogénéiser (Santoro et al 2011). Dans cet exemple, il est clair que la modularité et le coefficient de clustering sont liés (Santoro et al 2011).

Le phénomène 'Rich get richer'

La tendance général qui contrôle la façon dont laquelle l'OSN/SN évolue et s'agrandit peut être définie en terme sociologique par le phénomène '***the rich getting richer***' ou encore 'avantage cumulatif ((Chakrabartiet al 2004)). En effet, les nouveaux nœuds ont tendance à être attirés par des nœuds existants ayant des degrés élevés, pour former des liens avec eux. Autrement-dit, les auteurs trouvent que les nouveaux liens se forment suivant le régime (la loi de puissance) de la distribution des degrés (In-Degree) des nœuds ((Chakrabartiet al 2004)). Par conséquent, le phénomène 'rich get richer' ou l'attachement préférentiel est cité comme l'explication de la distribution des degrés selon la loi de puissance.

Selon (Santoro et al 2011), étudier l'évolution de cette caractéristique du SN dans le temps permet de découvrir l'arrivée ou le départ des hubs (Santoro et al 2011). Les auteurs montrent que la courbe suit une loi de puissance : Une loi de puissance à chaque intervalle de temps. La dégradation de la fréquence des degrés élevés montre que le processus d'interconnexion est guidé par des nœuds à faible degré qui jouent le rôle des hubs entre les groupes (Santoro et al 2011). Par ailleurs, selon l'étude de (McGlohon & Faloutsos 2008) sur le réseau "Physics paper & leurs citations" sur ArXiv (entre Janvier 1993 et Avril 2003), la croissance du nombre de liens E(t) par rapport l'évolution du nombre de nœuds N(t) suit une loi de puissance (McGlohon & Faloutsos 2008). Ils confirment également que la distribution de degré suit une loi de puissance qui persiste au fil du temps (McGlohon & Faloutsos 2008).

Le 'gelling point' est une autre observation sur l'évolution du SN. C'est le point de gélification qui arrive à un moment donné, quand nombreuses petites composantes connectés se relient les unes avec les autres en créant ainsi une composante plus grande (McGlohon & Faloutsos 2008) ((Srivastava & DeLong 2013)).

Une évolution qui tend vers un équilibre

Les études précédentes montrent que certaines caractéristiques comme le diamètre et loi de puissance évoluant dans le temps arrivent à un certain moment, à un état de sagesse (Santoro et al 2011) (Nettleton 2013) ((Leskovec et al 2007)). Ce sont des conclusions proches de celles de ((Kossinets & Watts 2006)). Ils ont étudié l'évolution d'un SN. Il s'agit d'un réseau construit à partir les données (Timestamp, l'émetteur/ récepteur) des courriers échangés entre les étudiants et le staff de faculté d'une grande université (Nettleton 2013) ((Snijders 2010)). Même si ce SN est influencé par la structure organisationnelle de l'environnement (de l'université), les auteurs trouvent que ses caractéristiques sont plus ou moins constantes mais ont tendance à atteindre un équilibre dans le temps ((Kossinets & Watts 2006)).

3. SNA et la dynamique temporelle des SNs

Tout le monde s'accorde sur le fait que les véritables SNs ont aussi des structures dynamiques. Le développement des Smartphones, les réseaux véhiculaires et les réseaux satellitaires ont récemment favorisé la recherche sur les réseaux dynamiques et stimulé les études des nouveaux concepts (Santoro et al 2011). Cependant, dès son apparition, SNA s'est basée sur des caractéristiques statiques de SN. La majorité des mesures, techniques et enquêtes ne supportaient pas la dynamique des interactions. (Tang et al 2010a) affirment que la longueur des chemins, coefficient de clustering, la centralité, etc., statiques donnent une vue trop grossière sur les SNs tant que la dynamique temporelle est une information inhérente là-dedans. La plupart des études analytiques se concentrent sur un instant donné du réseau dans le temps, ou en faisant une agrégation de toutes les interactions à plusieurs instants dans une représentation statique (Tantipathananandh et al 2007). Il s'agit donc d'une SNA classique qui ne parvient pas à exploiter des propriétés dynamiques importantes ni de saisir des phénomènes d'évolution. Un exemple simple d'un scénario d'interactions entre trois individus (1, 2 et 3) peut avoir deux interprétations différentes (Tantipathananandh et al 2007). Par exemple {1, 2}, {1, 2, 3}, {1, 2}, {1, 2, 3} ou {1, 2}, {1, 2}, {1, 2, 3}, {1, 2, 3}. Dans la première séquence, l'individu : 1 et 2 sont observés en interaction à n'importe quel point de temps. 3 interagit avec eux dans la moitié du temps. D'où il est probable que 3 est un membre à part entière de ce groupe. La deuxième séquence explique l'affiliation de 3 de manière plus plausible en montrant qu'il rejoint le groupe pendant la période d'observation (Tantipathananandh et al 2007).

Exigence et influence de la dimension temporelle en SNA

Les chercheurs sont rapidement convaincus que les relations sont variables. Elles se développent de manière plus ou moins forte dans le temps et présentent plus de dimensionnalité qu'une analyse statique est incapable de capturer (Tang et al 2010a). Par analogie, le facteur temps comme dans la physique, constitue une dimension d'analyse (Kazienko et al 2011). Analyser suivant une dimension temporelle peut fournir des informations précieuses sur l'évolution des caractéristiques dans le passé, prédire dans le futur et ajouter des nouveaux concepts aux systèmes d'analyse ((Erétéo 2011)) ((Golbeck 2011)). Par exemple la date/ l'heure des interactions entre acteurs est très utile pour affiner des mesures de centralité ou une détection de communautés.

L'incorporation de la dimension temporelle dans SNA ouvre la voie aux différentes possibilités/ applications de modélisation de l'évolution des structures sociales complexes, émergentes dans l'environnement des réseaux informatiques. Allant plus loin encore, c'est une occasion pour étudier l'impact des SNs dynamiques sur la qualité de service, performance et les propriétés fonctionnelles des systèmes d'information.

En conséquence plusieurs questions se posent: Cette information peut-elle être utile pour évaluer précisément les rôles des acteurs et les groupes dans le SN ? ((Golbeck 2011)). Comment peut-on suivre la croissance des SNs, l'évolution individuelle, celle des groupes et la stabilité des structures sociales? Est-il possible qu'un acteur change de positionnement en jouant un rôle plus important 'leadership' ? Comment peut-on anticiper et détecter le développement de telles positions stratégiques ((Erétéo 2011)). Comment ces rôles, ces groupes peuvent–ils évoluer au fil du temps? ((Golbeck 2011)). Comment ces changements peuvent-ils affecter la structure des SNs et mêmes des patterns ou structures sous-jacentes? Autres questions se posent également sur la prévision cette évolution, etc.

Ce sont des sujets qui se dégagent d'une SNA ayant plus de dimensionnalité (temporelle) et qui peuvent inspirer les chercheurs pour résoudre autres problèmes liés à la sociabilité (SN): Les phénomènes d'épidémies et de propagation de maladie dans les SNs, simulations basées

sur des agents de SNs, etc. Cependant, ***la modélisation ainsi que l'analyse du processus dynamique des SNs/OSNs au fil du temps reste un véritable challenge scientifique.***
Beaucoup d'approches sont proposées pour modéliser et analyser l'évolution des graphes de SNs/OSNs. Par ailleurs, certains auteurs se contentent d'analyser statistiquement cette évolution dans le temps. D'autres étudient des aspects et modèles spécifiques du processus d'évolution (Nettleton 2013).

3.1.Comment modéliser l'évolution des SNs dans le temps

Malgré le grand succès du modèle représentatif des interactions sociales (le graphe), son un inconvénient majeur est que l'information temporelle indiquant là où les interactions sociales ont eu lieu est négligée (Berger-Wolf & Saia 2006). ***Cette représentation statique est incapable de distinguer les différents scénarios*** (Berger-Wolf & Saia 2006). La figure suivante montre 2 scénarios de la dynamique du réseau qui donnent le même graphe statique (Berger-Wolf & Saia 2006).

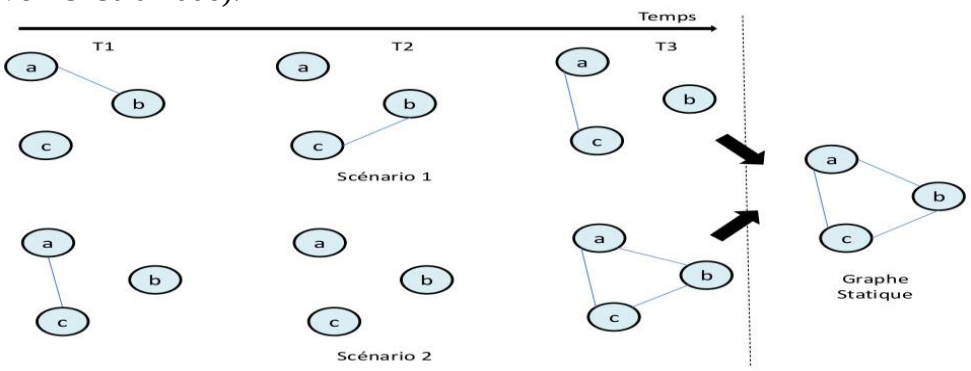


Figure 31. Un seul graphe statique qui représente deux scénarios d'interactions dynamiques différents, inspiré de ((Berger-Wolf & Saia 2006))

Les 2 premières lignes représentent le SN dynamique dans 3 instants et dans la dernière ligne, il se trouve un seul graphe statique correspondant (Berger-Wolf & Saia 2006). Supposons que le lien (**Figure 31**) représente le contact social qui provoque la transmission d'une maladie (Berger-Wolf & Saia 2006). Selon (Berger-Wolf & Saia 2006), *les décisions basées uniquement des données statiques peuvent être imparfaites*. Les auteurs donnent l'exemple d'une question qui cherche à vacciner uniquement une seule personne en minimisant le nombre total des personnes infectées (Berger-Wolf & Saia 2006). Étant donné que tous les individus peuvent être initialement infectés (Berger-Wolf & Saia 2006), le graphe statique (**Figure 31**) donne l'impression que n'importe quel individu peut être vacciné (il restera 2 personnes infectées à la fin), (Berger-Wolf & Saia 2006). Par contre, dans la première présentation (dynamique) (**Figure 31**), il semble que b est la seule personne à vacciner car, peu importe, si a, c, sont infectés d'abord, l'infection ne se propage pas selon ce premier scénario (Berger-Wolf & Saia 2006).

Besoin de données temporelles

Du fait qu'on considère tous les liens comme ils apparaissent au même temps**,** ***les représentations statiques sont trompeuses*** (ex. au niveau du processus de diffusion d'information)**.** Beaucoup de propriétés temporelles clés (Tang et al 2010a) sont absentes: la durée des contacts, 'inter-contact time', contacts récurrents, 'time order' des contacts sur un chemin, etc. Cela implique une surestimation des chemins potentiels et une sous-estimation des longueurs des plus courts chemins. En conséquence, beaucoup de métrique de SNA sont touchées et donnent ainsi des résultats trompeurs (Tang et al 2010b). Même si ce n'est pas

facile à les trouver ou les inférer, il est important que les données temporelles comme le temps de création des liens entre utilisateurs ou de suivi (ex. sur Twitter) soient intégrés dans les graphes sociaux. Sur un OSN comme Twitter le temps de création des liens n'est pas accessible par ses APIs. (Meeder et al 2011) proposent une méthode qui s'appuie sur un processus d'horodatage pour inférer le temps de création des liens Twitter. C'est une procédure à deux entrés, le temps de création de compte utilisateur (célébrité) noté C_u dans un 'Snapshot' du SN (graphe orienté) et la liste inversée des adeptes (des utilisateurs U) qui suivent C_u. F_u est donné comme le temps inconnu quand 'u' commence à suivre (crée sa relation avec la célébrité), donc $C_u \leq F_u$ (Meeder et al 2011). F_u n'est qu'une approximation (une borne inférieure) du temps réel de création de ce lien. F_u est équivalent au maximum des temps F_v tel que v ∈ B (u), $B(u) \subseteq U$ et $B(u) = \{v \in U : F_v \leq F_u\}$. L'utilisateur 'v' est appelé le 'record-breaker' pour 'u' (Meeder et al 2011). La procédure se réitère sur l'ensemble U pour identifier les record-breakers des adeptes de la célébrité et donc estimer le temps de création de ses relations dans le sous-graphe temporel 'Celebrity Follower Subgraph' résultant (Meeder et al 2011).

3.1.1. Modèles de représentation les plus connus

Se plonger dans les aspects dynamiques des SNs nécessite de définir des modèles mathématiques qui sont capables d'exprimer explicitement la composante temporelle et reproduire les propriétés observées dans les réseaux réels. Par exemple, dans le sujet du processus de diffusion d'information, des recherches antérieures ont proposé des modèles de réseau (graphe) temporel basées sur des liens étiquetés par le temps (Tang et al 2010a), (des liens pondérés par le temps de création ((Erétéo 2011))). Mais selon (Tang et al 2010a), ils ne permettent pas d'analyser les fréquences des contacts entre les nœuds ou les groupes. Une autre proposition s'est basée sur une mesure équivalente au délai de livraison entre les nœuds du graphe (Tang et al 2010a). Mais elle a monté ses limites en termes de délai global de diffusion d'information. Différentes propositions ont introduit la notion du graphe temporel (Tang et al 2010a) ou encore le graphe variant dans le temps ('Time-varying graph' TVG), (Santoro et al 2011), pour modéliser la dynamique temporelle du SN. Dans ce sens, ***on distingue deux types d'approches détaillées dans les deux sous-sections suivantes***

3.1.1.1.Séquence de traces (snapshots) de SN dans le temps

Beaucoup d'auteurs définissent le modèle d'un graphe temporel de SN dans un intervalle de temps par une séquence de graphes (traces) statiques. Chacun représente l'état (Kazienko et al 2011) du SN dans un laps de temps (Snapshot). En d'autre terme, chacun est une agrégation d'interactions, appelé l'empreinte d'un TVG dans un sous-intervalle de temps donné (Santoro et al 2011).

(Tang et al 2010a) ont formalisé une telle représentation par $G^{w}_{t\,(t\,min,\,t\,max)}$ (V, E) qui est une séquence de graphes $G_{t\,min}$, $G_{t\,min\,+\,w}$,…, $G_{t\,max}$., w est un pas ou l'unité de temps (la taille de fenêtre). Un contact entre i et j ∈V à l'instant s est noté: R_{ij}^{s} tel que $t \leq s \leq t+w^2$ (Tang et al 2010a). La figure (**Figure 32**) montre un exemple de graphe temporel $G^{1}_{t(0,3)}$ sur l'axe 'Time'.

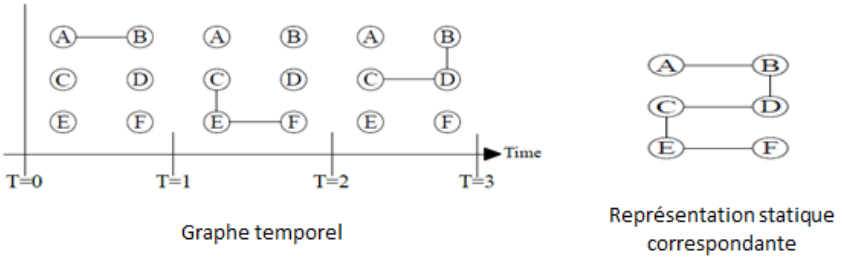


Figure 32. Exemple de graphe temporel, une séquence de traces dans le temps selon le formalisme de ((Tang et al 2010a)) **devant sa représentation statique**

La représentation statique reproduit les contacts sans considérer leur ordre temporel (Tang et al 2010a). Si le nœud A veut atteindre F ou l'inverse on suivra le chemin B, D, C, E. Cependant, il est remarquable que le contact entre A, B n'existe pas dans 2éme et la 3éme fenêtre de temps. Autrement dit, l'image donnée sur la possibilité que l'information puisse se propager entre A, F est mal exprimée dans le cas statique.

D'autre part, (Santoro et al 2011), se sont basés sur un formalisme un peu similaire en partitionnant une durée de vie T de TVG en sous-intervalles: (t_0, t_1),…, (t_i, t_{i+1}), …. Chaque (t_k, t_{k+1}) est noté pat T_i, i = 1..T (Santoro et al 2011). Par exemple, entre t_0 et t_1, une empreinte de G a été noté par $G^{(t1,\ t2)} = (V, E^{(t1,\ t2)})$. L'agrégation de toutes les interactions sur la séquence d'empreintes de G est notée par $SF(T) = G^{t0}, G^{t1}, \ldots$ (Santoro et al 2011). Cette séquence est aussi décrite comme une série temporelle (Dekker 2011).

Exemple du monde réel:

Dans la (Figure 33), (Tang et al 2010b) ont présenté un réseau de 'Enron' ((Leskovec et al 2009)) ((Klimmt & Yang 2004)) sous forme d'une séquence d'empreintes en snapshot (Tang et al 2010b). Les jeux de données de 'Enron' ((Leskovec et al 2009)) ((Klimmt & Yang 2004)) est l'un des meilleurs exemples utilisés pour modéliser et étudier l'évolution du comportement des SNs. En effet, 'The Enron Energy Corporation' était le plus grand fournisseur de gaz et services publics d'électricité aux États-Unis et même en Grande Bretagne (Tang et al 2010b) ((von Frentz 2003)). Les données ont été collectées dans le cadre de l'enquête du gouvernement américain sur le scandale de comptabilité d'Enron qui a déclaré une faillite record ((von Frentz 2003)) en décembre 2001 après avoir découvert des manipulations frauduleuses de comptabilité qui ont caché des billions de dollars de dettes (Tang et al 2010b) ((von Frentz 2003)). Ces données portent sur les échanges par emails (environ 250.000 emails) entre 151 employés de l'entreprise entre mai 1999 et Juin 2002.

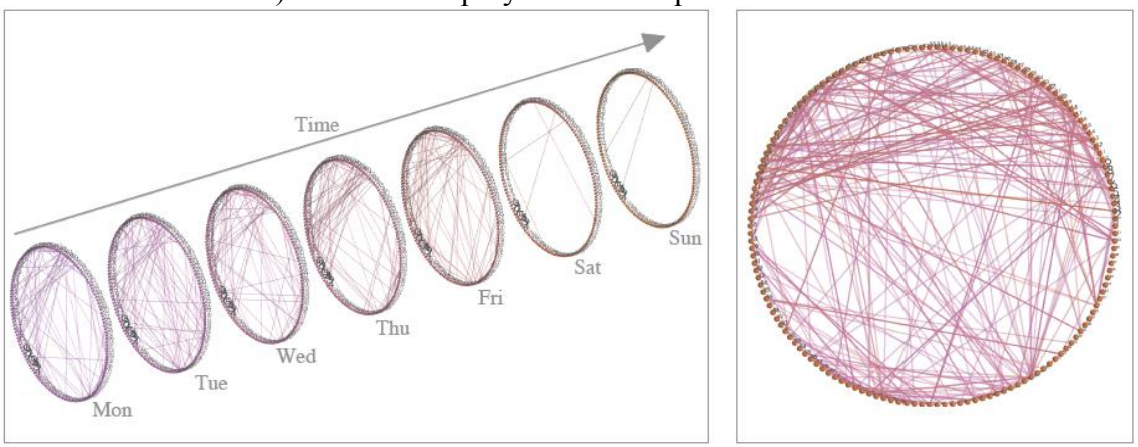


Figure 33. Les traces de communications entre les employés (SN) de 'Enron' chaque 24 heures, pendant une semaine dans le mois de novembre 2001 ((Tang et al 2010b))

Au niveau de chaque empreinte de ce SN étudié dans (Tang et al 2010b), un lien s'est créé entre 2 nœuds quand un courriel est envoyé par un employé vers un autre pendant 24h (Time-window), (Tang et al 2010b). À droite, on trouve le graphe agrégeant toutes ces interactions en une seule représentation statique (Tang et al 2010b).

3.1.1.2.Formalisme d'un graphe variant dans le temps

En SNA, proposer un formalisme d'un graphe variant le temps (Time-Varying Graph: TVG) est un pas supplémentaire pour améliorer la qualité de représentation des SNs. Il s'agit d'ajouter une ou des composantes temporelles dans le formalisme classique (ensemble de nœuds V et de liens E) d'un graphe social (séquence d'empreintes). Un éventail de modèles de réseaux dynamiques (Réseaux de transport –aviation-, de communication, sans fil mobile,

etc.) (Santoro et al 2011), évoquent le formalisme de TVG selon différents scénarios et restrictions (Santoro et al 2011).
G = (V, E, T, ρ, ξ) est un exemple de formalisme de TVG proposé par (Santoro et al 2011) pour modéliser un SN dynamique: $\rho: E \times T \rightarrow \{0,1\}$ est une fonction qui indique si un lien donné se présente dans une période de temps appelée aussi la durée de vie de système (Santoro et al 2011). $\xi: E \times T \rightarrow T$ est une fonction de latence qui indique le temps à prendre pour traverser un lien donné à partir d'une une date donnée (Santoro et al 2011). Mais généralement cette fonction n'a pas d'importance dans le cas des SNs (Santoro et al 2011). Donc TVG est défini en (Santoro et al 2011) par G = (V, E, T, ρ). ***Un nouveau formalisme de TVG, va changer beaucoup de notions et propriétés structurelles ainsi que la définition des indicateurs de SNA.***

Sous graphes temporels (chemins et géodésiques temporels)
La notion du chemin (géodésique) est cruciale dans SNA (classique). Des extensions temporelles sont proposées (Journey) en se basant par exemple sur le formalisme précédent (Santoro et al 2011). Un 'Journey' est défini par (Santoro et al 2011) comme un trajet au fil du temps : $\mathcal{J} = \{(e_1, t_1), (e_2, t_2), \dots, (e_k, t_k)\}$ tel que $\{e_1, e_2, \dots, e_k\}$ est déjà uj chemin dans G et $\forall i,\ 1 \leq i < k,\ \rho(e_i, t_i) = 1,\ t_{i+1} \geq t_i$. $|\mathcal{J}| = k$ est la longueur topologique, alors que $||\mathcal{J}|| =$ temps d'arrivé $(\mathcal{J}) -$ temps de départ$(\mathcal{J})$ représente la longueur temporelle (Santoro et al 2011). L'ensemble des chemins temporels de 'u' vers 'v' est noté par $\mathcal{J}^*(u, v)$ et l'ensemble de tous ces chemins du TVG est noté par $\mathcal{J}^*$ (Santoro et al 2011).
La géodésique entre u et v est ainsi défini par le 'Journey' minimal qui a différentes interprétations (Tableau 20).

Tableau 20. Géodésique statique et temporelle

Géodésique statique	**Géodésique temporelle à base de 'Journey' minimal** (Santoro et al 2011)
Chemin/ Distance le plus court $d(u, v)$	Distance la plus courte est le minimum des longueurs topologiques de u à v à l'instant t (Shortest) $d^t(u, v) = \text{Min}\{\|\mathcal{J}\| : \mathcal{J} \in \mathcal{J}^*(u, v) \wedge \text{départ} \geq t\}$, (Santoro et al 2011).
	Distance principale de de u à v à l'instant t (Foremost) : Min {l'arrivé$(\mathcal{J}) - t : \mathcal{J} \in \mathcal{J}^*(u, v) \wedge$ départ $\geq t\}$
	Distance la plus rapide est le minimum des longueurs temporelles de u à v à l'instant t (Fastest) : $Min\ \{\big\|\|\mathcal{J}\|\big\| : \mathcal{J} \in \mathcal{J}^*(u, v) \wedge \text{départ} \geq t\}$

Avec cette notion de chemin temporel, les chercheurs pourront envisager de définir une diversité d'indicateurs (Connectivité, centralité, etc.) temporels: Temps d'accessibilité (distance temporelle), proximité temporelle (vitesse minimale de propagation d'information), diamètre, densité, coefficient de clustering, modularité, etc. On peut même parler de 3 versions temporelles selon le type de distance 'shortest', 'foremost', fastest' (Santoro et al 2011). L'étude de l'évolution de ces indices peut se faire sur une séquence de TVG : Un sous-graphe temporel à un intervalle donné (Santoro et al 2011). Un sous-graphe G' de TVG est défini dans (Santoro et al 2011) en limitant la durée de vie T : G' = (V, E', T', ρ') tel que : $T' \subseteq T$ et $E' = \{e \in E: \exists\ t \in T' \text{ tel que } \rho(e, t) = 1\}$ et $\rho': E' \times T' \rightarrow \{0,1\} telque\ \rho'(e, t) = \rho(e, t)$.

3.1.1.3.La place des effets influant et la coévolution dans les modèles dynamiques temporels

La plupart des modèles des SNs notamment les modèles d'évolution se basent souvent sur des graphes de nœuds et des liens. Les effets influant ou menant à la création des relations sociales, les attributs des nœuds ne sont pas pris en considération (Jamali et al 2011). (Jamali et al 2011) citent l'exemple d'un modèle basé sur deux facteurs: l'influence sociale et de la sélection. (Zheleva et al 2009) ont proposé un modèle d'une coévolution comprenant

l'influence sociale et transitivité en SN (réseaux d'affiliation). Les nœuds sont affiliés à des groupes étiquetés (Zheleva et al 2009). Par exemple, l'influence sociale se présente dans les groupes à rejoindre, choisis par un utilisateur, qui sont en général parmi ceux sélectionnés par ses amis (Zheleva et al 2009).

Exemple de modèle probabiliste

Wikipédia, Épinions, Flickr, etc. sont des contextes exemplaires pour modéliser l'évolution d'un SN avec ces effets. Par exemple, les réseaux extraits de Wikipédia représentent souvent des activités d'édition d'un article donné, interactions, participations dans les discussions sur le profil d'un utilisateur. À cet égard, (Jamali et al 2011) ont mis l'accent sur des SRNs : Social Rating Networks. Pour eux, la force de chaque effet n'est pas constante pendant l'évolution d'un SRN (Jamali et al 2011). Ils ont proposé un modèle probabiliste génératif, qui modélise la force et la dynamique de chaque effet tout au long de cette évolution (Jamali et al 2011). En effet, un SRN (notation sociale) consiste en une séquence d'actions datées (Jamali et al 2011). Dans leur modèle, les auteurs (Jamali et al 2011) se sont basés sur 2 actions pour exprimer le comportement de l'utilisateur: Une action menant à la création d'une relation sociale (action sociale) et une autre d'évaluation (Une valeur binaire 1 pour évaluer une photo ajoutée). Ainsi, l'ensemble de données se présente par une séquence ordonnée A d'actions de 2 types selon un ordre chronologique. À un instant donné t, si un utilisateur 'u' exécute une action donnée, elle sera interprété selon une probabilité comme suivant (Figure 34).

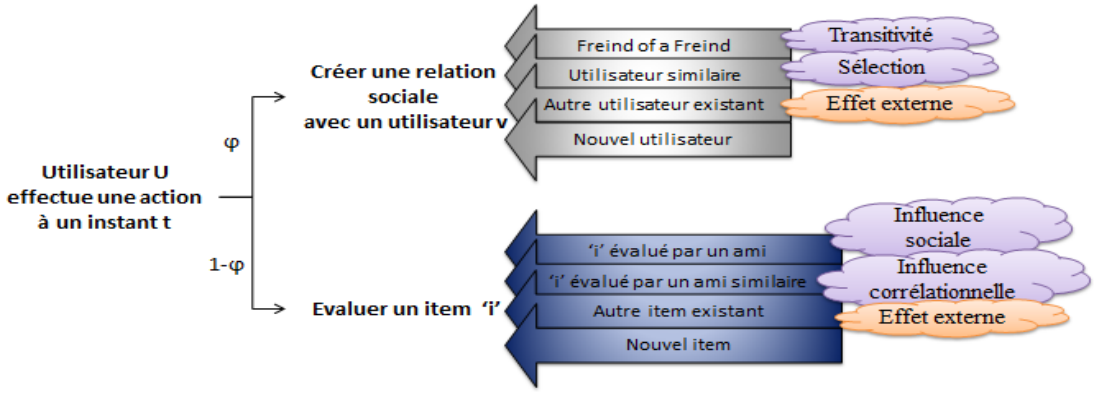


Figure 34. Aperçu sur un modèle probabiliste basé sur différents effets influant la dynamique temporelle d'un SRN proposé dans ((Jamali et al 2011))

La probabilité d'exécuter une action en créant une relation avec 'v' est φ (Jamali et al 2011). Si la transitivité est en cause, alors v doit être l'un des amis des amis de u. Si c'est la sélection cela signifie que v est l'un des utilisateurs les plus similaires à u. Sinon, il s'agit d'un effet externe inconnu influant sur cette action et dans ce cas, v peut être n'importe quel utilisateur existant ou un nouveau qui rejoint le réseau (Jamali et al 2011). D'autre part, la probabilité d'exécuter une action d'évaluation est effectivement 1-φ (Jamali et al 2011). Si l'influence sociale est en cause, alors 'i' devrai être l'un des éléments notés par les amis de 'u'. En conséquence, la valeur d'évaluation de 'u' est affectée par les valeurs exprimées par ses amis sur 'i'. S'il s'agit d'une 'influence corrélationnelle, 'i' devrai être l'un des items (articles) évalués par les utilisateurs les plus similaires à 'u'. La valeur d'évaluation de 'u' est également influencée (Jamali et al 2011). Sinon, 'i' est évalué par 'u' sous un effet inconnu. Dans ce cas 'i' peut être n'importe quelle ressource ou nouvel élément qui n'est pas encore évalué (Jamali et al 2011).

(Jamali et al 2011) ont montré à quel point un modèle probabiliste peut reproduire des SNs réalistes à partir deux datasets réelles de: Epinions et Flickr. Ils ont généré un réseau de similitude en calculant la similarité entre un ensemble d'utilisateurs. Cette similarité est estimée par la corrélation de Pearson appliquée sur les notes de leurs évaluations. Le nombre des premiers utilisateurs similaires à un utilisateur 'u' (par rapport à ses actions d'évaluation) a été trouvé le même que le nombre de ses voisins directs dans son SN (Jamali et al 2011).

Un modèle probabiliste peut servir à la prédiction des futurs liens, structures communautaire, etc. (Jamali et al 2011). La modélisation statistique permet de tester les théories sur les développement et comportements du réseau ((Snijders 2005)). Elle peut supporter l'incertitude dans les conclusions qui peuvent se généraliser sur des populations partant des données empiriques ((Snijders 2005)). Elle est particulièrement importante pour la recherche non expérimentale notamment quand on sait que certains effets sont difficiles à contrôler ((Snijders 2005)) avec des nœuds et des liens seulement.

3.1.2. Bilan

Aujourd'hui les représentations statiques des SNs ne répondent pas aux exigences des nouvelles tendances de SNA en parallèle avec l'émergence des nouveaux SNs et la richesse des données sociales. Quand il s'agit d'étudier la dynamique temporelle, deux types d'approches de modélisation ont été distingués.

Tableau 21. Comparatif entre deux approches de modélisation de SNs dynamiques

Modèles à base de séquence de traces d'empreintes	Formalisme de graphe variant dans le temps
Des indicateurs intemporels sont appliqués sur des séquences d'empreintes statiques (par fenêtre de temps) donnent des informations précieuses sur l'évolution.	Envisager de définir des extensions d'indicateurs et mesures temporels basés sur les chemins et les sous-graphes temporels
Des représentations simples	Les modèles sont plus riches et plus complexes.
Le temps est discret : La composante du temps n'est pas suffisamment explicite pour comprendre des phénomènes complexes (persistance)	La composante du temps est plus explicite.
On peut avoir un chemin entre x et y dans toutes les empreintes de SF	Possibilité de ne pas avoir un seul Journey entre x et y
Des changements microscopiques peuvent passer inaperçu entre 2 fenêtre de temps. Souvent, les modèles ne fournissent pas des informations sur le retard du processus de diffusion d''information (Tang et al 2010a).	Possibilité d'estimer le temps d'accessibilité, vitesse de propagation d'information, prédiction de relations (Jamali et al 2011) , etc.

Quand un modèle est proposé pour représenter la dynamicité temporelle des SNs, l'ensemble des nœuds peut être supposé consistant (uniforme) pendant les différentes périodes d'observation. Cependant, les réseaux dynamiques peuvent pratiquement présenter des ensembles variables de nœuds dans le temps. Par ailleurs, la plupart des modèles dynamiques apparaissent comme des modèles d'évolution. Les nœuds et les liens peuvent être ajoutés, mais pas toujours supprimés au cours du temps (Berger-Wolf & Saia 2006).
En dehors de ces deux types d'approches de modélisation, on trouve que 'The Dutch Soccer Team as a Social Network' de ((Kooij et al 2009)) est l'un des modèles de SN exceptionnels qui semblent être statiques, mais peuvent porter de l'information temporelle. Ici les nœuds représentent des joueurs de générations successives dans le temps, ce qui a permis à ((Kooij et al 2009)) d'examiner l'évolution des propriétés topologiques au fil du temps.

Le calcul et les outils analytiques de la dynamique temporelle constituent aussi un autre défi. En général, la recherche en modélisation et analyse des réseaux dynamiques varient. D'un point de vue mécanique statistique, ce type de recherche considère les réseaux comme des systèmes physiques complexes. On s'intéresse à décrire les lois qui régissent leur évolution, le comportement et leurs propriétés (Berger-Wolf & Saia 2006), comme dans les deux approches. Un point de vue plus calculable intègre les probabilités et l'incertitude dans les structures sous l'influence de certains effets dynamiques comme le modèle probabiliste de (Jamali et al 2011). On peut avoir un aperçu sur ces influant, par exemple l'effet de

transitivité a été trouvé beaucoup plus important que la sélection dans la création de relations sociales sur Epinions et Flickr (Jamali et al 2011). Des résultats comparables avec les autres modèles de la littérature ont été obtenus par des métriques d'évaluation appropriées (Jamali et al 2011). Ce type d'approche calculable combine récemment l'analyse d'un SN avec les systèmes multi-agents. Dans ce sens, les simulations constituent la principale technique de calcul pour intégrer ces informations (Berger-Wolf & Saia 2006). Dans les modèles à base d'agent (entité sociale, groupe), l'agent est un ensemble de règles pour décider et agir ((Scott 2011), Social network analysis: developments, advances, and prospects). Son action se déclenche en concaténant l'ensemble des conséquences d'actions des autres sur le réseau ((Scott 2011), Social network analysis: developments, advances, and prospects). Par conséquent, la connaissance des règles décrivant comment l'agent va agir, peut servir à prédire le changement de la structure du réseau.
En revanche, la dynamicité simultanée entre le réseau et le comportement (la coévolution) n'est pas encore claire. Outre, les modèles les plus connus montrent globalement la dynamicité temporelle du SN comme le résultat de changement de comportements d'acteurs basés sur les relations. Or, ce n'est pas exactement l'image réelle. Des éléments importants et des changements implicites qui s'impliquent dans le comportement de l'acteur ne sont pas pris en compte dans un modèle dynamique du SN.

3.2.Analyser et fouiller la dynamique temporelle du SN

Après avoir défini un motif d'analyse sérieux, la pertinence et la richesse des données sociales vont aider à définir un modèle dynamique de SN plus réaliste. La dimension temporelle est aujourd'hui censée être une priorité dans les travaux de recherche et les applications récentes de SNA. Par la suite, des exemples illustratifs seront abordés pour montrer comment étudier la dynamique des SNs évoluant dans le temps suivant les deux types d'approches de modélisation décrites ci-dessus. La dynamique temporelle sera abordée selon différents points de vue (mesures et technique de SNA) et niveaux de granularité (l'évolution des individus et des groupes).

3.2.1. Analyser l'évolution avec des indicateurs atemporels sur une séquence d'empreintes dans le temps

Selon différentes études (McGlohon & Faloutsos 2008), beaucoup d'indicateurs de SNA (classique) peuvent être appliqués directement sur une séquence de traces (graphes) d'un SN évoluant dans le temps. Comme montré dans (Figure 30), le suivi de la densité, loi de puissance, diamètre, modularité, coefficient de clustering, etc. dans le temps permet d'expliquer le phénomène d'évolution au niveau de la structure globale du SN. Même s'il s'agit d'un concept qui est souvent lié aux communautés, la conductance du graphe social, mesurée dans le temps est aussi informative. La conductance d'un graphe est tout d'abord définie comme la conductance minimale sur toutes les coupures (Cut) possibles (Santoro et al 2011). Au niveau du processus de diffusion d'information, l'évolution de cette mesure caractérise le temps de convergence d'une marche aléatoire vers sa distribution uniforme (Santoro et al 2011). Elle reflète indirectement un processus d'auto-optimisation (ou de détérioration) de l'efficacité du réseau (Santoro et al 2011).

L'évolution des centralités (popularités) individuelles

Les changements d'interactions dans un SN ont un impact direct sur l'évolution des rôles/ positionnements (centralités individuelles) des acteurs dans le temps. Dans un cas d'étude, (Dekker 2011) a analysé l'évolution des scores de centralités individuelles dans un groupe (SN) de discussion (commentaires) internet (sci.math) dans le cadre d'un atelier de

planification gouvernementale (Dekker 2011). Il a suivi le changement de centralités (normalisées) de 6 principaux participants, les plus centraux (Dekker 2011).

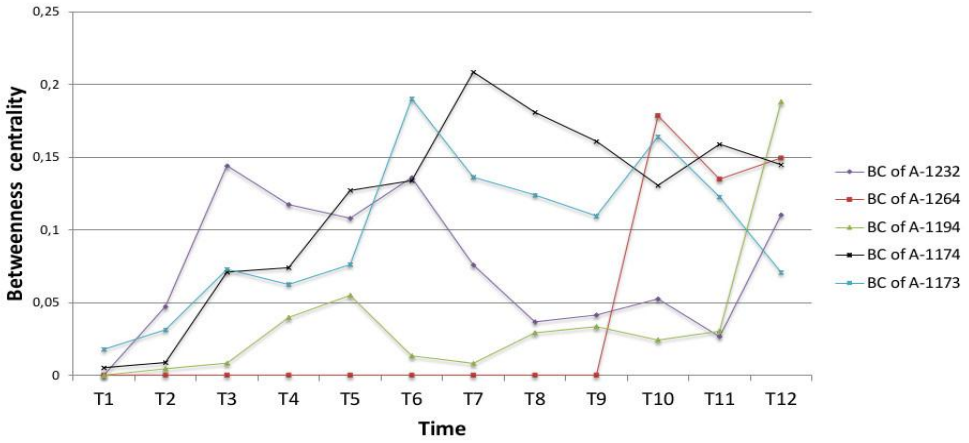


Figure 35. Evolution des scores de centralités (intermédiarité) des acteurs les plus centraux

D'autre part, sur un échantillon de SN de communication par mails entre des employés de 'Enron' ((Leskovec et al 2009)) ((Klimmt & Yang 2004)), nous montrons également (**Figure 35**) le changement de l'intermédiarité des acteurs centraux dans le temps. Il est remarquable par exemple que A-1174 est l'acteur qui domine la communication le plus, perdant 5 'time-steps'. Avec telles étude, les animateurs/ managers auront une rétroaction devant le flux dynamique de leurs ateliers ou autres processus organisationnels sociaux pareils (Dekker 2011). Par exemple, l'animateur a besoin d'identifier ceux qui travaillent ou sont actifs constamment et qui présentent un modèle plus équilibré de participation (Dekker 2011).

La hausse et la baisse de la potentialité des individus ont un rapport avec des évènements du monde réel. Après avoir inféré approximativement le temps de création des liens sur Twitter, (Meeder et al 2011) ont proposé une formule (une probabilité $f_i(t)$) pour évaluer la popularité d'une célébrité i dans le temps. Ils ont suivi la popularité du top 50 des célébrités sur une séquence de snapshots. La **Figure 36** montre comment la popularité des 5 célébrités les plus populaires varie dans le temps et comment elle est loin d'un attachement aléatoire (courbe noire).

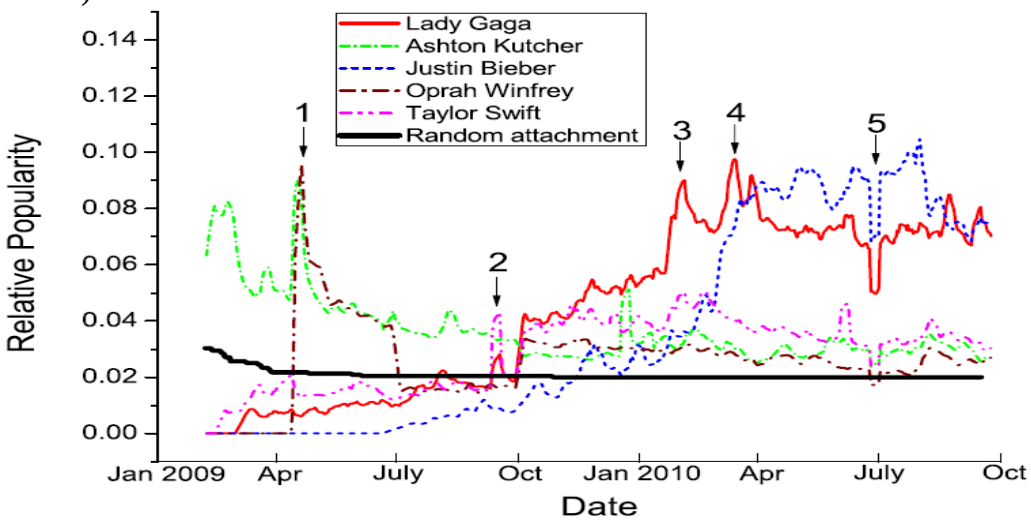


Figure 36. La popularité de 5 célébrités sur Twitter, mesurée par ((Meeder et al 2011)) **en fonction du temps**

Par exemple Oprah Winfrey a affiché un pic de popularité plus rapidement que Justin Bieber qui commence avec une popularité faible mais elle progresse dans le temps (**Figure 36**).

Plusieurs pics et baisses de popularité sont identifiés sous l'influence des événements du monde réel qui expliquent plausiblement ces changements. Par exemple les deux pics de popularité (3 et 4) de Lady Gaga correspondent respectivement à sa nomination pour les récompenses Emmy de télévision américaine (le 1 février 2010) (Meeder et al 2011) et à la sortie de son vidéo clip 'Telephone' le 13 Mars 2010 (**Figure 36**). Il y a des évènements qui impliquent deux ou plusieurs célébrités, ce qui explique l'apparition des pics de popularité simultanés (**Figure 36**). Le vendredi 17 Avril 2009 Ashton Kutcher qui a réussi à atteindre un million d'adeptes sur Twitter (pic 1) est apparu dans l'émission d'Oprah au cours de laquelle Oprah Winfrey a rejoint Twitter ((Oprah 2009)). Cet évènement a contribué à augmenter la popularité des deux et à une plus forte hausse jamais vue dans le nombre des comptes Twitter ce jour-là (Meeder et al 2011). La même augmentation simultanée a été remarquée pour Lady Gaga et Taylor Swift (le pic 2) en participant ensemble à l'évènement 'MTV Video Music Awards'. Dans cette soirée, Kanye West a été impliquée dans un infâme incident (interruption de discours d'acceptation de Taylor Swift) qui aurait pu baisser sa popularité mais il n'était pas sur Twitter à ce moment.

3.2.2. Analyser l'évolution par des mesures temporelles à base d'un formalisme de TVG

Si des nouveaux formalismes de TVG est une solution pour modéliser la dynamique temporelles des SNs, les chercheurs devraient définir des extensions de mesures de SNA applicables sur tels modèles. On parle d'indicateurs temporels qui requirent souvent de remplacer la notion de chemin classique par le chemin temporel (Journey). L'évolution de ces indices est plus compliquée et l'une des solutions propose de l'étudier sur une séquence de sous-graphe temporel (ST) (Santoro et al 2011). Prenant l'exemple de formalisme de TVG et Journey décrits ci-dessus, certains indicateurs ont été définis comme suivant.

3.2.2.1.Distance temporelle

Il ne s'agit pas seulement d'un concept de base pour la plupart des métriques temporelles. Le calcul et l'évolution des distances temporelles donnent un aperçu sur le processus de diffusion d'information (la vitesse de diffusion), la proximité/ l'accessibilité des nœuds (Santoro et al 2011). Comme montré dans le (Tableau 20), la distance (géodésique) temporelle peut avoir 3 définitions (Santoro et al 2011).

D'autres auteurs définissent la plus courte distance temporelle entre le nœud 'i' et 'j' par d^h_{ij}(t min, t max) parmi les l'ensemble des chemins noté p^h_{ij}(t min, t max) (Tang et al 2010a): h est le temps ou les sauts temporels à prendre pour que l'information se propage de i à j (le temps maximal est appelé l'horizon). Dans le cas où la paire de nœuds est temporellement déconnectée (Tang et al 2010b), l'information de i n'atteint jamais j. donc la distance temporelle $d_{ij} = \infty$ (Tang et al 2010a).

Les auteurs se servent par exemple de l'algorithme DFS: 'Depth First Search' pour calculer ce type de distance entre chaque paire de nœuds à partir deux listes D et R (Tang et al 2010a): D est une liste qui compte le nombre de sauts temporels nécessaire pour atteindre chaque nœud 'i'. R indique si les nœuds sont atteints ou pas. D et R sont initialisés respectivement par '1' et 'False'. Pour chaque fenêtre de temps et à partir d'un nœud source i tel que R(i) = T. une recherche en profondeur (DFS) est effectuée pour vérifier si les nœuds non atteints ont un chemin vers un nœud qui est atteint dans une fenêtre précédente (Tang et al 2010a). Si le nœud j est accessible, R(j) = 'True' sinon la distance D(j) est incrémentée (Tang et al 2010a).

3.2.2.2.Excentricité (Accessibilité temporelle)

L'excentricité e(u) d'un nœud 'u' dans un TVG : $e(u) = Max\{d^t(u,v), v \in V\}$ représente sa capacité d'accessibilité étant donné que la distance temporelle est le 'Shortest' (Santoro et al

2011). Les autres extensions 'Foremost' ou 'Fastest Eccentricity' sont utiles dans des phénomènes importants comme les épidémies. Des nœuds ayant une forte excentricité temporelle sont associés à la probabilité qu'un virus survive à très long terme pour réinfecter des gens (Santoro et al 2011).

3.2.2.3.Diamètre temporel

Il se base sur l'excentricité avec 3 versions : $\max\{e(u), u \in V\}$ (Santoro et al 2011). Le diamètre temporel représente aussi la diffusion maximale de l'information dans le réseau au fil du temps (Tang et al 2010a). Selon (Santoro et al 2011), l'évolution de diamètre/ excentricités temporels n'a pas retenu suffisamment d'attention pour le moment (Santoro et al 2011) même si elle peut révéler des paramètres sociaux complexes.
Outre, des nouveaux indicateurs temporels (globaux) sont proposés comme :

3.2.2.4.L'efficacité temporelle

L'efficacité temporelle est inversement proportionnelle à la distance temporelle (plus la distance est grande plus l'efficacité est faible) (Tang et al 2010a). Elle est notée dans (Tang et al 2010a) par :

$$\mathrm{E}_{\mathrm{ij}}^{\mathrm{h}}\,(\mathrm{t}_{min}, \mathrm{t}_{max}) = \frac{1}{\mathrm{d}_{\mathrm{ij}}^{\mathrm{h}}(\mathrm{t}_{min}, \mathrm{t}_{max})} \qquad (\mathbf{30})$$

Si les nœuds sont temporellement déconnectés alors E = 0. D'où l'efficacité globale (Tang et al 2010a) :

$$\mathrm{E}_{\mathrm{glob}}^{\mathrm{h}}(\mathrm{t}_{min}, \mathrm{t}_{max}) = \frac{1}{\mathrm{n}(\mathrm{n}-1)} \sum_{ij} \mathrm{E}_{\mathrm{ij}}^{\mathrm{h}}\,(\mathrm{t}_{min}, \mathrm{t}_{max}) \qquad (\mathbf{31})$$

3.2.2.5.Centralité temporelle

Regardant l'état de l'art courant, l'influence individuelle des acteurs est souvent évaluée par des métriques de centralité définies sur des modèles statique (une seule agrégation) ou des agrégations du réseau dans le temps). Cependant, il est également important d'évaluer ces métriques sur une topologie de SN/OSN évoluant dynamiquement dans ces systèmes (Tang et al 2010b). L'adaptation temporelle des concepts de centralité devient significative (Santoro et al 2011) à partir du moment où les nœuds centraux changent de statut d'un point de vue temporelle. D'où la notion de centralité temporelle a émergé (Santoro et al 2011).

3.2.2.5.1. L'intermédiarité temporelle (Temporal Betweenness: Tb)

En gardant le concept général de l'intermédiarité statique, l'intermédiarité temporelle d'un nœud 'q' désigne une fraction de plus courts chemins temporels traversant 'q' (Tang et al 2010b). L'une de ses versions est définie comme suivant (Santoro et al 2011) :

$$\mathrm{Tb}\ (\mathrm{q}) = \sum_{v \neq u \neq q \in V} \frac{|d'(u, v, q)|}{|d(u, v)|} \qquad (\mathbf{32})$$

Tel que : $|d'(u, v, q)|$ est le nombre des géodésiques temporelles (Shortest Journeys) qui passent par 'q' (Santoro et al 2011). C'est un indice qui est aussi utilisé pour distinguer les individus agissant comme médiateurs clés sur les chemins de communication dans le temps. Mais aussi, son calcul ne se contente pas seulement du nombre des géodésiques passantes par un nœud. Ici, pour calculer la centralité d'intermédiarité, les auteurs considèrent aussi la longueur du temps dans lequel un nœud sur ces chemins conserve un message avant de le

transmettre à un autre nœud (Tang et al 2010b). Ça permettra de répondre aux questions posées sur la perturbation et l'efficacité de communication :
Soit $p_{kj} = (j^{t0}, i^{t1}, k^{t2})$ le seul plus court chemin entre j et k, tel que la transmission d'un message de j vers k doit passer par i au temps t1 (Tang et al 2010b). i joue un rôle important de médiateur dans la communication entre j et k, mais sa vulnérabilité dépend fortement des intervalles [t0, t1], [t1, t2]. Plus le message transmis par j prend du temps (d'attente) sur i, il y a plus de chance de perturber la transmission vers k (Tang et al 2010b). D'où, Tang et al, définissent la centralité d'intermédiarité de i comme :

$$\mathrm{CB_i(t)} = \frac{1}{(n-1)(n-2)} \sum_{\substack{j \in V \\ j \neq i}} \sum_{\substack{k \in V \\ k \neq i \\ k \neq j}} \frac{U\,(i,t,j\,,k)}{|S^h_{jk}|} \qquad (\mathbf{33})$$

C'est une fraction du nombre des géodésiques temporelles qui passent par i entre chaque paire j et k à l'instant t : U (i, t, j, k), par rapport le nombre total des géodésiques $|S^h_{jk}|$ (Tang et al 2010b). Il y a deux situations, soit i reçoit un message de j allant vers k à l'instant t ou bien il conserve déjà ce message après une fenêtre de temps jusqu'à rencontrer le prochain nœud à t1 > t (Tang et al 2010b). Si S^h_{jk} est vide, le nœud i est totalement isolé : $CB_i(t) = 0$ (Tang et al 2010b).

D'un autre côté, ((Min-Joong et al 2016)), proposent l'une des versions les plus récentes qui adapte le problème de calcul de centralité d'intermédiarité (algorithme de (Brandes 2001) adapté) sur des graphes entièrement dynamiques. Ce qui réduit aussi considérablement le nombre les plus courts chemins qui devraient être recalculée tant que le graphe change.

3.2.2.5.2. La proximité temporelle (Temporal Closeness: Tc)

Suivant le principe de sa version statique, la proximité temporelle d'un nœud 'u' est définie dans (Santoro et al 2011) comme suivant:

$$\mathrm{Tc(t)} = \sum_{v \in V \setminus u} \frac{d^t(u,v)}{|\{w \in V : \exists\, \mathcal{J} \in \mathcal{J}^*(u,w)\}|} \qquad (\mathbf{34})$$

C'est une moyenne des distances géodésique temporelle vers les autres nœuds. Ce concept est fortement lié à l'excentricité temporelle, et pourtant, chacun a émergé dans un domaine de recherche différent. Selon le point de vue de diffusion d'information, la proximité temporelle mesure à quelle vitesse un utilisateur peut diffuser une information (Tang et al 2010b). Ce qui est praticable dans le marketing viral et l'étude de la propagation de rumeurs (Tang et al 2010b).

Récemment, des auteurs comme ((Taylor et al 2016)) ont osé toucher les mesures de centralité vectorielle ('eigenvector-based centrality measures') pour les étendre sur des réseaux variant dans le temps.

3.2.3. Bilan sur les métriques temporelles et atemporelles

Même si les métriques de SNA sont à l'origine statiques, elles donnent des informations précieuses sur l'évolution, le faite d'être appliquées sur une séquence d'empreintes de SN dans le temps. Cependant, quand la composante de temps devient plus explicite dans les modèles de SNs (Formalismes de TVG), ces indicateurs peuvent être remplacés par des extensions temporelles et même des nouveaux concepts. Jusqu'à présent, les métriques temporelles notamment globales sont proposées pour caractériser les réseaux dynamiques en général et les processus de diffusion d'information.

Capturer la dynamique de l'ensemble du SN n'est pas le seul objectif (Tang et al 2010a). (Tang et al 2010a) montrent que les métriques temporelles disposent d'un avantage comparatif devant leurs homologues statiques. Ils ont testé deux graphes sociaux variant dans le temps. Le premier se base sur la connectivité des périphériques mobiles. Le deuxième est formé par deux types d'interactions au sein d'un groupe Facebook géographiquement limité (à Londres) : Poster un contenu sur le mur d'un utilisateur, ou commenter une de ses photos publiées (Tang et al 2010a). Dans une autre étude les auteurs (Tang et al 2010b), ont évalué des centralités (proximité (C), intermédiarité (B), degré (D)) temporelles (T) sur le jeu de données d'Enron ((Leskovec et al 2009)) ((Klimmt & Yang 2004)): SN dynamique, en comparant les résultats avec leurs homologues statiques (S). Les résultats ont été bénéfique aussi bien pour enquêter le scandale d'Enron (en identifiant les acteurs centraux de l'entreprise), que pour montrer les différences entre une analyse statique et temporelle (Tang et al 2010b) :

Dans le cas statique, le classement des acteurs les plus centraux est quasiment le même avec SC, SB, SD. L'analyse statique ne favorise que ceux qui interagissent avec le plus grand nombre d'acteurs (Tang et al 2010b), sachant que le degré influence considérablement sur les autres centralités. Ce n'est pas le cas avec les versions temporelles ou le classement est différent d'une métrique à une autre.
En outre, les centralités statiques sont montrées fortement corrélées (En utilisant le coefficient de corrélation de Kendall-tau) par rapport les centralités temporelles (Tang et al 2010b). On peut déduire que: $\forall \text{Cen}_1, Cen_2 \in \{D, C, B\}, Cor(SCen_1, SCen_2) > Cor(TCen_1, TCen_2)$

D'un autre côté, les employés centraux dans le cas statique (le secrétaire, le directeur général) ne sont pas les mêmes dans le cas temporel (Tang et al 2010b). Pour découvrir le résultat le plus réaliste, les auteurs ont simulé le rôle du top N d'acteurs centraux (top N) dans deux processus dynamiques: Diffusion et de médiation de l'information par analogie avec le processus de vaccination contre les épidémies (Tang et al 2010b). Dans ce cas, la diffusion se caractérise par la contagion. Il s'agit de mesurer le taux de propagation d'information (la contagion) sur le réseau avant et après que les nœuds centraux (le top N) soient supprimés (vaccinés contre la contagion) (Tang et al 2010b). Ce sont les nœuds centraux selon TC (les plus accessibles) et TB (les médiateurs). La suppression des médiateurs génère une baisse globale de diffusion et d'efficacité de communication (Tang et al 2010b) par rapport le cas statique (Tang et al 2010b).

Les mesures de centralité temporelles complètent l'ensemble des indicateurs atemporels. Elles sont efficaces pour expliquer des phénomènes et des processus dynamiques (diffusion et médiation d'information). Par exemple la notion de vitesse (distance temporelle) et d'accessibilité permettent de découvrir les nœuds importants dans la diffusion d'information (Tang et al 2010b). D'autre part, la centralité temporelle montre que la médiation d'information est aussi vulnérable au temps et non seulement au positionnement (sur une proportion importante de canaux de communication). Cependant, ces propriétés d'évolution sont étudiées à une échelle de temps raisonnable (Santoro et al 2011), tandis qu'il reste difficile de les expliquer sur un SN qui évolue sur des périodes d'observation plus longues.

Tableau 22. Métriques temporelles et statiques

Extensions temporelles	**Métriques de SNA (statiques/ atemporelles)**
Elles sont définies sur un modèle de SN dynamique basé sur un formalisme de TVG	Elles sont d'origine statique définies sur un graphe agrégeant toutes les interactions. Elles sont applicables sur une séquence d'empreintes

	(statique) de TVG.
Le calcul de distance temporelle révèle des chemins temporels (géodésiques) asymétriques. D'où l'accessibilité temporelle est aussi asymétrique (Tang et al 2010a).	Une estimation parfois trompeuse de la symétrie de chemins. La distance géodésique entre A et B est souvent symétrique (Mise à part le cas orienté)
Elles s'appliquent sur des chemins qui se forment temporairement (Des nœuds temporellement connectés)	Le nombre des paires de nœuds connectés (les chemins) est surestimé. Donc une estimation trompeuse de la connexité de réseau. Ainsi, la longueur des chemins est sous-estimée ce qui donne des résultats imprécis
Une faible corrélation entre les métriques. On ne favorise pas seulement des nœuds qui interagissent beaucoup avec d'autres nœuds	Une forte corrélation entre métriques. Le degré influence sur les autres métriques. Souvent, on ne favorise que les acteurs ayant des degrés élevés
Plus de précision dans la détection des acteurs clés. Les nœuds centraux préservent des degrés élevés tout au long de la période d'observation	Une estimation moins précise: Un acteur central n'exprime pas tout le temps un degré élevé.
Les nœuds centraux ont plus d'influence sur les processus dynamiques de diffusion et de médiation d'information	Les nœuds centraux ont moins d'influence par rapport le cas temporel.

3.2.4. Dynamique des communautés (groupes)

L'incorporation de l'aspect temporel dans le sujet de détection des structures communautaires en SNAM suscite un intérêt croissant dans les domaines de sociologie et de sciences comportementales (Gilbert et al 2010). La découverte des communautés dans les SNs dynamique est un domaine émergent (Parthasarathy et al 2011). À ce niveau-là, il ne s'agit pas d'analyser juste l'évolution du SN, mais c'est une fouille plus profonde, menant à étudier aussi la dynamique des groupes : l'évolution du comportement de collectivité au fil du temps. La dynamique des communautés/ groupes est un niveau de dynamique plus élevé et complexe et attire l'attention des chercheurs en SNAM et dans plusieurs autres sujets en sciences comportementales (Gilbert et al 2010) (Cuvelier & Aufaure 2011). Beaucoup de question se posent: S'il n y a pas un accord complet sur la définition des communautés statiques, comment se forment-elles dans le temps ? Comment peut-on les découvrir dans un SN dynamique ? Comment évoluent/persistent-elles au fil du temps ?

3.2.4.1.Notions strictes en extensions basées sur un modèle de TVG

Certaines propositions ont tenté d'étendre des notions théoriques (composante) liées aux collectivités pour définir une communauté 'temporelle' sur des formalismes de TVG. Tant que le 'time ordering' des contacts n'est pas considéré, la connexité/ la connectivité statique est trompeuse et ne permet pas de capter pas des vraies composantes connexes ou connectées. Etant l'un des critères important pour décrire le concept de communauté statique, l'accessibilité a été ciblée par (Tang et al 2010a) en proposant des définitions étendues de ce concept et les notions liées sur un modèle de TVG (Tang et al 2010a). Les propriétés des chemins temporels vont affecter directement les notions liées à l'accessibilité, notamment:

- **Définition** 29. Une composante temporellement connectée est un sous-ensemble de nœuds où il y a un chemin temporel entre chaque paire.
- **Définition** 30. Out-component temporelle notée $OUT_{ti}^{h}(t_{min}, t_{max})$ d'un nœud i est le sous-ensemble de nœuds que i peut atteindre dans un intervalle de temps $[t_{min}, t_{max}]$.
- **Définition** 31. In-component temporelle notée $IN_{ti}^{h}(t_{min}, t_{max})$ d'un nœud i est le sous-ensemble de nœuds qui peuvent atteindre i dans un intervalle de temps $[t_{min}, t_{max}]$.

Etant donné que les chemins temporels sont asymétriques, l'accessibilité peut se décrire par une composante faiblement ou fortement connectée temporellement. Si un nœud n'appartient qu'à une seule composante statique, (Tang et al 2010a) montrent que les composantes connectées temporellement se chevauchent.

3.2.4.2.Communautés dynamiques sur une séquence de snapshots

Nombreux travaux récents qui portent sur la découverte des communautés temporelles partagent en commun une idée clé qui consiste à enfiler une découverte de communautés sur des snapshots (empreintes) d'un SN dans le temps. La principale raison est que la majorité des techniques de partitionnement ne supportent pas l'aspect temporel. Toutefois, des approches récentes ont abordé différents aspects et propriétés d'évolution en proposant des cadres et des modèles appropriés et en étendant les notions (ex. connectivité) et les paramètres (ex. modularité) connus dans la détection classique de communautés (Berger-Wolf & Saia 2006) (Parthasarathy et al 2011). Par exemple, ((Min-Joong et al 2016)) ont récemment adapté l'algorithme de (Newman & Girvan 2004), (Parthasarathy et al 2011), en adaptant notamment le calcule de centralité d'intermédiarité des liens (algorithme de Brandes adapté) sur des graphes entièrement dynamique.

Sur ce plan dynamique la définition d'une communauté est plus flexible. D'ailleurs, les concepts de groupe et de communauté peuvent se distinguer. Selon (Tantipathananandh et al 2007), un groupe capture les interactions affichées à un moment donné par une communauté qui est un concept plus latent. Cela se confirme par la définition de (Reda et al 2009):

- **Définition 32.** Une communauté temporelle est un regroupement d'individus qui persistent au fil du temps, tout en permettant à des nouveaux membres de joindre ou de sortir de cette communauté vers d'autres (Reda et al 2009).

Suivant une approche basée évènements, ((Asur et al 2007)), ont structurée des évènements critiques qui caractérisent le comportement (l'évolution) des communautés et leurs membres. Il y a des évènements impliquant des communautés: 'k-merge', 'k-split', 'form' et 'dissolve' comme il y a des événements individus: apparaître, disparaître et rejoindre ((Asur et al 2007)). D'où, on parle de l'affiliation chronologique des individus.

3.2.4.2.1. Affiliation chronologique des individus

L'affiliation chronologique des individus présente un dernier niveau de granularité permettant d'expliquer facilement la formation et l'évolution des groupes et des communautés dans le temps.

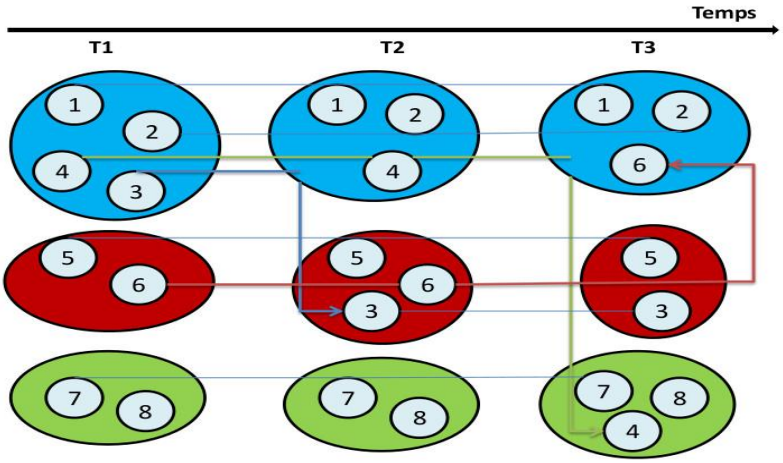


Figure 37. La dynamique des groupes. Des individus qui persistent, autres rejoignent ou quittent la communauté dans le temps

La **Figure 37** montre 3 communautés (bleu, rouge, vert) qui évoluent dans le temps, chacune possède une trace (groupe) à un time-step. Il remarquable que la paire des individus (1,4) persiste dans la même communauté (bleu). (7, 8) persiste dans la communauté en vert. Ces sous-ensembles (groupes/sous-groupes) qui persistent dans le temps constituent le 'noyau' ou l'identité de chacune de ces communautés. Les chercheurs pensent à représenter la dynamique de ces groupes en décrivant le déroulement d'affiliation chronologique des individus dans le temps (Reda et al 2009).
Comme ils ont fait (Reda et al 2009), on peut représenter aussi le déroulement de cette affiliation chronologique **Figure 37** de chaque individu en lui associant à une ligne. Cette ligne peut être contiguë pour représenter son départ ou son arrivé à une communauté dans les time-steps (Reda et al 2009). Lorsqu'un individu change son affiliation, sa ligne est intégrée en faveur de la nouvelle communauté.

Le suivie de l'appartenance des individus au fil du temps est l'une des tendances pour expliquer et comprendre la dynamique des communautés (Zhou et al 2007). Dans ce sens, (Kang et al 2007) ont proposé par exemple l'outil C-GROUP (Kang et al 2007) qui est capable d'explorer visuellement l'évolution des groupes à partir d'une paire de nœuds sélectionnés dans le SN. En naviguant sur un intervalle de temps, il permet de percevoir l'évolution de leur appartenance aux groupes, les groupes partagés et non partagés entre les deux (Kang et al 2007). Autres chercheurs ont encadré l'étude de l'appartenance des individus aux groupes dynamiques par un ensemble d'hypothèse et des propriétés (Tantipathananandh et al 2007).

Si les individus changent d'affiliation ils animeront la dynamique des groupes, sinon les groupe persiste et donne un sens à la communauté temporelle. (Zhou et al 2007) donnent ***une définition moins discrète*** des communautés temporelles (Zhou et al 2007).

- **Définition 33.** Une communauté temporelle est une séquence de communautés statiques. La structure de chacune à un moment donné dépend de N snapshots précédents.

D'où ils proposent un partitionnement conditionné par l'historique d'appartenance (d'affiliation) pour les détecter (Zhou et al 2007).

3.2.4.2.2. Partitionnement conditionné par l'historique d'appartenance

Il y a un type d'approches alternatives qui ont une vue plus globale de la découverte de communauté sur une séquence de snapshots. (Zhou et al 2007) proposent une approche pour découvrir des communautés temporelles mais cette fois-ci dans les réseaux hétérogènes (basés sur des documents). Les auteurs affirment qu'un SN hétérogène a également une structure dynamique dans le temps et la modélisation la mieux adaptée se présente par une séquence de snapshots d'un graphe triparti (**Figure 38**)

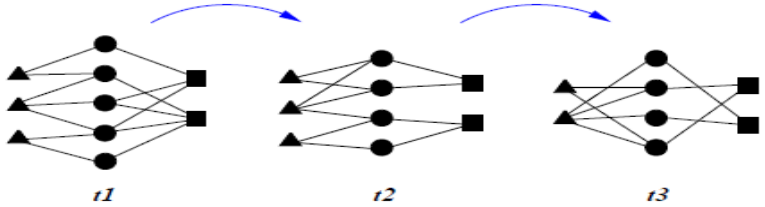


Figure 38. Un SN hétérogène dynamique. Dans 3 snapshots, l'ensemble des auteurs X (triangles), des lieux Y (rectangles) et des mots Z (des cercles) varient dans le temps ((Zhou et al 2007))

Le principe consiste à exécuter un regroupement (partitionnement) périodique dans des périodes de temps consécutives en utilisant la sortie de la période précédente comme connaissance préalable (Zhou et al 2007). A chaque période de temps t_i, cette connaissance se

réfère à l'historique d'appartenance des sommets aux communautés précédentes (Zhou et al 2007). Un historique d'appartenance va coder non seulement les informations de la période qui précède immédiatement t_i. C'est aussi une combinaison d'informations provenant de toutes les périodes précédentes (Zhou et al 2007). Cependant, la période initiale ajoute une contrainte supplémentaire. Par ailleurs, l'ensemble des nœuds peut être consistant comme il peut évoluer dans le temps. Dans ce cas il n'y aura aucune connaissance préalable sur l'appartenance des nouveaux nœuds (Zhou et al 2007). Par conséquent, il s'agit d'un partitionnement par contraintes. Dans le cas statique, on sait qu'une méthode spectrale de partitionnement est utilisée pour enregistrer l'appartenance des nœuds: Un vecteur associé à chaque nœud pour enregistrer son appartenance à k communautés par 0 ou 1 (Zhou et al 2007). Les auteurs ont dû généraliser sa fonction coût (à minimiser) sur le cas des réseaux hétérogènes tripartis et l'adapter pour supporter la dimension temporelle (Zhou et al 2007). Le partitionnement a été formulé comme un problème NP-difficile (quadratically constrained quadratic programming problem) dont la solution approximative a été donnée par un algorithme appelé 'fractional subspace iteration' (fsi) (Zhou et al 2007) suivi par l'exécution du 'K-means' pour regrouper les objets hétérogènes à un moment donné (Zhou et al 2007). L'approche a été évaluée sur des données synthétiques (Deux graphes connexes générés G_{XY}, G_{YZ}) selon différents paramètres (densité de liens, les proportions de X / Y / Z, nombres de cluster k) (Zhou et al 2007). Ensuite, l'évaluation a été effectuée sur des données réelles échantillonnée depuis 'CiteSeer (Un moteur de recherche publique et une bibliothèque numérique des documents scientifiques et universitaires).

Selon (Zhou et al 2007), le processus est montré efficace pour détecter des communautés 'temporelles' (basées sur des connaissances préalables) plus fiables par rapport à une communauté discrètement découverte dans un instantané unique. Ils ont étudié le changement de taille de 4 communautés pendant 6 périodes de temps (Zhou et al 2007). À une période donnée, la taille de la collectivité est normalisée par rapport à son groupement uniforme des années. Ce qui a été remarquable est que chaque communauté conserve un nombre minimal d'acteurs (d'auteurs) relativement stable. Ce sont des acteurs principaux qui persistent et forment les membres '***noyau'*** de la communauté et qui conservent son identité. Au cours des six périodes de temps, les auteurs observent aussi que les clusters découverts émergent avec des nouveaux mots (Zhou et al 2007). Par conséquent, ***la communauté change de taille et aussi d'intérêt dans le temps.***

3.2.4.2.3. Phénomène de persistance

La question sur persistance des groupes sociaux est fondamentale. C'est une propriété ou indice comportemental qui conserve l'identité et l'existence d'un groupe évoluant pendant une durée de temps et qui montre que le concept de communauté est plus vaste qu'un groupe. Donc, la recherche d'une communauté temporelle se penche sur la recherche d'un groupe persistant dans le temps. À cette fin, on a besoin de modèles plus expressifs où la composante de temps est plus explicite pour comprendre la persistance des groupes d'individus ayant un noyau stable et une périphérie qui change progressivement dans le temps.

(Berger-Wolf & Saia 2006) ont étudié le phénomène de persistance en proposant un modèle de graphe multiparti $G = (V_1, ..., V_T, E)$, orienté et pondéré. Il représente la dynamique des groupes d'un SN pendant une période d'observation (de 1 à T). D'abord, le réseau est observé chaque time-step sous forme d'une partition d'individus en groupes: $P_1, P_2,..., P_T$, tel qu'une partition P_i est un ensemble disjoint de groupes qui est en fait le sous-ensemble V_i (Berger-Wolf & Saia 2006). Si un groupe $g \in P_i$ alors son indice $P(g) = i$. $(g_i, g_j) \in E$, si $P(g_i) < P(g_j)$ et la similarité $SIM(g_i, g_j) \geq$ Beta (un seuil de similarité) (Berger-Wolf & Saia 2006). De ce

fait, w (g_i, g_j) = Sim ((g_i, g_j), (Berger-Wolf & Saia 2006). La similarité entre 2 groupes g et h est évaluée par une mesure de similarité d'ensemble (ex. indice de Jaccard).

(Berger-Wolf & Saia 2006) ont utilisé ce modèle pour représenter l'un des datasets référence classique en SNA : 'SOUTHERN CLUB WOMEN' [29,75] ((Breiger 1974)) ((Davis et al 1941)). Ces données ont été recueillies en 1933 à Natchez, une ville située au bord du fleuve Mississippi, par un groupe d'anthropologues. Ils ont effectué un suivi pendant une période de 9 mois sur 18 femmes participant à 14 événements sociaux informels (Des partis en jardin, jeux de cartes et autres) (Tantipathananandh et al 2007).
À partir d'une matrice d'incidence, ces évènements ont permis de distinguer intuitivement 2 communautés (Tantipathananandh et al 2007) : Groupes de femmes par événement social (14 groupes) qui ont défini les sommets de leur modèle (Berger-Wolf & Saia 2006) (**Figure 39**). Les groupes (composante connectées) sont classés par ordre chronologique et numérotées (**Figure 39**). Dans cette figure, seulement les liens de poids de similarité = 6 sont pris (Berger-Wolf & Saia 2006).

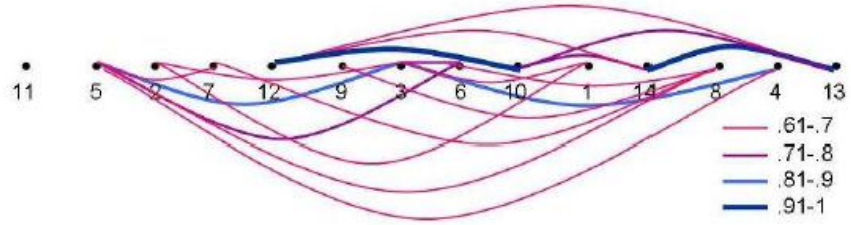


Figure 39. Le réseau de 'Southern Club Women' représenté par un graphe de méta-groupes avec un seuil de similarité Beta = 6 (Berger-Wolf & Saia 2006)

Ce graphe, appelé aussi Beta-graphe, est acyclique car toutes les arêtes sont orientées dans le temps vers le dernier time-step (Berger-Wolf & Saia 2006). Un chemin dans ce graphe est appelé un méta-group (MG) de longueur au moins Alpha(T) (Berger-Wolf & Saia 2006). Dans la **Figure 39**, un MG est par définition une séquence de groupe similaires MG = (g_1,….g_l) (Berger-Wolf & Saia 2006).

Un point de vue formel

Un individu x ∈ X qui appartient à un chemin MG, il appartient à un certain nombre de groupes g_1,..., g_l. Si le nombre d'occurrences de x dans MG est supérieur d'un seuil d'appartenance k a priori choisi (Berger-Wolf & Saia 2006), il sera considéré comme l'un des membres qui persistent dans le temps. Formellement la persistance a été définie en cherchant le plus long chemin (méta-groupe) MG. Autrement dit, le groupe le plus persistant (le plus stable) est celui qui correspond au MG qui maximise le nombre de groupes associés (Berger-Wolf & Saia 2006), qui maximise la moyenne : somme des poids des arêtes divisée par la longueur du chemin (Berger-Wolf & Saia 2006). C'est-à-dire le sous-ensemble qui apparait le plus dans les time-steps (dans au moins k groupes de MG). Il est vrai que les groupes changent progressivement et des sous-groupes persistent avec ce type de réseau, mais un lien qui relie par définition deux groupes similaires de 2 partitions différentes n'assure pas toujours une intersection non vide entre g_1 et g_l dans MG. Donc la persistance dépend ici des seuils choisis. Voici des exemples de trois de méta-groupes (le plus persistent/ stable et le plus grand) à partir la **Figure 39** :

MG = {11} est composé par un seul groupe, un événement pas très semblable aux autres. MG1 = (12, 10, 14, 13) est une séquence d'événements, de groupes plus semblables. Mais le

groupe le plus persistant correspond à MG2 = (5, 3, 6, 8) (Berger-Wolf & Saia 2006), ce qui montre que la notion de communauté et plus vaste qu'un groupe.

3.2.4.2.4. Problème d'une identification formelle des communautés dynamiques et sa complexité

Tout en se basant sur une séquence de sous-graphes (snapshots) d'interactions, l'ensemble des individus X peut varier dans le temps. Tant que certains individus n'apparaissent pas à tout moment, chaque P_t dans la séquence de (Berger-Wolf & Saia 2006) noté $H = \{P_1, P_2,\ldots, P_T\}$ dans (Tantipathananandh et al 2007) n'est pas une partition mais plutôt une collection non vide d'ensembles disjoints à l'instant t (Tantipathananandh et al 2007). Étant donné les affiliations des individus à chaque instant, le problème d'identification des communautés dynamiques a été bordé par (Tantipathananandh et al 2007) comme un problème d'optimisation combinatoire. Il s'agit d'un problème adapté à un problème de coloration dont la solution consiste globalement à minimiser une fonction objective basée sur des contraintes. Ces contraintes (numérotées dans le Tableau 23) sont pratiquement tirées depuis les propriétés des communautés, leur évolution (affiliation chronologique des individus), persistance, etc.

Tableau 23. Quelques propriétés des communautés dynamiques à formaliser

Propriétés	**Descriptions (contraintes dérivées)**
Distinction entre groupe et communauté	À chaque time-step, un groupe est un représentant une trace d'une communauté distincte (1) Un individu est souvent présent dans le groupe représentant la communauté à laquelle il est affilié (5)
Persistance	Un individu possède une tendance à ne pas changer sa communauté très fréquemment Il interagit fréquemment avec les membres de sa communauté: liens intracommunautaire (3)
Promiscuité	Sur le plan topologique chaque individu est affilié seulement à une communauté à un moment donné. Mais il peut changer son appartenance au fil du temps (2)
Oscillation	Un individu qui change son affiliation à plusieurs reprises, est une oscillation entre les communautés plutôt qu'une promiscuité (4)

Selon (Tantipathananandh et al 2007), la fonction objective est différente de celle de la coloration traditionnelle d'un graphe classique. La **Figure 40** montre que le modèle de graphe proposé contient deux types de sommets (Tantipathananandh et al 2007). $v_{i,\,t}$ un sommet individu (rond) qui représente l'individu i à l'un instant t et $v_{g,\,t}$ un sommet groupe (carré) qui représente un groupe de P_t.

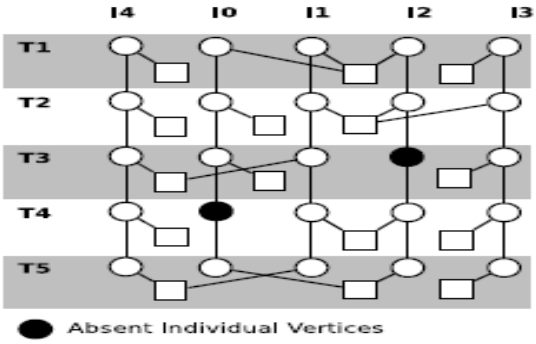


Figure 40. Modèle de graphe proposé dans ((Tantipathananandh et al 2007)) **en interprétant la détection des communautés dynamiques comme un problème de coloration de graphe. Les carrés sont des sommets groupes et les cercles sont des sommets individus.**

Une arête est créée de $v_{i,\,t}$ à $v_{i,\,t+1}$ si i continue à apparaitre à t+1. Une arête (d'appartenance) entre $v_{i,\,t}$ et $v_{g,\,t}$ si $i \in g$ à l'instant t tel que $t \leq T\text{-}1$ (Tantipathananandh et al 2007). Une coloration 'f' de ce graphe consiste à définir la couleur f ($v_{i,\,t}$) d'un sommet individu $v_{i,\,t}$ (et la couleur f ($v_{g,\,t}$)) qui représentera l'affiliation communautaire de i (la communauté que g représente) à un time-step t. Une coloration valide va intuitivement vérifie les deux contraintes 1 et 2 (Tableau 23). Si un sommet individu a une seule couleur c'est parce qu'il appartient exactement à une communauté à l'instant t. Un sommet groupe possède également une couleur unique car il représente une seule communauté (Tantipathananandh et al 2007). Cependant, la violation des contraintes de 3 à 5 a été pénalisée par des coûts (i-coût, g-coût, et c-coût) (Tantipathananandh et al 2007) :

Algorithme 2. Coût de violation des contraintes d'une interprétation de communautés dynamiques étant un problème de coloration

Pénaliser la violation des contraintes liées à certaines propriétés des communautés dynamiques

SI f (vi, t) ≠ f (vi, t+1) ALORS
- i-coût engage un facteur α

SI f (vi, t) = f (vg, t) & i ∉ g ALORS
- g-coût engage β1

SINON
- g-coût engage β2

SI i se présente à l'instant t mais pas dans son groupe ALORS
- g-coût implique β1 et β2

c-coût est attribué à chaque couleur d'un individu, différente de sa première couleur.

Le coût optimal de coloration d'un individu i avec une couleur $x \in S$ à un time-step t

$\Gamma(t,S,x) = G(t, x) + \min(\Gamma(t\text{-}1,R,y) + I(t, x, y) + C(x, R))$

- G (t, x) coût de coloration de i par la couleur x
- I (t, x, y) coût de la coloration de i entre t et t-1 par x et y
- C (x, R) coût d'utilisation de la couleur x par rapport R
- R est l'ensemble des couleurs utilisées dans les étapes précédentes de 1 à t-1.

Les paramètres $\alpha, \beta 1, \beta 2 \geq 0$. Donc, le problème d'optimisation consiste à trouver la coloration f (l'interprétation communautaire) qui minimise le coût total résultant (Tantipathananandh et al 2007). C'est une coloration d'un graphe qui représente en effet la dynamicité temporelle du SN. Par conséquent, elle permet de découvrir la séquence d'affiliation de chaque individu i $Ai = \{f(vi, 1), \ldots, f(vi, t)\}$. Partant du problème de coloration qui est à l'origine un problème NP-complet, il a été démontré (avec des théorèmes) que la résolution d'une interprétation communautaire dans un SN dynamique est également NP-complet (Tantipathananandh et al 2007). Le problème de décision correspondant a été défini comme suivant : Y a-t-il une interprétation communautaire sur (X, H) de coût total d'au plus B?

Théorème *: Le problème d'interprétation communautaire est NP-complet et APX dur. Autrement dit, il y a une constante ϵ, aucun algorithme en temps polynomial ne peut obtenir une approximation garantie mieux que $(1 + \epsilon)$ pour une cette interprétation de Communautés* (Tantipathananandh et al 2007).

Ce théorème a été prouvé en rapprochant le problème (dans les 2 sens) à un problème 'Minimum Multiway Cut problem' (Tantipathananandh et al 2007). Par conséquent, les chercheurs (Tantipathananandh et al 2007) ont pensé à résoudre le problème en le

décomposant en deux sous-problèmes dont la solution est une combinaison d'heuristiques gloutonnes approximatives et la programmation dynamique. Il s'agit d'abord de procéder par un algorithme A (fg) menant à trouver une coloration optimale des sommets groupes. C'une recherche exhaustive (sur f*g) sur toutes les assignations valides de couleur à ces groupes (Tantipathananandh et al 2007) qui est accélérée par la technique 'Branch & Bound', puisque le temps d'exécution pour trouver une coloration optimale de groupes n'est pas polynomial (Tantipathananandh et al 2007). Pour les grandes instances, les auteurs préfèrent d'utiliser des heuristiques à la place de la recherche exhaustive comme 'le matching biparti' ou des heuristiques gloutonnes (greedy) qui sont plus efficaces (Tantipathananandh et al 2007): 'Backward Greedy algorithm' ou 'Least Delay Greedy' (Tantipathananandh et al 2007).

Ensuite, la coloration des sommets individus a été décomposée en sous-étapes :

***Lemme**: Etant donné une coloration des sommets groupes, le coût minimal de coloration des sommets individus consiste en coûts minimaux de coloration de chaque individu i, indépendamment des autres* (Tantipathananandh et al 2007).

Le coût optimal de coloration d'un individu i est le minimum de ses coûts de coloration à chaque time-step. Le coût optimal de coloration de i avec une couleur x à un time-step t est calculé par une procédure récursive (Algorithme 2). Ce cadre formel a été d'abord évalué sur 2 ensembles de données synthétiques 'Assembly Line' et 'Dutiful Children' avec des scénarios dynamiques différents (Tantipathananandh et al 2007). Après ils l'ont évalué sur des données sociales de référence 'Southern Women' [29,75] ((Breiger 1974)) ((Davis et al 1941)) et 'Grevy's zebras' ((Sundaresan et al 2007)). Sur ce dernier dataset, les résultats ont montré que la structure communautaire trouvée en s'appuyant une heuristique gloutonne ne s'accorde pas avec la configuration manuellement identifiée par les biologistes (Reda et al 2009). Cependant cette interprétation dynamique aide à observer des phénomènes intéressants (changement d'appartenance d'un individu au cours du temps) qui ont été masqués dans un graphe statique (Tantipathananandh et al 2007).

3.2.4.2.5. Evolution spatio-temporelle des groupes

Comme beaucoup de données sociales sont spatialement référencés, on a déjà noté que l'environnement a une influence aussi sur l'évolution des structures sociales, la formation des groupes et leur persistance. Dans ce sens, (Reda et al 2009) avaient tendance de combiner l'espace (l'environnement) avec le temps pour étudier la dynamique de ses groupes dans le temps et obtenir un aperçu plus grand sur leur évolution. Ils ont étudié les données de 'Grevy's zebras' ((Sundaresan et al 2007)), une structure sociale animale d'un troupeau de zèbres sauvages de Grévy en Kenya. Suivant des boucles de recensements approximatifs à deux fois par semaine (Tantipathananandh et al 2007), les circulations de ces animaux sont repérées par GPS (Reda et al 2009) (Tantipathananandh et al 2007). Les liens sociaux entre eux ont été enregistrés comme étant la proximité physique au sein du troupeau pendant une période d'observation de 3 mois en 2002 (Tantipathananandh et al 2007). Cette étude était supervisée par des experts écologistes intéressés par la recherche dans le comportement social des zèbres de Grévy (Reda et al 2009).

Le SN est composé par 28 individus en interactions, pendant 44 time-steps (Tantipathananandh et al 2007) ((Sundaresan et al 2007)). Ces animaux forment des petits

groupes (serrés) pendant des périodes courtes (Reda et al 2009). Les auteurs ont fournis une visualisation plus intuitive de l'évolution spatiotemporelle de ces groupes en utilisant l'outil Socio Scape. Grâce à une carte topographique en 3D, les auteurs ont montré l'évolution des groupes (en état de reproduction des individus), influencée par leur environnement (les besoins de ressources) (Reda et al 2009). En effet, les individus choisissent des associations qui maximisent l'accès aux ressources : l'eau et de l'herbe (Reda et al 2009). Ils représentent de manière plus détaillée les mouvements et les associations (affiliation chronologique) des individus (males/ femelles) dans l'espace et le temps.

3.2.4.2.6. Autres approches alternatives

((Chakrabarti et al 2006)) sont parmi les premiers qui ont voulu éviter que la division de réseau à un instant donné soit trop divergente des divisions antérieures comme c'est le cas dans la plupart des travaux précédent (Tantipathananandh et al 2007) (Zhou et al 2007) (Berger-Wolf & Saia 2006). Dans ce sens, des approches alternatives comme le clustering évolutif pour des réseaux dynamiques peuvent servir aussi de cadres formels pour l'identification des communautés dynamiques. ((Chakrabarti et al 2006)) évoquent le problème de 'Clustering évolutif': Au lieu d'extraire dans un premier temps des 'communautés' dans chaque snapshot et puis trouver des liens entre ces communautés de différents snapshots, ils ont étudié la structure communautaire et son évolution au même temps. Ils se basent sur la qualité de snapshot: Comment le clustering C_t représente les données à t) et le coût de de l'historique: Comment le clustering C_t est différent de C_{t-1}, et adaptent un clustering hiérarchique agglomératif et 'k-means' à cette proposition ((Chakrabarti et al 2006)). Un clustering de graphe évoluant dans le temps a été prouvé aussi par ((Sun et al 2007)) comme un problème NP-difficile qui avait besoin d'une heuristique gloutonne appelée GraphScope ((Sun et al 2007)). En outre, ((Chi et al 2007)) ont étendu un clustering spectral sur une configuration dynamique de réseau en proposant deux cadres 'preserving cluster quality' (PCQ) et 'preserving cluster membership' (PCM) qui supportent la variation du nombre des nœuds. D'autre part, ((Lin et al 2008)) ont utilisé un modèle d'appartenance probabiliste pour proposer FacetNet 'a framework for analyzing communities and their evolutions in dynamic networks.' Avec un modèle probabiliste, ils étalaient capables d'attribuer un individu à plusieurs communautés selon un poids qui indique le degré d'appartenance ((Lin et al 2008)).

3.2.4.3.Bilan sur la dynamique des communautés

La dynamique temporelle ajoute plus de problématiques et de complexité à la découverte des structures communautaires dans des SNs dynamiques. Cette complexité s'explose le faite d'étendre les concepts théoriques (composante) liés au concept de communauté sur des formalismes de TVG (chemin temporel) pour définir une communauté temporelle. En outre, la majorité des techniques de partitionnement/clustering de graphe sont développées dans un cadre statique. C'est l'une des raisons principales pour lesquelles la découverte des communautés temporelles s'appuie souvent sur une séquence de snapshots du SN comme étant le squelette des modèles proposés. D'autre part, des auteurs affirment qu'une découverte de communautés basée sur la discrétisation du graphe social dynamique peut radicalement déformer la structure communautaire réelle et son évolution (Tantipathananandh et al 2007). Traiter chaque cliché de réseau indépendamment peut entraîner parfois des fluctuations indésirables au niveau des adhésions de la communauté d'un time-step à l'autre (Parthasarathy et al 2011). De ce fait, des cadres formels proposent de modéliser des liens entre les

collectivités d'un snapshot à l'autre pour pouvoir capter les changements temporels: Partitionnement conditionné par l'historique d'appartenance (Zhou et al 2007), liens de ressemblance entre groupes d'un MG (Berger-Wolf & Saia 2006) D'autres sont basés sur le clustering évolutif ou des modèles probabilistes, etc. ***Ce sont des représentations là où on peut appliquer ou adapter des techniques et heuristiques de Graph Mining*** (Comme la coloration d'un graphe dans (Tantipathananandh et al 2007)). C'est une bonne manière pour comprendre des indices comportementaux (masqués dans le cas statique) comme la persistance, l'influence, la diffusion, etc., et prédire le comportement future d'une communauté (ex. collaboration entre les groupes) dans des réseaux du monde réel qui évoluent progressivement ((Asur et al 2007)). ***Ce type d'approches a montré aussi que le concept de communauté et de groupe se distinguent***. Par exemple, un MG dans le modèle de (Berger-Wolf & Saia 2006) ou les contraintes de (Tantipathananandh et al 2007) ont montré que la notion de communauté est plus vaste qu'un groupe.
La dynamique des communautés/groupes est un problème difficile à formaliser et même à tester notamment sur de grands SNs évoluant pendant des longues périodes d'observations. Cependant, elle ouvre la voie pour comprendre profondément la dynamique du SN et répondre à des questions comme ***la fragilité du SN*** dans le temps. C'est une question importante dans un contexte épidémiologique où on doit trouver les acteurs/structures influents sur l'évolution (pour les mettre en quarantaine) afin de limiter la propagation des maladies contagieuses. Formellement, selon le modèle de (Berger-Wolf & Saia 2006) la fragilité du SN a été liée à la recherche du plus petit ensemble de groupes (critiques) dont la suppression ne laisse pas des MG, autrement dit ne laisse aucun groupe évolue dans le temps (Berger-Wolf & Saia 2006). C'est un problème NP-Complet, qui permet de revenir au problème de 'Min k-Path Vertex Shattering Set': On cherche le plus petit sous-ensemble de sommets $U \subseteq V$ tel que le sous-graphe induit par $V \setminus U$ n'a pas de chemin plus long que k-1. (Berger-Wolf & Saia 2006).
Au-delà de l'évolution topologique des structures communautaires, la dynamique d'une communauté évoque aussi l'évolution de ses intérêts.

3.2.5. Conclusion partielle

Les SNs/OSNs évoluant dans le temps stimulent le développement des modèles, métriques et techniques qui supportent de plus en plus la composante temporelle et qui peuvent servir à analyser les autres réseaux dynamiques : Réseaux de capteurs, routage Internet, des modèles de mobilité, etc., même en dehors de l'informatique : En sociologie, épidémiologie (ex. l'étude des épidémies qui se propagent dans les réseaux mobiles et sociaux), (Tang et al 2010a).
(Tang et al 2010a) affirment que les interactions des utilisateurs sur les OSNs ont des caractéristiques et une efficacité de diffusion différentes et même plus lente par rapport des contacts humains (Tang et al 2010a) (ex. Dans réseaux mobiles). Dans un OSN, les interactions se produisent instantanément car les relations en ligne ne nécessitent pas que les utilisateurs se connaissent pendant des longues périodes de temps contrairement aux contacts humains. Cependant les utilisateurs peuvent ne pas répondre ou commenter immédiatement par exemple sur ce qui est posté (photos) sur un mur. Donc un retard naturel est introduit entre les interactions (Tang et al 2010a).
En général, les études sur les SNs dynamiques sont communément affectées par la résolution des fenêtres de temps.

La résolution des fenêtres de temps

Souvent, la durée d'observation d'un SN est décomposée en fenêtres de temps (une décomposition ponctuelle). L'étude se focalise à chaque fois sur un état donné (un snapshot) du SN. C'est les nœuds et les liens existant dans une fenêtre de temps ayant une taille bien

définie. Ces fenêtres de temps sont des observations réduites et discrètes d'un processus de développement en temps continu ((Snijders 2005)). En effet, définir la résolution de ces fenêtres est une phase importante pour comprendre et saisir la dynamique temporelle du SN dans un premier plan. C'est le schéma selon lequel la dimension temporelle est structurée (Kazienko et al 2011). Les fenêtres qui sont assez larges influencent la précision/ l'exactitude des mesures et l'identité et la qualité des communautés (Tang et al 2010a). D'abord, les informations sur les dépendances temporelles et l'ordonnancement des liens dans le temps (Tang et al 2010a) (Kazienko et al 2011) sont de plus en plus perdus. L'ordre d'apparition des liens est de moins en moins respecté. Ce qui risque de surestimer la connectivité surestimée et sous-estimer des chemins et les géodésiques, car la représentation tend à devenir plus statique. D'autre part, si les fenêtres de temps sont étroites, un bruit sera relativement introduit dans ces métriques (Kazienko et al 2011). En résumé, le développement d'un modèle de réaliste d'un SN dynamique nécessite une résolution optimale pour obtenir des interprétations significatives sur des phénomènes plus profonds.

Il vrai que les acteurs sont susceptibles d'apporter des modifications qui peuvent passer inaperçues entre 2 observations consécutives (snapshots). C'est pour cette raison les métriques standard de SNA appliquée sur une séquence de snapshots ont été montrés moins efficaces que leurs extensions temporels sur des formalismes de TVG. Cependant, un modèle élaboré sur une observation quasi-continue semble infaisable. C'est pour cette raison aussi que l'étude du comportement dynamique des communautés est beaucoup plus concentrée sur ces séquences de snapshots tout en cherchant à minimiser la discrétisation dans des modèles améliorés.
La modélisation et l'analyse des structures dynamiques des SNs, métriques temporelles et dynamique des communautés ne sont pas les seuls sujets d'actualité dans la dynamique temporelle. Il y a des tendances comme la prédiction des liens futurs et ***des questions posées sur des phénomènes plus profonds***, etc.

3.3.Méthodes de visualisation des SNs dynamique et softwares

La visualisation ou l'analyse visuelle des SNs est l'un des objectifs de SNA (Kang et al 2007). L'une des tâches spécifiques de la visualisation est de mettre en évidence les informations pertinentes (Kang et al 2007). Mais devant la progression de leur taille il est difficile maintenant de tirer des sociogrammes lisibles et précis ((Scott 2011)). En plus, la dynamique temporelle du SN rend sa visualisation plus compliquée. Mais, c'est un outil puissant pour analyser l'évolution des SNs, la dynamique des communautés, etc. (Ahn et al 2011). Le défi est comment analyser visuellement la dynamique du SN au fil du temps (Ahn et al 2011). Nombreuses méthodes proposées ont été couplées avec la visualisation, même sur plusieurs logiciels commerciaux et autres applications.

Pourquoi visualiser :

En utilisant des représentations graphiques, statistiques et la navigation temporelle (Ahn et al 2011), une méthode, un prototype ou un système donne l'avantage d'apercevoir l'évolution et comparer interactivement des réseaux au fil du temps. Avec la visualisation les gestionnaires 'Community manger, 'SN manager', etc. ont une meilleure compréhension de la dynamique des SNs. Ils ont l'opportunité de découvrir les éléments déclencheurs et les phases de croissance, décroissance, stabilité, utilisation active ou des activités malveillantes, etc. D'où, ils peuvent élaborer des stratégies pour promouvoir la croissance et prévenir la décroissance et l'interférence destructive (Ahn et al 2011).

3.3.1. Approches utilisées pour visualiser la dynamique temporelle

Intuitivement, si les SNs sont représentés par des matrices ou graphes sur des plans à deux dimensions, l'ajout d'une dimension temporelle est susceptible de produire des représentations tridimensionnelles (Dekker 2011). Deux approches sont utilisées pour visualiser la dynamicité temporelle d'un SN.

3.3.1.1.Visualisation statistique descriptive

La visualisation statistique est utile pour afficher l'évolution du SN (des courbes) au fil du temps (Ahn et al 2011). Dans nombreux systèmes (ex. les appels de téléphones cellulaires) les méthodes statistiques sont avantageuses pour détecter de la croissance et la diminution de certaines caractéristiques (Ahn et al 2011). L'étude de (Dekker 2011) sur l'évolution des discussions au sein des ateliers de planification gouvernementale, illustre des techniques bidimensionnelles de visualisation statistiques des SNs dynamiques. Souvent ces techniques produisent des diagrammes à 2 dimensions (Dekker 2011) mais elles se distinguent selon l'axe de temps, soit il est explicite ou non.

Diagrammes avec axe de temps explicite

Le temps se présente dans le diagramme sur un axe horizontal explicite. Sur l'axe vertical, il se trouve des propriétés numériques (degré, centralité, etc.) des acteurs, des liens comme c'est illustré dans la Figure 35 et Figure 36 ou de l'ensemble du réseau comme c'est le cas dans Figure 30. Ce type de diagramme qui affiche par exemple l'évolution des scores de centralités (Figure 35) est facile à comprendre et permet d'identifier les principaux participants à divers points de temps. Une autre représentation alternative en 3 dimensions appelée 'temporal social surface' affiche à chaque point de temps, une liste ordonnée de centralités des nœuds. Elle montre plus clairement l'évolution des nœuds les plus centraux sans les identifier (Dekker 2011).

Le ***diagramme à barres empilées*** a été utilisé par (Dekker 2011) pour représenter le degré des participants les plus actifs dans le temps. Le degré de chaque participant à un moment donné se présente par une barre colorée. Cependant cette visualisation ne donne pas des indications sur la structure du réseau (Dekker 2011). Ce diagramme ne montre pas avec qui le nœud interagit (vers qui ex. ses messages sont envoyés vers qui?). D'où, des auteurs ont pensé à remplacer les barres colorées par des barres avec flèches 'Time Axis with Arrow counts' mais l'idée semble infaisable dans les réseaux dense (Dekker 2011). Une autre alternative consiste à représenter toutes les interactions (des flèches entre les nœuds) tout au long de l'axe du temps, ce qui est connu en industrie logicielle comme un ***diagramme de séquence***. Le diagramme de séquence est assez explicite pour présenter l'évolution temporelle des petits réseaux, mais il risque de devenir encombré (Dekker 2011).

3.3.1.2.Visualisation graphique

Généralement une approche ponctuelle représente le SN par des snapshots séparées à chaque point dans le temps (Ahn et al 2011) (Dekker 2011). C'est en quelques sortes une série temporelle de réseaux qui peut être convertie en une animation ou une vidéo. Les animations affichent concrètement sur un écran ordinateur l'évolution dynamique du SN avec la possibilité de faire l'avance rapide ou des pauses (Dekker 2011), ex. l'outil Commetrix[106]. Mais elles ne peuvent pas être imprimées et ne sont pas les plus adaptées à l'analyse.

Un encodage qui remplace l'axe de temps

Il est vrai que les graphes (à deux dimensions) continuent d'être clairement la principale approche pour visualiser les SNs. Un encodage peut être appliqué pour indiquer par exemple

les nouveaux nœuds arrivants avec des triangles et les sortants avec des cercles ou un encodage de taille qui peut encoder aussi le degré des nœuds (Ahn et al 2011).
Par ailleurs, il y a des méthodes comme 'Temporal Coloring Arrows' qui permet d'encoder l'information temporelle dans les attributs spécifiques du graphe, les attributs des nœuds ou des liens (Dekker 2011), sans avoir besoin d'un axe de temps explicite.

(Dekker 2011) utilisent aussi la couleur d'un arc pour indiquer à quelle l'heure l'interaction a eu lieu (Dekker 2011): Le temps d'activité. La saturation indique la progression dans le temps. La couleur et la taille d'un nœud distinguent respectivement un participant et sa centralité. Mais le schéma résultant semble plus difficile à interpréter et moins utile (Dekker 2011) par rapport les modèles précédents. Une autre technique appelée 'la corrélation temporelle des acteurs' utilise aussi un encodage d''information temporelle par position (Dekker 2011). Elle se base sur le calcul de similarité temporelle entre les acteurs, entre leur historique d'activation (Dekker 2011). Chacun est associé à un historique d'activation à partir de ses messages envoyés. Le processus vise à une mise en échelle multidimensionnel pour afficher la similarité entre les activités temporelles des acteurs (Dekker 2011). Il regroupe ceux qui sont actifs en même temps, ce qui donne une représentation plus abstraite de l'évolution temporelle, difficile à interpréter par les non-spécialistes.

3.3.1.3.Bilan

Pour voir ce qui se passe réellement derrière les phases d'évolution du SN, les visualisations descriptives comme courbes d'évolution de centralités sont plus intéressantes. En général, la topologie des graphes simples ne saisit pas bien les structures dynamiques comme les groupes influencés par les changements d'interactions dans les SNs. Mais quand même, le temps serait mieux représenté explicitement par rapport un autre type d'encodage. Quand la dimension temporelle est plus explicite, elle entraine une amélioration. C'est l'image de l'exemple donnée par la **Figure 37** qui montre le déroulement d'une affiliation chronologique des individus à des communautés dans le temps. (Ahn et al 2011) donnent des principes pour mettre en œuvre une visualisation interactive de l'évolution temporelle SNs et qui aide à percevoir les changements au fil du tps.

- Il y a des parties statiques du graphe qui restent figées en évitant les distractions.
- Les changements (des nœuds, des liens et leurs attributs) doivent être visuellement manifestés et détectable pour des comparaisons plus faciles explorables et aussi interactives.
- Les changements temporels d'un sous-graphe (communauté/ groupe/ sous-groupe) et ses attributs doivent être également détectables.

Les auteurs distinguent 3 états de changement de ces composantes: ***addition (ajout), enlèvement (suppression), vieillissement*** (Ahn et al 2011).

3.3.2. Application, outils et softwares (avancés) visuels-analytiques de la dynamique des SNs

Parmi les outils et softwares de SNA cités précédemment, il y a ceux qui supportent la dynamique temporelle des SNs, représentent ou analysent des données sous des formats puissants : .net, .gml, cmx, etc. Par exemple, Gephi ((Bastian et al 2009)), Pajek ((Batagelj & Mrvar 2012)) ((Batagelj & Mrvar 2003a)) ((Batagelj & Mrvar 1998)) ((Beauguitte 2011)), NetworkX ((Hagberg et al 2008)), NetVis, Commetrix, etc. La majorité qui est utilisée pour visualiser les SNs (notamment dynamiques), affiche l'ensemble du réseau (Kang et al 2007). Ils sont axés sur l'ajout et la suppression des nœuds et des arêtes (Kang et al 2007). Pour certains auteurs, c'est une visualisation limitée mais associée à des analyses statistiques sophistiquées. Pour montrer et exploiter les avantages de leurs approches et idées proposées,

les chercheurs veulent les implémenter dans des prototypes et des systèmes plus flexibles, faciles à utiliser et hautement configurables (Ahn et al 2011). Cela facilitera l'exploration de manière interactive les changements temporels des SNs (Ahn et al 2011).

(Ahn et al 2011) ont présenté '***Spreadsheet-based approach***' en utilisant ***NodeXL*** ((Hansen et al 2010)) (Une extension de Microsoft Excel). à travers une barre glissante du temps il est possible de calculer la centralité de chaque nœud dans le temps (Ahn et al 2011).
Dans la **Figure 41**, on trouve un aperçu sur '***TempoVis***', un prototype plus autonome qui a un avantage plus axé sur le temps (***Adding Time-based Interactive Exploration***) (Ahn et al 2011). Il est équipé d'une barre glissante permettant aux utilisateurs de naviguer dans le temps. La taille du nœud est proportionnelle aux statistiques du graphe qui sont calculées dynamiquement dans le temps (Ahn et al 2011). En déplaçant le curseur vers la gauche ou la droite (**Figure 41**), l'utilisateur peut distinguer les nouvelles conversations (nouveaux liens) d'un mois donné (Juin 2010) et les plus anciennes qui se représentent par des couleurs à une faible intensité (vieillissement) (Ahn et al 2011). Il peut sélectionner des parties (groupes) de graphe en indiquant les liens concernés (sélectionnés) en bleu. Par exemple, le grand cluster situé au centre (d'une couleur à faible intensité) se réfère à un groupe qui s'est formé plus tôt que l'autre cluster situé en haut de la **Figure 41** qui est plus récent, ayant une couleur plus intense.

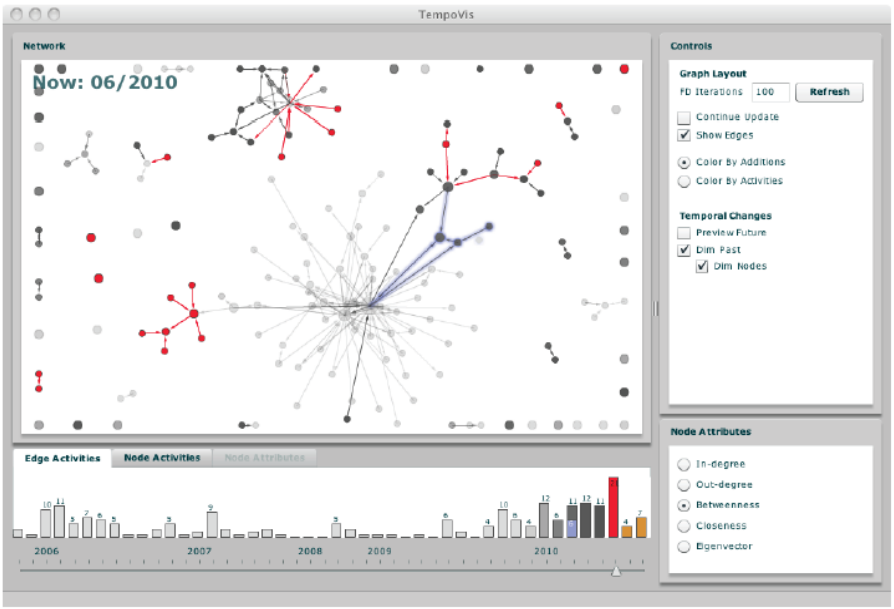


Figure 41. TempoVis un prototype pour visualiser la dynamique temporelle d'un SN ((Ahn et al 2011))

La visualisation des SNs dynamiques évoquent des suggestions sur comment visualiser la dynamique de ses groupes. Par exemple, il y a la visualisation de l'appartenance à un groupe pour un seul acteur. C'est une vue égocentrique du réseau, centrée autour d'un seul individu (Kang et al 2007). Toutefois, on (Kang et al 2007) qui ont proposé l'outil '***C-GROUP***' qui permet d'analyser visuellement l'évolution des appartenances aux groupes dans le temps à partir d'une paire d'individus. Cela montrera si les 2 acteurs (pair d'acteurs) se trouvent dans des groupes similaires ou différents, et comment la structure des groupes partagés change au fil du temps (Kang et al 2007). C-GROUP est inspiré de '***D-Dupe'***, un outil analytique visuel qui se concentre sur les acteurs qui sont très semblables (Kang et al 2007). Il présente deux aspects de visualisation. Il visualise le changement dans les groupes partagés et les comportements du groupe en utilisant l'animation pour signaler les parties pertinentes de ce changement (Kang et al 2007).

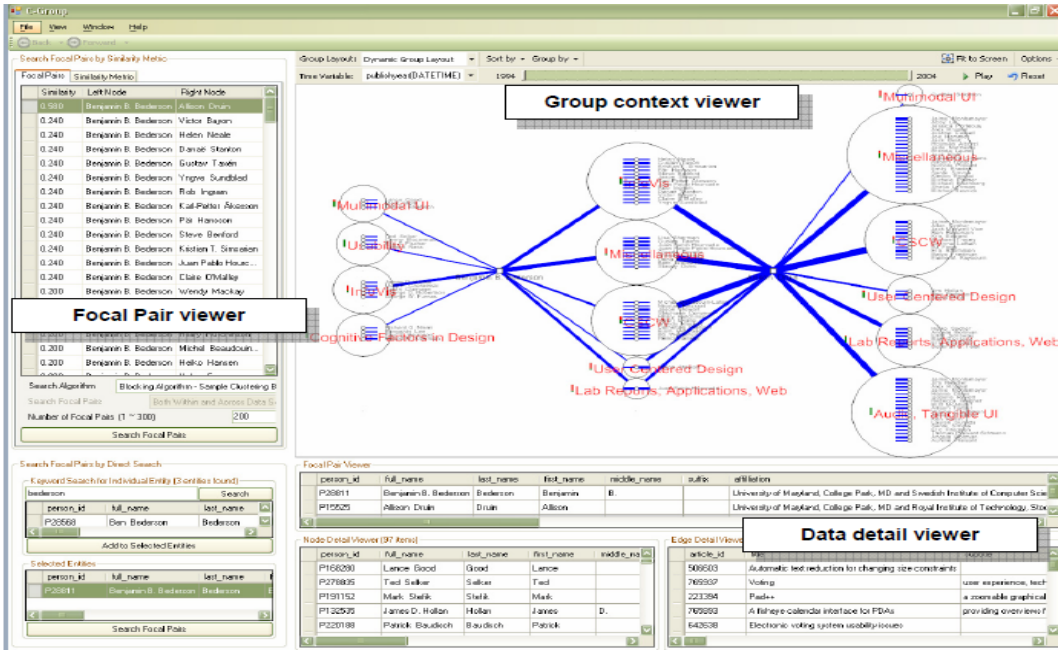


Figure 42. L'interface de C-GROUP, un outil pour analyser visuellement l'évolution et les appartenances aux groupes à partir d'une parie d'acteurs sélectionnés ((Kang et al 2007))

La **Figure 42** montre un cas d'étude de C-GROUP appliqué sur un dataset qui est un réseau de coauteurs extrait à partir la bibliothèque numérique ACM: 4.073 documents de conférence ACM CHI entre 1982 et 2004 rédigés par 6.358 auteurs qui sont connectés avec 12.727 relations coauteurs. L'interface se compose de trois fenêtres :

- 'Focal pair viewer' (**Figure 42**): Selon l'objectif analytique de l'utilisateur, il est possible de sélectionner les pairs (A1, A2) directement ou selon des critères : par similarité, ou autre combinaison entre les deux.
- 'Group context viewer' (**Figure 42**): fournit deux visualisations de l'appartenance de deux auteurs ('Ben Bederson' et 'Allison Druin') aux groupes. il y a 3 régions. Région des groupes partagés et les 2 autres régions des groupes non partagés de A1 et A2 qui sont à coté (Kang et al 2007). L'outil est basé sur ***un mécanisme souple (une certaine sémantique) de regroupement*** par attributs auteurs, par attributs de papier (évènement) ou par attributs de participation. qu'on peut choisir par 'Group by' dans la barre d'outils (**Figure 42**). Les groupes ne sont pas forcément mutuellement exclusifs. Ils se chevauchent. Il est possible de naviguer dans de temps (contrôler la granularité du temps) en constatant l'évolution de ces groupes.
- 'Data detail viewer' (**Figure 42**): montre les valeurs d'attributs des auteurs et des documents sélectionnés, affichés dans les groupes obtenus (Kang et al 2007), avec d'autres éléments statistiques sur la structure réseau.

Une autre conclusion est que C-GROUP permet ***d'explorer visuellement l'évolution de la dynamique des SNs hétérogènes*** (auteurs, articles, conférences). Alors que la conception des outils et des tâches analytiques visuelles spécifiques aux SNs hétérogènes dynamiques est difficile, C-GROUP est une étape vers une analyse visuelle plus orientée (Kang et al 2007).

TeCFlow 'A Temporal Communication Flow Visualizer for Social Network Analysis**'** est un autre exemple d'application, proposé par (Gloor & Zhao 2004). C'est un visualiseur de flux de communication temporelle. Il fusionne la visualisation animée avec l'analyse temporelle des SNs (les réseaux de courrier électronique) (Gloor & Zhao 2004). Il importe et traite d'abord les archives (les journaux : fichiers logs) de communications (emails, d'appels, de messagerie instantanées, transcriptions sur blogs, etc.) et les stockent dans une Base de données (BDD) SQL. La BDD est interrogée pour sélectionner les messages envoyés ou reçus

au sein d'une équipe (groupe) donnée, dans une période de temps donnée (From, to, Timestamp, etc.). Par conséquent, il génère automatiquement des visualisations statiques et des films interactifs de ces flux de communication sélectionnés sur un navigateur visuel en utilisant netgraph ((Varghese & Allen 1993)). À travers une barre glissante, l'utilisateur peut naviguer sur l'intervalle de temps sélectionné (une trame de n jours) de manière flexible (Gloor & Zhao 2004), et observer les nouveaux liens (des messages échangés) ajoutés au graphe. Tandis que les anciens liens en se déplaçant en dehors de cette trame de temps, sont grisés (vieillissement). Plus il y a d'interactions entre deux acteurs, plus ils sont proches et plus connectés et donc de plus en plus positionnés au centre du graphe (Gloor & Zhao 2004).

La visualisation est accompagnée par certaines évaluations (métriques de SNA comme 'Betweenness centrality' d'un groupe (GBC), sa densité, etc., dans le temps (Gloor & Zhao 2004), ainsi que l'indice de contribution (comme dans la Figure 9) de chaque participant (Gloor & Zhao 2004). Ainsi, on distingue les individus les plus actifs, le coordinateur qui envoie plus de messages (en recevant moins) et le chef d'équipe ou de communautés qui reçoit beaucoup plus de messages (en envoyant moins). Les auteurs ont découvert aussi des phases de collaboration (Des pics : des périodes d'activité de communauté) liées à des événements intéressants dans la durée de vie d'une équipe virtuelle (Gloor & Zhao 2004). D'où l'outil donne ***un éclairage précieux sur la dynamique organisationnelle*** (Gloor & Zhao 2004). C'est un moyen pour ***détecter l'émergence, le potentiel, centre / périphérie des clusters denses (des équipes de collaboration)*** (Gloor & Zhao 2004). À travers cet outil, (Gloor & Zhao 2004) visent ***une compréhension plus avancée de l'évolution des groupes en ligne et de développement d'une théorie de rôles joués par les membres des communautés virtuelles***.

On ajoute aussi l'outil '***SocioScape***' conçu par (Reda et al 2009) pour visualiser l'évolution spatiotemporelle des groupes. Les auteurs ont utilisé ***SocioScape*** pour visualiser les mouvements des zèbres de Grévy et le leurs structures communautaires dans l'espace et le temps (Reda et al 2009). Donc, ce type d'outils ouvre un autre paradigme de visualisation qui tend à combiner l'environnement avec la dimension temporelle. Il offre un aperçu plus intuitif sur le rôle de l'environnement dans la formation des structures sociales (des groupes) dynamiques dans le temps (Reda et al 2009).
Il y a aussi des applications moins connues comme ***Condor*** qui crée des films dynamiques de l'évolution des SNs à partir de nombreux types d'archives de communication : emails, forums, journaux d'appels, chat, les blogs), (Berger-Wolf & Saia 2006) ou ***Coolhunting*** qui permet de prévoir les tendances (trendsetters) des films (Prédictions Oscar), Tendances scientifiques ((Gloor 2007)) à partir leurs SNs correspondants.

3.4.Analyse multidimensionnelle des SNs dynamiques (Cas d'étude)

(Kazienko et al 2011) affirment que la complexité de calcul des SNs avec leur dynamique temporelle (et la richesse sémantique), qui ne cesse pas d'augmenter, remet en question les techniques/métriques de SNA traditionnelles et même leurs extensions. ***On analyse le SN sous différents angles (temps, groupes, sémantique, etc.)***. Les auteurs pensent à une analyse multidimensionnelle pour mettre SNA au courant de développement rapide des données sociales sur le web en taille et en richesse (Kazienko et al 2011). Ils considèrent que tous les SNs dynamiques peuvent être vus comme des structures multidimensionnelles (Kazienko et al 2011). Mais il y a peu de travaux de recherche abordent la multi-dimensionnalité des SNs dynamiques (Kazienko et al 2011). (Kazienko et al 2011) mettent l'accent sur ***le temps, le type de relations et la structure communautaire qui constituent en effet, les dimensions multiples ou le modèle multidimensionnel générique d'un SN***.

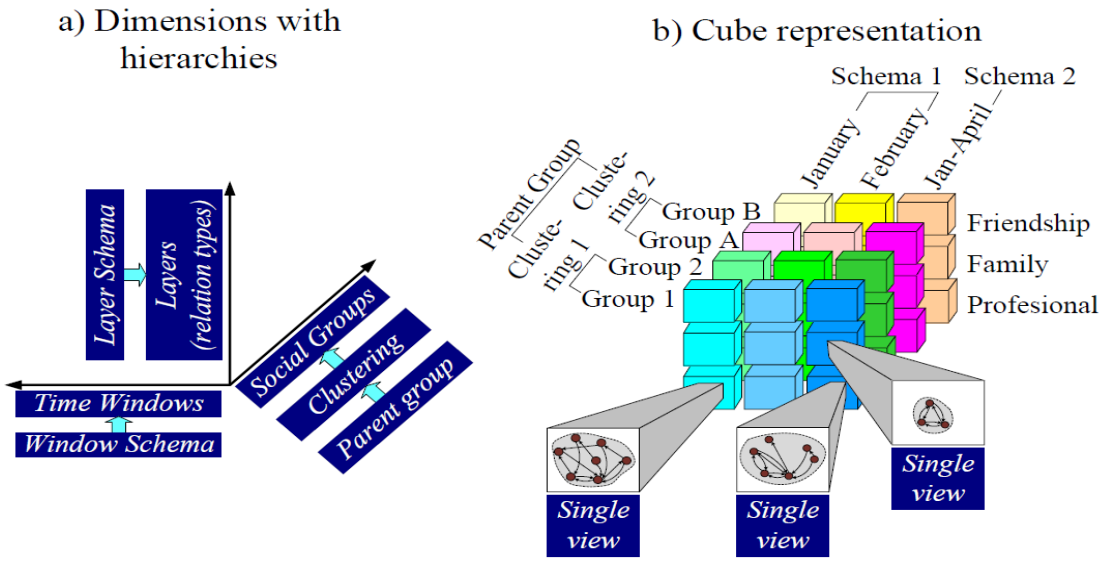


Figure 43. Un exemple de modèle multidimensionnel (tridimensionnel) d'un SN dynamique ((Kazienko et al 2011))

Dans la **Figure 43** (a), on trouve d'abord les 3 dimensions principales structurée selon des schémas hiérarchiques. Par conséquent le SN se présente par un cube **Figure 43** (b). L'évolution du SN dans le temps est captée par une série de snapshots. D'où, la résolution des fenêtres de temps constitue le schéma de la dimension temps. Dans un système donné un SN évolue maintenant avec différents types de relations {R1, R2,…, Rn} qui sont des canaux de communication. Même s'il y a des approches pour représenter les relations multiples, les auteurs (Kazienko et al 2011), définit un modèle de SN appelé multicouches L = {l_1, l_2,…l_n}. ***Ce sont des couches sémantiques, chacune représente un type de relation*** (Kazienko et al 2011). Le schéma de la dimension couche peut être par exemple Schéma A: layer1-l'amitié, layer2 - affaires, layer3 – Famille (**Figure 43** (b)) ou Schéma B: layer1 - emails, layer2 - appels téléphoniques (Kazienko et al 2011). La troisième dimension représente la configuration du SN en groupe (**Figure 43** (a)). Elle est structurée également selon un schéma qui décrit dans un deuxième niveau le type d'algorithme ou technique (de clustering) ayant produit le partitionnement du réseau. Dans le troisième niveau un objet virtuel racine est proposé pour conserver les informations sur les relations intergroupes.

Une vue dans ce modèle (**Figure 43** (b)) est une intersection des 3 dimensions, définie par un sous-réseau appartenant à une couche donnée dans une fenêtre de temps précise (Kazienko et al 2011). ***C'est un sous-ensemble de nœuds formant un group, interconnectés par le même type de relation dans la même période.*** Ainsi, cette vue dans ce modèle tridimensionnel d'un SN dynamique est comparable avec des hypothèses principales de l'architecture logique de ***Data Warehouse*** (Kazienko et al 2011). D'où, les données sociales reçoivent presque les mêmes étapes de prétraitement de Data Warehouse: Nettoyage des données, la validation pour vérifier l'intégrité des données et les conserver dans une structure unifiée, détection des couches, génération du SN multicouches, définition des fenêtres de temps, la création du concept 'Parent group' (Kazienko et al 2011) et la définition des objets correspondants. L'objet principal est le réseau. 'Edge' est aussi un objet dans une couche. 'Window', 'Group', 'Layer' sont des objets qui représentent les 3 dimensions, chacun a ses propriétés (son propre schéma). Donc, ***l'agrégation des vues (suivant une, deux ou trois dimensions) offre un moyen pour analyser un OSN*** (ex. en Facebook), (Kazienko et al 2011).
(Kazienko et al 2011) ont expérimenté leur modèle sur un réseau à quatre couches. Deux du monde réel: liens familiaux et collègues de travail, et deux autres de l'univers virtuels : amis

sur Facebook et un type de relations dans un jeu (Kazienko et al 2011). Ils ont agrégé par exemple les couches du monde réel et comparé les agrégations avec celles des couches du monde virtuel, en montrant à chaque fois les avantages de la dimension temporel, et les différentes méthodes de regroupements.

PARTIE II : CONTRIBUTIONS

Chapitre 3: Comment caractériser et identifier une identité significative d'un noyau dans un SN évoluant dans le temps

1. Introduction

De nos jours, les interactions médiatisées par ordinateur entre personnes, employées, apprenants, institutions, etc. favorisent d'une manière ou d'une autre la socialisation de l'individu. Les médias sociaux prolifère et envahit non seulement les plates-formes sociales, mais aussi les différents environnements et organisations. Des phénomènes complexes sont actuellement discutés à partir des 'Online & Organizational social networks' OSNs émergeants ((Rosen et al 2011)). Nous sommes intéressés par l'une des structures sous-jacentes dans les SNs, un phénomène pas comme les autres, c'est ***la structure noyau.*** Nous posons la question sur le noyau d'un SN devant sa dynamicité temporelle (et même sa sémantique). Comment un noyau peut-il être significativement caractérisé et identifié particulièrement dans une organisation survivant dans un contexte social?

En économie, affaires, entreprises, gouvernement, organisations, réseaux téléphoniques mobiles ou de conspiration criminelle, etc., les enquêteurs, les analystes ont besoin de savoir plus sur une structure noyau, afin d'alimenter les stratégies commerciales, les décisions, les politiques et la sécurité intérieure. Le noyau est une notion phare, un sujet de recherche commun dans les SNs, évoqué dans ***la reconnaissance des élites*** ainsi que d'autres systèmes complexes. Néanmoins, moins d'attention est accordée à la structure noyau d'un SN. Souvent, les auteurs l'utilisent intuitivement de différentes manières, non seulement parce qu'il s'agit d'un phénomène complexe et sous-jacent, mais il est aussi à la mesure dans laquelle le SN est centralisé. Des débats méthodologiques commencent à se formaliser et s'installer autour d'une structure noyau, mais la plupart sont basés sur des conceptions abstraites et purement statiques ((Borgatti & Everett 2000)).

Comme tous les systèmes dynamiques, le SN évolue au fil du temps, ses acteurs changent leurs interactions et leurs affiliations. Même si les obligations institutionnelles et sociales sont fusionnées comme (ex. les interactions entre les employés d'une entreprise, systèmes de recommandation ((Lathia et al 2008))), et rendent les connexions plus limitées dans les réseaux qui semblent être stables. Cependant, les transitions entre les contextes sociaux impliquent souvent des changements ((Mollenhorst et al 2014)). La dynamique temporelle et notamment le comportement dynamique des collectivités nous inspire à révéler une identité significative d'un noyau dans un SN évoluant dans le temps. Nous proposons une approche méthodologique en suivant trois phases: Caractérisation, Modélisation et Identification. Tout d'abord, nous partons depuis trois principales caractéristiques: Cohésion, Dominance et résistance vont définir théoriquement une identité noyau. Cette identité sera conceptuellement moulée dans le concept groupe. En d'autres termes, nous explorons ses caractéristiques par des paramètres dérivés de la dynamique des groupes: partant de la cohésion interne, la persistance, la centralité du groupe, la stabilité de sa centralité dans le temps. Nous examinerons pendant combien de temps un groupement d'individus sous-jacent peut persister, s'il joue un rôle central/ efficace dans le réseau et dans quelle mesure son rôle/influence est stable. En effet, la phase de caractérisation impliquera d'autres défis: Comment pouvons-nous modéliser le SN et quel type de réseau est plus adapté? Devant ces paramètres, comment pouvons-nous identifier des structures sous-jacentes pertinentes et laquelle sera en mesure de présenter une identité de noyau? De ce fait, dans la phase de modélisation, nous formaliserons un méta-modèle d'un SN du monde réel qui évolue au cours d'une période d'observation. Ça sera 'A temporal weighted directed acyclic graph' qui représentera un processus évolutif de

groupes dans des 'time-steps' successifs. Chaque arc pondéré impliquera un chevauchement temporel entre deux groupes appartenant à deux points (snapshots) de temps successifs et qui se chevauchent. Nous définissons la fonction de pondération comme étant la composante la plus importante qui incarnera les paramètres précédents, axés sur ce type de chevauchements. Les résultats vont montrer des poids lourds qui se réfèrent à des structures sous-jacentes pertinentes: des chevauchements larges et centraux. Nous proposons dans une phase d'identification de réaliser une recherche de patterns 'critical pattern-based research' dans ce méta-modèle pour détecter le pattern critique: Le chemin le plus lourd qui couvre tous les points de temps. Nous allons découvrir que le chemin critique comprend une succession de chevauchements temporels pertinents. Lorsqu'un regroupement persiste profondément à l'intérieur, les résultats montrent qu'il présente une plus grande composition, durable avec un rôle central, le plus stable possible au fil du temps. Cette découverte indique un pattern significatif d'une structure (noyau) qui évolue profondément à l'intérieur du SN. Des tests seront nécessaires pour valider son calibre et dans quelle mesure le SN y est sensible par rapport aux autres régions.

2. Motivation qui prend source dans le monde réel

En économie et monde d'entreprises

Dans plusieurs domaines, ***la reconnaissance des élites*** évoque la structure noyau comme une classe leader qui a un large contrôle de contenu et de régulation des médias. Par exemple, les directions imbriquées '*interlocking directorates*', là où les obligations institutionnelles et sociales sont fusionnées a besoin de reconnaitre ces élites. Dans les directions imbriquées, un directeur appartient à plusieurs conseils d'administration de plusieurs entreprises '*multiple director*'. C'est l'exemple qu'on trouve avec les données collectées sur les directions imbriquées d'un ensemble de sociétés en Norvège 'Norwegian Interlocking Directorate', en Août 2009 ((Seierstad & Opsahl 2010)), (**Figure 44**).

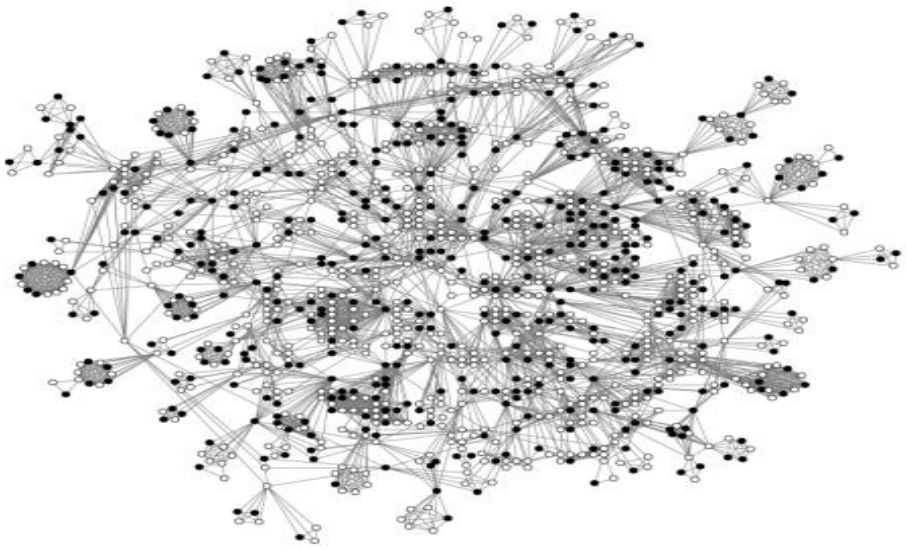


Figure 44. Structure de réseau de directions imbriquées entre 1 495 dirigeants (directeurs/administrateurs) de 367 entreprises (Norwegian Boards (Aug'09)). 2 directeurs sont liés s'ils sont membres du même conseil d'administration. Les nœuds en noir et blanc distinguent respectivement les femmes des hommes ((Seierstad & Opsahl 2010))

En effet, '*interlocking directorates*' s'est produit depuis l'apparition du 'Joint Stock Companies', les sociétés anonymes, société de capitaux ou sociétés par actions qui présentent aujourd'hui la forme principale des entreprises commerciales. Ce schéma d'imbrication, on le

trouve au début de XXe siècle (1904-5) en Ecosse avec le dataset 'Corporate interlocks in Scotland (1904-5)' ((Batagelj & Mrvar 2006)) ((De Nooy et al 2004)) ((Yongcheng Xu et al 2013)). Au XIXe siècle, la révolution industrielle a contribué au développement des chemins de fer écossais, son industrie (lourde) et l'industrie textile ((Scott & Hughes 1980)). Les entreprises s'agrandissaient et le capital dépassait les moyens des familles propriétaires ce qui a conduit à créer les sociétés par actions ((Scott & Hughes 1980)). Ainsi, les sociétés se sont représentées par des actionnaires constituant le conseil d'administration de chacune et produisant également '*interlocking directorates*'.
La découverte d'un ensemble de '*multiple director*s' peut se réfère à une classe d'élite (un noyau) qui n'a pas seulement un impact économique, mais aussi ces acteurs sont souvent nommés à des postes gouvernementaux et susceptibles de traiter des problèmes politiques en faveur de leurs entreprises. Pour ((Asimakopoulos 2009)), c'est la classe régnante, dirigeante et dominante dans la politique moderne.

En politique

On prend l'exemple d'un réseau de pouvoir mexicain 'Mexican political elite' entre 1910 et 1994 ((Gil-Mendieta & Schmidt 1996)) ((De Nooy et al 2004)). Pendant la quasi-totalité du XXe siècle, le pouvoir politique mexicain a été dans les mains d'un groupe restreint de personnes qui ont été liées par des relations d'affaires, adhésion d'institutions politiques, liens familiaux et amitié. C'est un réseau où les présidents et leurs plus proches collaborateurs constituent ***le noyau de cette élite politique*** (composé par 37 acteurs) qui a été étudié par ((Gil-Mendieta et al 1997)) ((Gil-Mendieta & Schmidt 1996)). L'effet marquant et même frappant se trouve dans les élections présidentielles, la nomination des candidats pour la succession des présidents. Depuis 1929, et suivant trois générations de politiciens, les présidents se succèdent, mais chaque candidat avait maintenu des liens avec ces anciens présidents et leurs collaborateurs (ses prédécesseurs). Chaque nouveau président était même un secrétaire dans le cabinet du président précédent (il collabore même avec lui) ((Gil-Mendieta & Schmidt 1996)). 'Partido Revolucionario Institucional' était le parti politique qui a remporté toutes ces élections même s'il contenait deux groupes en compétition pour atteindre pouvoir. Par conséquent, ***une sorte d'élite politique a maintenu le contrôle du pays avec le temps***. Après la révolution, l'élite politique ***a été dominée*** par les militaires qui semblent être l'opposition réelle mais les civils ont progressivement repris le pouvoir ((Gil-Mendieta & Schmidt 1996)).

Réseaux illégaux

La détection de noyau pèse sur les enquêtes menées sur les réseaux illégaux évoluant dans le temps, dissimulés et s'installant même dans certaines organisations. C'est une notion importante dans l'étude des réseaux des criminels, les crimes organisés et ceux qui bénéficient et affectent des entreprises elles-mêmes par des comportements frauduleux malgré leurs réseaux d'interactions strictes. L'étude d'une telle structure est exportable vers les chercheurs et experts dans la lutte antiterroristes à l'image des travaux du centre CASOS qui se présente comme illustration d'analyse de systèmes organisationnels/ sociaux ((Frankenstein et al 2015)) ((Fellman et al 2011)) ((Il-Chul et al 2008)) ((Il-Chul & Kathleen 2007)) ((Max & Kathleen 2005)). L'un de ses réseaux étudiés est construit à partir les six grands groupes qualifiés comme 'terroristes' qui s'opèrent en Cisjordanie (West Bank) : 'Al Aksa Martyrs Brigades, Al Fatah, Al Qaeda, Hamas, Hezbollah, Islamic Jihad'. Ces réseaux sont construits à partir 18 textes extraits via 'LexisNexis Academic', une société pionnière dans l'accessibilité électronique des documents juridiques et journalistiques ((Stephen 2012)) depuis 'The Economist, The Washington Post et The New York Times'. 'Madrid Train

bombing' est un autre exemple de réseaux de contacts entre terroristes ((Hayes 2006)). Ce sont des contacts de terroristes présumés, impliqués dans les attentats de Madrid le 11 Mars 2004, en incluant des relations d'amitié, co-participation à des camps d'entraînement ou des attaques précédentes.

Remarque : L'identification d'un noyau dans tels SNs n'est pas toujours faisable. Dans nombreux réseaux illégaux (criminels), les besoins de secret et de dissimulation se croisent avec les besoins d'efficacité de taches à réaliser et de coordination, ce qui mène à centraliser ou décentraliser ces réseaux, dans ce dernier cas, leurs classes dominantes sont susceptibles d'être cachées.

Réseaux de Citations

De plus en plus, les études des réseaux de citations scientifiques ont besoin de plus en plus de découvrir des documents noyau parmi un ensemble d'articles académiques.

3. Problématique, motivations techniques et orientation

Qu'est-ce qu'un noyau de SN sur le plan topologique statique:

Pour les réseaux informatiques en général, un noyau se distingue comme la partie solide (difficile), interne ou centrale dans les échanges d'informations. Il donne au système son existence, un caractère ou son équilibre. Pour un SN qui est intuitivement modélisé par des graphes non-aléatoires avec des caractéristiques spécifiques et des structures sous-jacentes (McGlohon & Faloutsos 2008) (Nettleton 2013), le noyau est une notion commune mais souvent informelle en SNAM selon différents points de vue. En se basant sur des conceptions intuitives, il est principalement défini comme la partie dense et cohérente du réseau ((Borgatti & Everett 2000)). Il est étonnant que peu de travaux de recherche soient intéressés par la structure noyau d'un SN et ses propriétés ((Borgatti & Everett 2000)) par rapport à ce qui a été réalisé en SNAM à l'heure actuelle.

Premières conceptions

((Borgatti & Everett 2000)) proposent l'un des rares œuvres complétement focalisés sur ce phénomène et qui constituent un point de départ lançant un débat méthodologique pour le formaliser et l'identifier. ((Borgatti & Everett 2000)) ont essayé de formaliser l'une des descriptions intuitives du noyau étant une région cohésive. Ils se sont basés sur la décomposition de la matrice d'adjacence du SN. Une modélisation par blocs 'block modeling' ((Borgatti & Everett 2000)), suivant la densité des liens qui permet de distinguer les zones cohésives. Pour eux, *un modèle idéale de blocs permettra de distinguer le noyau comme la région la plus cohérente formée par un sous-ensemble d'individus densément liés*, unis dans un seul bloc (modèle discret) ((Borgatti & Everett 2000)). Dans une version plus étendue de ce modèle idéal, les auteurs vont au-delà de cette propriété en s'intéressant aux individus qui forment cette région. Chacun est associé à un score de degré de noyauté 'coreness' index' ((Borgatti & Everett 2000)). C'est un score dont la définition est théoriquement basée sur la notion de K-core ((Malliaros et al 2016)) qui se présente par le *sous-graphe maximal* qui contient des nœuds ayant un degré k ou plus. Pour certains d'autres auteurs, 'coreness' index' est inspiré de la centralité de proximité par rapport à un centre structurel (centroide) ((Borgatti & Everett 2000)). Par conséquent, ((Borgatti & Everett 2000)) considère *une région noyau comme un sous-ensemble d'individus ayant les degrés de noyauté ou encore les centralités les plus élevés* (modèle continu). Pour une partition susceptible de représenter un noyau et le reste du réseau, un test statistique, (coefficient de corrélation), est utilisé pour mesurer sa corrélation avec son modèle idéal ((Borgatti & Everett 2000)). Dans le cas où il n'y a pas de

partition à proposer, le coefficient de corrélation se transforme en une fonction objective d'un algorithme d'optimisation utilisé pour rechercher un modèle idéal, une partition la plus susceptible de représenter une structure noyau.

Problèmes, autres concepts et objectifs

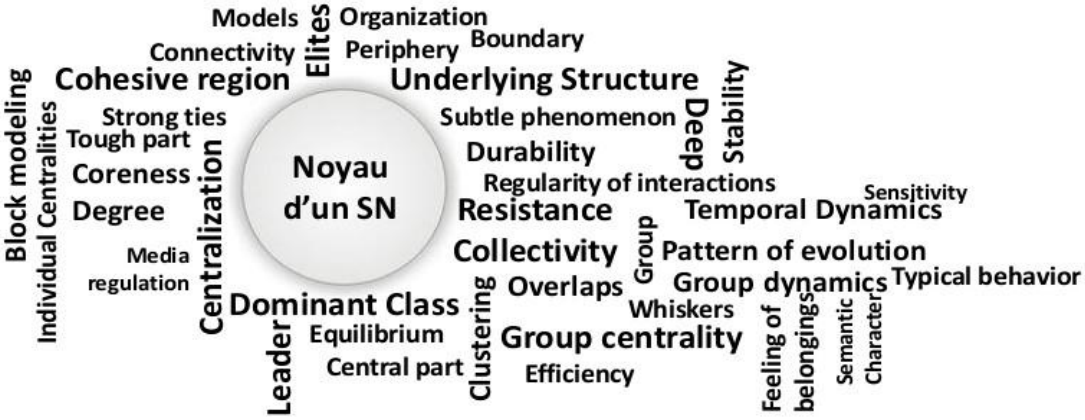


Figure 45. Nuage des propriétés, concepts, notions autour d'une structure noyau d'un SN. Certains ont été abordés dans un contexte statique et d'autres on va les introduire suivant nos exigences: Contexte dynamique et une vue plus conceptuelle 'de collectivité'

- Cependant, ce type d'études a été expérimenté sur des données d'interactions comme les proximités physiques (ex. troupe de singes ((Borgatti & Everett 2000))) qui ne semblent pas aujourd'hui assez bénéfiques (en termes de pertinence) pour une exploration plus profonde (comme le cas des interactions humaines les plus orientées).
- En outre, ces études sont fondés dans des ***cadres statiques*** (et structurelles) et n'abordent pas ***la structure noyau dans son ensemble***. Récemment, ((Wang et al 2013)) localise la structure noyau comme une zone d'intersection des groupes ((α, β)-communautés) qui se chevauchent, dans un graphe social statique et dense. La structure noyau est abordée aussi par ((Leskovec et al 2009)) quand ils décrivent la plupart des structures des SNs par un modèle de 'core-and-whiskers' ((Leskovec et al 2009)) (Parthasarathy et al 2011). Dans ce cas, le réseau se compose d'un noyau qui est en effet un groupe entouré par des communautés appelées 'whiskers' (Padgett 1994) (Robins et al 2005).
- Par conséquent, ***nous pouvons distinguer deux orientations pour décrire une structure noyau***. ***La première se contente sur sa structure interne et ses éléments individuels qui la composen***t. Dans ce sens, le degré de noyauté ou d'autres mesures de centralité ajustées, utilisées pour identifier les éléments qui forment noyau, sont simplement focalisés sur les individus. Alors que les questions se posent sur le potentiel du noyau en tant qu'une collectivité.
- Des réponses clés peuvent être fournies à travers des mesures généralisées sur les groupes 'Group centrality' (sous-section: Métriques pour les groupes) même si la plupart d'entre eux considèrent tous les nœuds externes (Everett & Borgatti 1999) ((Everett & Borgatti 2004)). Une approche plus précise aura besoin d'une configuration de réseau en groupes pour distinguer les acteurs qui forment les frontières, ceux qui forment des périphéries et même un noyau, alors que certains d'autres peuvent être évités en diminuant le coût de calcul. On peut dire que la centralité de groupe a du potentiel qui permet de trouver des regroupements centraux et expliquer des phénomènes liés à la réussite et l'efficacité des groupes (Everett & Borgatti 1999).

- ***La seconde orientation explique la composition d'un noyau, d'un point vue de collectivité, c'est dans ce sens-là qu'on retrouvera l'une des considérations de notre proposition***. Néanmoins, tous les indices et les modèles qui ont été utilisés sont statiques et ignorent souvent la dynamique du SN. Alors que nombreuses illustrations considérant la dynamique temporelle ont fait preuve de réalisme, en obtenant plus de précision et des interprétations plus significatives par rapport au contexte statique qui peut être trompeur. Même s'il y a des extensions de mesures de centralités adaptées aux graphes dynamiques on parle rarement de l'importance des groupes dans le temps.

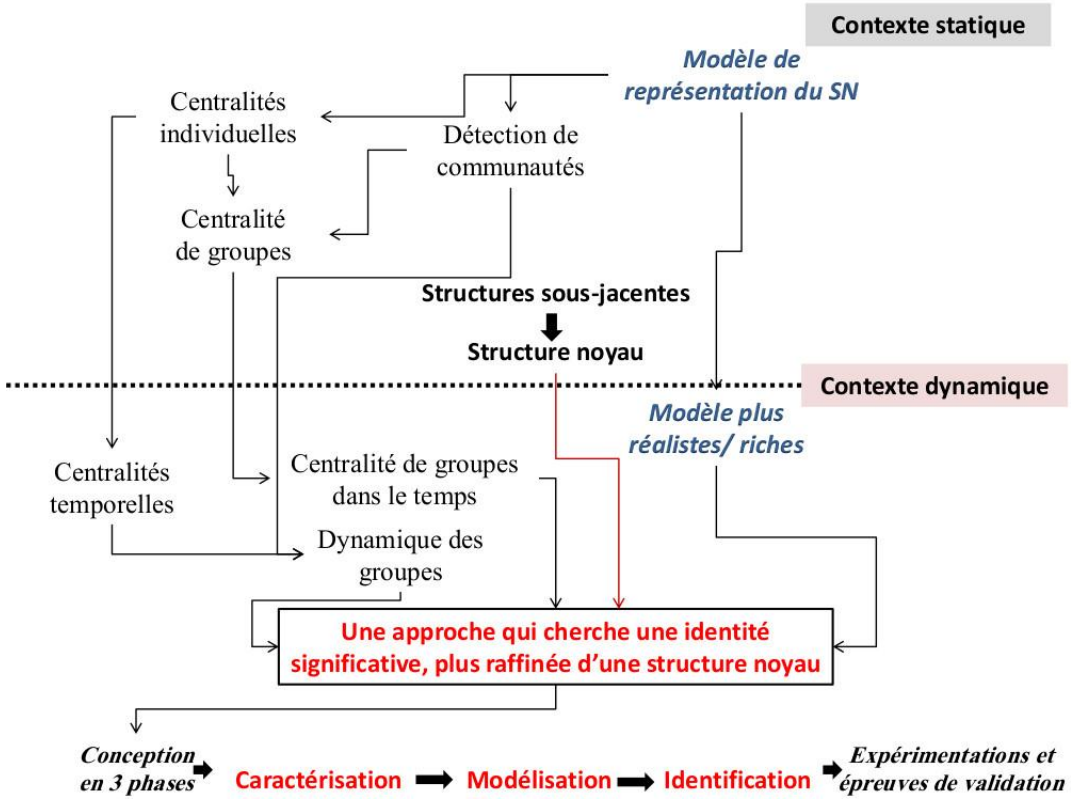


Figure 46. Les sujets abordés dans les parties précédentes et qui sont les sujets connexes liés à notre approche proposée avec ses étapes principales

- Donc, la détection de noyau doit prendre en compte les travaux de recherche menés récemment sur la manière dont le graphe social évolue: Partant des analyses statistiques vers la modélisation des patterns d'évolution, dynamiques des groupes, qui peuvent contribuer à expliquer le comportement/ l'influence d'une structure noyau significative.
- Il est vrai qu'il s'agit d'une région dense composée par des individus centraux. Cependant, même les études récentes ((Wang et al 2013)) ((Taylor et al 2016)) ((Min-Joong et al 2016)) ((Smalla 2013)) n'ont pas clarifié la définition de ce phénomène, comment il se caractérise, comment il se produit et domine le réseau dans le temps sous une vue de collectivité. Par conséquent, nous pensons à exploiter plusieurs concepts, aspects temporels et des paramètres liés à la dynamique des groupes (persistance, centralité de groupe, etc.) afin de pouvoir qualifier une région cohésive comme une structure noyau (une classe élite) plus significative : Un regroupement qui affiche un comportement spécifique profondément à l'intérieur du SN dynamique.
- Partant des tentatives de repérage vers les épreuves de validation, ***nous proposons une approche pour caractériser et identifier une identité significative d'un noyau dans***

un SN évoluant dans le temps. Il y a 3 phases principales : Caractérisation, modélisation et identification.

4. Préliminaires, caractéristiques clés et paramètres

On se base sur trois concepts (caractéristiques) clés pour caractériser une identité noyau à l'intérieur d'un SN qui évolue dans le temps (**Figure 47**). Cette identité sera moulée dans un concept de groupe/ sous-groupe. En d'autres termes, chacune de ses caractéristiques sera explorée par un paramètre dérivé (tiré) du concept de groupe et sa dynamique (**Figure 47**).

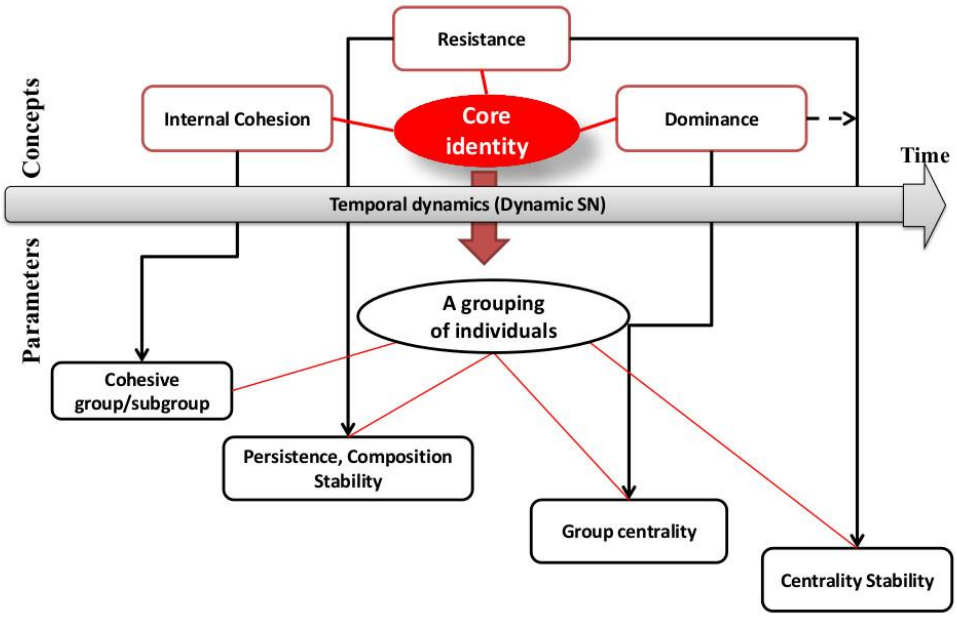


Figure 47. Caractéristiques clés (cadres rouges) décrivant une identité significative d'une structure de noyau dans un SN évoluant dans le temps. Un noyau est moulé dans le concept de groupe de telle sorte que ses caractéristiques soient décrites par des paramètres (cadre noir) liés à la dynamique des groupes

4.1.Cohésion interne décrite par la cohérence d'un regroupement durable

Une structure noyau est communément connue sous forme d'une zone dense et cohésive, dans les modèles statiques. Dans le cas où les liens sont pondérés ou hétérogènes (Zhou et al 2007), cette cohésion interne peut se renforcer dans des structures comme 'the core discussion network' ((Smalla et al 2015)). C'est-à-dire, un sous-ensemble d'individus fortement liés peut présenter un noyau du SN. Cependant, selon les récentes perspectives qu'on trouve dans ((Smalla et al 2015)) ((Smalla 2013)), la densité ou la force des relations entre les acteurs ne pourra pas incarner seule cette cohésion, à moins qu'ils interagissent régulièrement dans le temps. En effet, les relations notamment dans les OSNs, s'exposent au risque d'interruption et disparition avec le temps, à cause de la distance physique et l'absence des opportunités de rencontres. D'autre part, le SN tend à se densifier au fil du temps, sous l'influence de certains phénomènes comme 'rich get richer'. Certaines parties sont susceptibles d'être plus cohérentes, plus saillantes. Donc, les interactions régulières ((Smalla et al 2015)) est un critère essentiel dans la formation et maintien des régions cohésives (des groupes, des noyaux, etc.) et susceptibles d'être durables. Comme la montre la **Figure 47**, la cohésion d'une structure noyau sera prouvée comme un aspect temporel, en s'inspirant du concept de collectivité (groupe) dans le temps. Elle sera certainement exprimée à un moment donnée par un regroupement d'individus, mais il sera nécessaire de savoir si cette région cohésive peut être préservée, en suivant sa durabilité et sa dominance dans le temps.

4.2.Dominance (apparence) décrite par un regroupement central

Il est vrai qu'une structure noyau peut rassembler des acteurs qui jouent individuellement des rôles efficaces et stratégiques sur les flux de communication (des centralités de communication ((Zhai et al 2013))). Cependant, d'un point de vue global, la dominance d'un probable noyau dans son ensemble reste en fait inconnue. Souvent, ces individus n'ont pas une motivation collective pour dominer stratégiquement le SN. Nous n'aurons pas besoin de connaître le rôle individuel de chacun des membres. On cherche plutôt à évaluer le rôle joué par le regroupement pour savoir s'il présente la dominance d'un noyau. Comme c'est présenté dans la **Figure 47**, cette dominance sera évaluée par le paramètre de centralité de groupe ('Group Centrality' GC), en utilisant soit: 'Group degree centrality' (GDC)/ 'closeness' (GCC) ou betweenness (GBC) (Everett & Borgatti 1999), mais en lui donnant une dimension temporelle.

Le paramètre GC est lui-même un défi, car le groupe A (**Figure 48**) en question peut ne pas conserver tout le temps la même composition et donc il ne s'agit pas du même groupe. Par conséquent, nos estimations seront basées sur certaines hypothèses (**Figure 48**).

- D'abord, on va se baser à un moment donné sur une configuration de réseau en groupes pour évaluer le GC de chaque groupe A (**Figure 48**). Certains individus externes en dehors de A (Outside group) pourraient être évités dans le calcul.
- L'évaluation de la centralité (GC) temporelle du groupe doit être strictement conditionnée par sa composition stable dans le temps.
- En outre, on sait que la partie frontière d'un groupe, représente le canal le plus important pour interagir avec les autres parties du réseau. Ainsi, on ne considère que les interactions externes des membres frontières de A pour évaluer son GC.

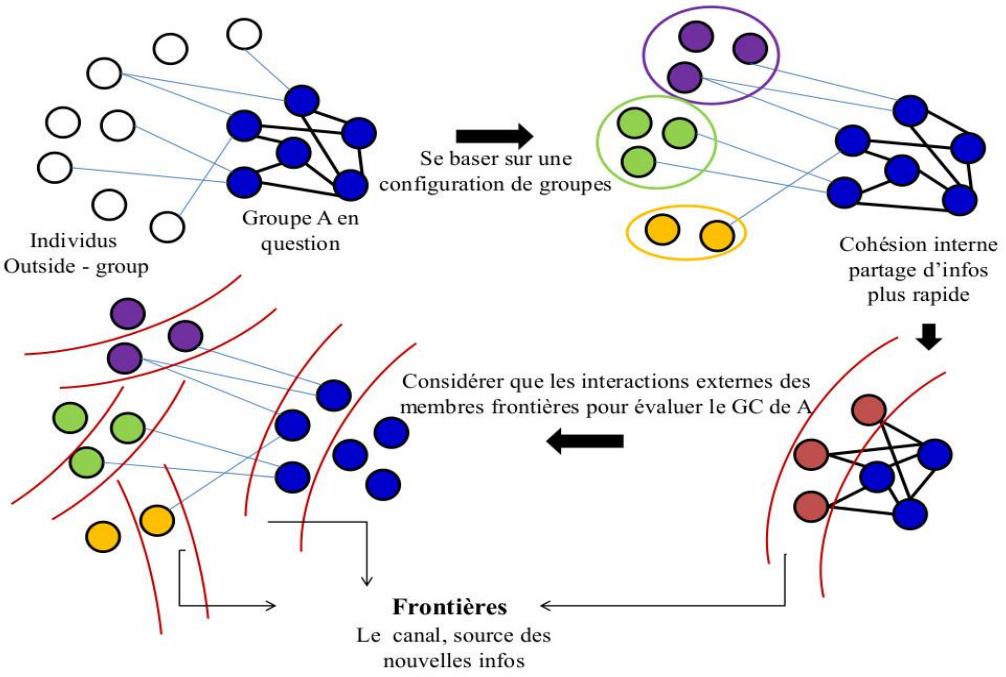


Figure 48. Hypothèses enchainées pour évaluer le GC d'un groupe à un moment donné

Par conséquent, la dominance d'une structure ayant une identité noyau sera prouvée par un regroupement (stable) central. Mais, nous devrons montrer à quel point ce rôle central peut être maintenu au cours d'une période d'observation (**Figure 47**).

4.3.Résistance (durabilité) décrite par la stabilité en termes de composition et de centralité

La résistance est un caractère significatif pour une identité noyau. Elle affecte significativement la cohérence d'une représentation de SN dans le temps par rapport au cas statique. Le noyau doit être le point de résistance devant la dynamicité temporelle du SN. Quand il est affecté par un changement dit profond, la représentation sociale peut se disloquer ((Aïssani 2009)). La **Figure 47**, montre qu'on va explorer la résistance d'un noyau dans des régions cohésives (regroupements), à deux niveaux (deux paramètres):

- **Stabilité de composition :**

  Nous cherchons pendant combien de temps, un regroupement donné persiste (dure) profondément. Ce qui veut dire qu'on cible premièrement la stabilité de sa composition. Pendant un intervalle de temps, un groupe sera qualifié stable/durable, s'il conserve strictement tous ses éléments qui le composent. Un groupement durable représentera le premier niveau de résistance d'une identité noyau et aussi son infrastructure pendant toute la période d'observation du SN (**Figure 47**). Ainsi la stabilité de composition sera le paramètre critique, par laquelle les paramètres sont conditionnés.

- **Stabilité de centralité**

  La stabilité de centralité dans le temps est le deuxième niveau (paramètre) de résistance d'un noyau (**Figure 47**). Mais, on ne pourra pas évaluer le degré de stabilité de centralité d'un regroupement donné, à moins qu'il persiste dans le temps. Il convient de noter aussi que ce paramètre est plus ou moins perturbé. Un groupe/ sous-groupe G peut conserver plus facilement l'intégralité de sa composition que de jouer un rôle central parfaitement stable dans temps. Lorsque G persiste, sa centralité de groupe est plus sensible aux changements microscopiques effectués par les acteurs (membres ou externes) par rapport aux affiliations. Nous proposons le concept d'***Amplitude de Centralité*** afin de capturer cette perturbation et évaluer ainsi la stabilité de centralité.

  Définition de l'Amplitude de centralité de groupe :

  L'amplitude de centralité ('Centrality Amplitude' : CA) d'un groupe G qui persiste dans un intervalle de temps $[T_k, T_l]$ au cours d'une période d'observation $[T_1\ T_t]$, est défini par la différence entre la valeur maximale et minimale de son et GC, $1 \leq k < l \leq t$. Si $k = 1$ et $l = t$, alors CA désignera une amplitude globale de centralité ('Overall Centrality Amplitude' : OCA) de G.

D'après la **Figure 47**, les trois caractéristiques clés d'une identité noyau seront explorées à la lumière des paramètres d'évolution des groupes dans le temps. ***Le but consiste à trouver le comportement typique d'une structure noyau plus raffinée et significative***. Elle se présente par un grand regroupement cohérent, équilibré, qui persiste et joue un rôle central, le plus stable possible. Partant de la cohésion qui est topologiquement incarnée par un groupe/ sous-groupe, les autres paramètres sont dépendants et significativement ordonnés (**Figure 47**): La persistance est une condition préalable pour passer à explorer la centralité du groupe et la stabilité de centralité.

5. Approche de modélisation (Conceptions)

Après avoir abordé la phase de caractérisation qui dégage un ensemble de paramètres (**Figure 47**, **Figure 49**), nous concevons une approche méthodologique appropriée de modélisation. Il

s'agit de modéliser un processus évolutionnaire du SN afin de pouvoir identifier cette l'identité noyau profondément à l'intérieur. Donc, ça ne sera pas un simple modèle de graphe social. Devant ce contexte de dynamicité temporelle, devant les exigences d'une fouiller de SN en mettant l'accent sur ses structures sous-jacentes, essayant de comprendre et percer dans sa profondeur, nous pensons plutôt à développer un formalisme de méta-modèle appelé TW-DAG.

Le TW-DAG (Temporal Weighted Directed Acyclic Graph) se présente par un graphe temporel acyclique, orienté et pondéré. Sa conception se fait en sous-étapes comme la montre la **Figure 49**. Nous partons des données/informations temporelles d'un SN évoluant dans le temps, qui seront une abstraction de bas niveau de sa dynamique temporelle. Nous mettons une étape préliminaire pour examiner la tendance des individus à se regrouper dans le temps, avant la modélisation (**Figure 49**). Ensuite, un formalisme sera proposé pour définir formellement les composantes de TW-DAG : partitions en groupes, chevauchements temporels, fonction de pondération (**Figure 49**). Cette dernière sera le support nécessaire pour exprimer les paramètres dérivés, là où on arrivera à un niveau d'abstraction plus élevé (**Figure 49**).

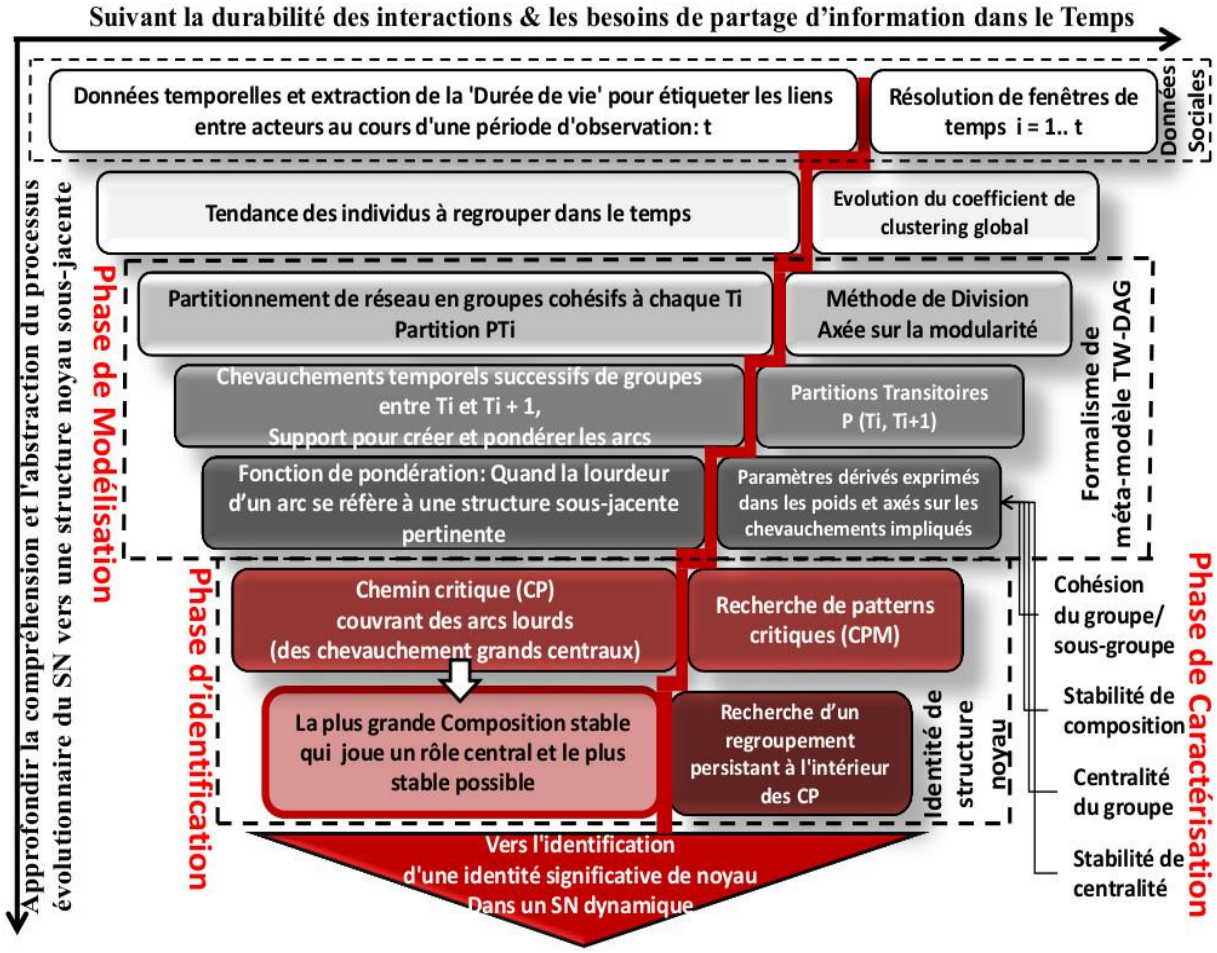


Figure 49. Approche méthodologique de conception : Phase de modélisation du processus évolutionnaire du SN par un méta-modèle et phase d'identification d'une identité significative d'une structure noyau profondément à l'intérieur

La phase d'identification d'une structure noyau sera axée sur une recherche de patterns critiques appliquée dans ce méta-modèle TW-DAG (**Figure 49**). Dans chaque ligne une paire de rectangles représente une étape décrite dans le grand rectangle alors que le plus petit désigne la technique ou le schéma utilisé ou proposé (**Figure 49**). Le contraste des couleurs indique

plusieurs niveaux de dynamicité temporelle du SN vers la profondeur (**Figure 49**). Chaque étape ou phase sera détaillée dans les parties suivantes.

5.1.Données sociales temporelles, une simple abstraction de la dynamicité temporelle

Malgré leur richesse, les données sociales temporelles sont soumises sous différentes restrictions de confidentialité et différents formats de données compliqués (ex. sous le format CMX dans l'outil Commetrix). Pour un SN donné qui évolue dans le temps, la période d'observation, le temps de création ou de suppression des liens, leur ordre chronologique, etc., permettent de capter une simple abstraction de sa dynamique temporelle.

Définition de la durée de vie des liens

Nous considérons la durée de vie de la relation entre deux acteurs une information temporelle fondamentale pour comprendre aussi bien la dynamique temporelle du SN que la conception de son méta-modèle TW-DAG (**Figure 49**).

Au cours d'une période d'observation $[T_1\ T_t]$, la durée de vie d'une relation donnée est un intervalle de temps $[T_k, T_l]$ tel que $1 \leq k < l \leq t$, pendant lequel la relation est maintenue par des événements (des activités) de liaison temporelle entre les deux acteurs impliqués. Cet intervalle de temps peut contenir plusieurs snapshots et chaque snapshot est limité par deux points de temps successifs.

Par conséquent, chaque lien sera étiqueté par une durée de vie dans le réseau. Etant le schéma qui va structurer la dimension temporelle, la résolution des fenêtres de temps (Snapshots, les instants, les points de temps, etc.) affecte l'extraction de la durée de vie des liens (**Figure 49**). Même si un modèle continu est infaisable, mais on doit se rapprocher d'une résolution optimale pour pouvoir développer un TW-DAG plus réaliste menant à des interprétations plus significatives.

5.2.Tendance à former des groupes, les sommets de TW-DAG

Pour un SN donné, la tendance de ses acteurs à se regrouper dans le temps est une condition préliminaire avant de développer les composantes de TW-DAG (**Figure 49**). À chaque point de temps, cette tendance sera évaluée à travers la version généralisée de l'indice de coefficient de clustering (CC) sur l'ensemble du réseau: La moyenne des coefficients locaux ou le CC de Watts-Strogatz. Cela permettra d'examiner l'évolution du CC, comment il varie et ainsi comment la connectivité du réseau change au fil du temps. Le schéma de cette étape se base sur une visualisation descriptive de l'évolution du CC pour trouver des phases d'évolution intéressantes avant la modélisation du SN étudié : Son CC est-t-il plus, moins, ou assez stable au cours de la durée d'observation?

5.3.Formalisme spécifique pour la conception d'un méta-modèle TW-DAG

S'il existe deux méthodes pour concevoir des modèles de SNs dynamiques (comme c'est déjà abordé dans la partie : Modèles de représentation les plus connus), nous définissons un méta-modèle qui fusionne les deux méthodes et répond aux exigences de caractérisation, plutôt qu'un simple graphe social communément connu. Il sera en mesure de supporter explicitement l'aspect temporel. En plus, il permettra d'exprimer et suivre non seulement le paramètre de persistance, mais aussi des paramètres comme la centralité et la stabilité e centralité des regroupements sous-jacents dans le temps. Etant donné un SN observé pendant une durée divisée en point temps i = 1...t, on définit son méta-modèle par TW-DAG ($\{V_1, V_2 ... V_t\}$, A (T_i, T_{i+1}), W).

5.3.1. Des sommets impliquant des groupes à chaque point de temps T_i

L'ensemble des sommets V de TW-DAG est décomposé en t parties. Chaque partie ou sous-ensemble de sommets V_i correspond à une partition PT_i du SN à un point temps T_i. PT_i est formée par un ensemble de groupes G_x-T_i, x = 1... X, (X ∈ N) l'indice du groupe dans la partition. Ainsi, chaque sommet de V_i représente un groupe de PT_i.

En effet, une fois que les acteurs du SN tendent à se regrouper, des régions cohésives (groupes) émergent (un sujet intéressant en science comportemental des collectivités (Gilbert et al 2010) (Cuvelier & Aufaure 2011). En utilisant l'algorithme de division 'Louvain Method' (Nettleton 2013) (Blondel et al 2008), nous partitionnons le réseau à chaque point de temps T_i ce qui donne la partition PT_i (**Figure 49**). En termes de cohésion, la méthode assure la découverte des groupes de bonne qualité, optimisée par une fonction de modularité.

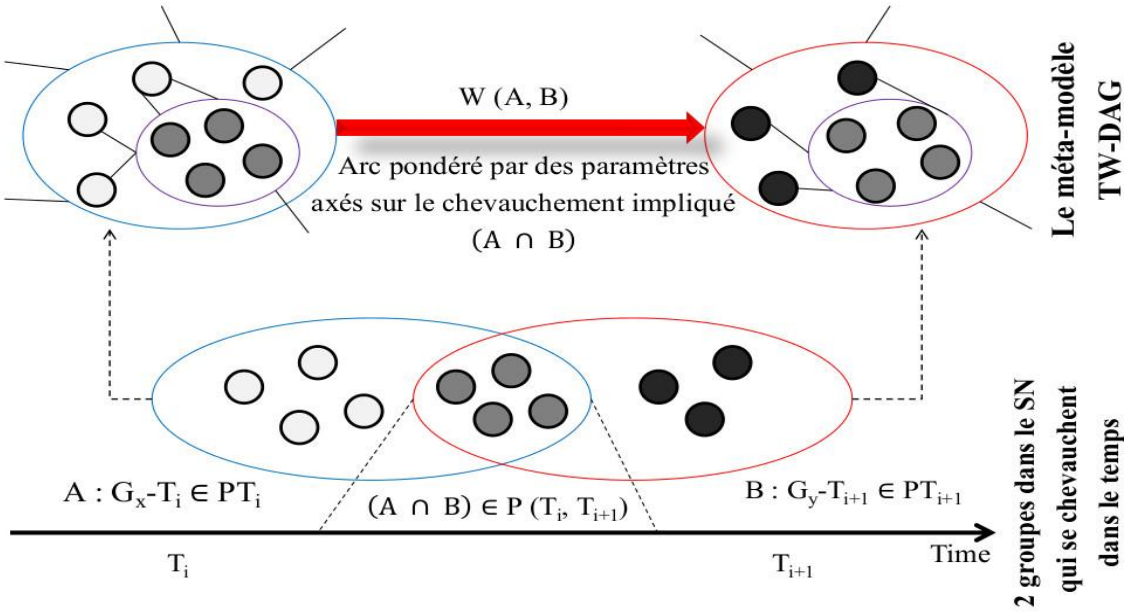


Figure 50. Aperçu sur deux groupes appartenant à deux partitions successives et qui se chevauchent, dans le temps. Ce chevauchement produit un arc pondéré entre ces deux groupes (sommets) dans le méta-modèle TW-DAG

5.3.2. Des arcs impliquant des chevauchements temporels

Chaque arc de l'ensemble des arcs A de TW-DAG, est notée comme a (G_x-T_i, G_y-T_{i+1}) ∈ A (T_i, T_{i+1}). Il connecte deux groupes (deux sommets) appartenant à deux partitions successives dans le temps (**Figure 50**). Autrement dit, G_x apparaît à T_i, un instant précédent juste avant G_y à T_{i+1}. Mais selon la **Figure 50**, un arc ne peut pas être créé, à moins que G_x-T_i et G_y-T_{i+1} se chevauchent et le poids correspondant W (G_x-T_i, G_y-T_{i+1}) ≠ 0, i = 1...t-1.

Lorsqu'un individu se socialise, il peut rejoindre différent groupes avec le temps. Il est clair que l'affiliation chronologique des individus permet aux groupes de se chevaucher dans le temps. Autrement dit, les groupes provenant de différentes partitions PT_i peuvent se chevaucher. Cependant, seulement les chevauchements temporels entre les points de temps successifs sont considérés pour créer les arcs de TW-DAG. On parle aussi de ***chevauchements temporels successifs***. Nous définissons ce type de chevauchement comme un regroupement (un sous-groupe) ou une structure sous-jacente ***transitionnelle***. Etant donné deux partitions successives dans le temps PT_i et PT_{i+1}, composées respectivement par X et Y

groupes. L'ensemble de chevauchements temporels possibles qui se produisent entre PT_i et PT_{i+1} est défini par un autre type partition P (T_i, T_{i+1}) appelée partition transitionnelle ou transitoire (**Figure 49**) tel que P (T_i, T_{i+1}) = {O_z–(T_i, T_{i+1})\ O_z–(T_i, T_{i+1}) = G_x-T_i ∩ G_y-T_{i+1}, G_x-T_i ∈ PT_i, G_y-T_{i+1} ∈ PT_{i+1}, x = 1...X, y = 1...Y, z = 1...Z, Z ≤ X×Y}.
Pratiquement, le nombre |P (T_i, T_{i+1})| = Z ne peut pas atteindre le nombre théoriquement prévu de chevauchements X × Y entre PT_i et PT_{i+1}. Dans un tel contexte dynamique réduit, il est très fréquent de trouver une large proportion de G_x-T_i conservée dans l'un des groupes G_y-T_{i+1}, tandis que le reste est dispersé dans un ou plusieurs autres groupes G_w-T_{i+1} (w=1..Y). Parfois même G_x-T_i = G_y-T_{i+1}, ce qui veut dire que G_x-T_i persiste localement entre T_i et T_{i+1}. Ainsi, G_x-T_i se chevauche rarement avec tous les groupes de PT_{i+1}. Si on suppose l'inverse, il faut que |G_x-T_i| ≥ Y et ses membres soient dispersés sur tous les groupes PT_{i+1}. Ce sont des conditions pratiquement irraisonnables, difficilement à remplir entre deux instants successifs. En outre, la résolution des fenêtres de temps peut avoir aussi une influence sur ce taux de chevauchement. Si elles sont assez larges, il est probable que la composition d'un groupe G_x-T_i sera plus différente à T_{i+1}. C'est le contraire qui se produit lorsque les fenêtres sont plus étroites car la composition devient de plus en plus statique.

Pourquoi les chevauchements temporels successifs ?

- Ce type de chevauchement permet de fouiller verticalement dans la profondeur du SN et se projeter au même temps horizontalement dans son processus d'évolution au fil du temps (**Figure 49, Figure 50**).
- Il concrétise localement le caractère de durabilité entre T_i et T_{i+1}
- Etant une composition sous-jacente stable, il est bien adapté pour appliquer et explorer les hypothèses de calcule de GC en profondeur ainsi que les autres paramètres étape par étape dans le temps (**Figure 50**). C'est une structure transitionnelle, avantageuse pour suivre l'évolution de ces paramètres
- Il fera l'objet de la fonction de pondération proposée.

5.3.3. Fonction de pondération

La fonction de pondération des arcs W: A (T_i, T_{i+1}) → R, (i = 1...t-1) est la composante critique dans TW-DAG en faisant le lien avec les dérivés de la phase de caractérisation (**Figure 49**). Etant donné deux groupes G_x-T_i noté A ∈ PT_i et G_y-T_{i+1} noté B ∈ PT_{i+1}(**Figure 50**), W est formulée de telle sorte que chaque poids W (A, B) soit déterminé à partir les paramètres (Composition, centralité, stabilité de centralité) du chevauchement temporel impliqué A∩B, étant un regroupement cohérent et stable entre T_i et T_{i+1}(**Figure 50**). Autrement dit, la fonction W est le support avantageux qui nous permettra d'exprimer, afficher et suivre l'évolution des paramètres des structures sous-jacentes étape par étape dans le temps. Mais, ***les paramètres seront ajoutées progressivement un par un dans les poids afin d'arriver à une combinaison justifiée, bien équilibrée et représentative***. En effet, nous avons tendance à définir cette combinaison de telle sorte que la lourdeur des poids soit proportionnelle à la pertinence des structures impliqués. Ainsi, l'arc le plus lourd W (A, B) sera en mesure de refléter une structure sous-jacente A∩B ∈ P (T_i, T_{i+1}) pertinente comme il est montré dans la **Figure 53**. D'abord, si A∩B est vide, W (A, B) est effectivement nul 0, sinon W (A, B) est défini comme suivant:

5.3.3.1.Formule (1) de pondération (exprimer le paramètre de centralité GC)

Même si A∩B persiste entre T_i et T_{i+1} (**Figure 50**), il n'a pas forcément la même GC à T_i et à T_{i+1}. Donc ce paramètre est en premier plan exprimé dans W (A, B) par une moyenne de centralité temporelle de groupe qui est en effet une moyenne de centralité de chevauchement A∩B, affichée à T_i et T_{i+1}. GC peut être spécifié comme GDC, GCC ou GBC.

$$W(A,B) = \frac{GC_{T_i}(A \cap B) + GC_{T_j}(A \cap B)}{2}, \quad (35)$$

$$j = i+1,\ i = 1...t-1.$$

D'où, on peut citer trois cas particuliers :

- *Composition stable mais isolée*: C'est un cas qui s'illustre avec cette pondération lorsque W (A, B) = 0. Cela veut dire que : A∩B est isolé à T_i et à T_{i+1}. Alors que selon la **Figure 50** le principe de cohésion dans A ou B confirme que tout intersection A∩B a normalement des liens avec A (à T_i) et avec B (à T_{i+1}). En effet, W (A, B) = 0 signifie que A et B semblent présenter la même composition (A∩B = A = B), le même groupe stable mais isolé à T_i et à T_{i+1}.
- *Chevauchement singleton* : Lorsque A∩B est singleton, il peut jouer un rôle plus central par rapport un autre chevauchement C∩D ∈ P (T_i, T_{i+1}) qui n'est plus singleton. Donc même si |A∩B| = 1 << |C∩D|, on peut tomber dans le cas où W (A, B) > W (C, D). Cette pondération montre que le paramètre GC seul ne peut pas représenter la pertinence d'une composition (sa taille).
- *Poids Trompeur*: Si la centralité de A∩B est nulle à T_i ($GC_{T_i} = 0$), tandis qu'il est beaucoup plus central à T_{i+1} (ou vice versa), la moyenne peut être désavantageuse car W (A, B) peut afficher une valeur élevée qui ne reflète pas vraiment une structure pertinente à T_i et T_{i+1}.

Bien que le premier cas semble incontournable, les deux autres nécessitent d'introduire plus d'ajustements et de paramètres

5.3.3.2.Formule (2) de pondération (Déterminer des structures plus larges et centrales)

Le deuxième cas particulier de la première formule de pondération affirme que les poids n'afficheront que des structures sous-jacentes (chevauchements) centrales entre T_i et T_{i+1}. De ce fait, la moyenne sera pondérée par la taille de la structure (nombre des membres de sous-groupe), étant une composition stable entre T_i et T_{i+1}. Ainsi chaque poids W (A, B) devient plus représentatif et équilibré :

$$W(A,B) = |A \cap B| \times \frac{GC_{T_i}(A \cap B) + GC_{T_j}(A \cap B)}{2}, \quad (36)$$

$$j = i+1,\ i = 1...t-1$$

Cependant, cela n'empêche pas que W (A, B) soit trompeur le selon le troisième cas précédent.

5.3.3.3.Formule (3) de pondération (annuler les poids trompeurs)

Afin d'éviter qu'un poids surestime un chevauchement A∩B qui n'est pas en réalité pertinent, nous proposons le facteur $\alpha \in \{0, 1\}$. α est utilisé pour pénaliser/ annuler ce type de poids:

Tableau 24. Les valeurs prises par le facteur α de chaque chevauchement suivant son GC entre T_i et T_{i+1}

A	$GC_{T_i}(A \cap B)$	$GC_{T_{i+1}}(A \cap B)$
0	0	0
0	>0	0
0	0	>0
1	>0	>0

Pour chaque chevauchement A∩B, α = 1 si la centralité de A∩B ne s'annule jamais entre T_i et T_{i+1}. Sinon α = 0 (Tableau 24). Donc :

$$W(A,B) = |A \cap B| \times \frac{GC_{T_i}(A \cap B) + GC_{T_j}(A \cap B)}{2} \times \alpha, \quad (37)$$

$$j = i+1,\ i = 1...t-1,\ \alpha \in \{0, 1\}$$

Néanmoins, le facteur α prend des valeurs discrètes qui ne sont pas en mesure de capter le changement de centralité GC de A∩B à T_i vers T_{i+1}

5.3.3.4.Formule (4) de pondération (pénaliser les poids suivant la stabilité de centralité)

Etant donné deux chevauchements A∩B, C∩D ∈ P (T_i, T_{i+1}) qui peuvent avoir la même moyenne de centralité : par exemple $\frac{GC_{T_i} + GC_{T_{i+1}}}{2} = 5$ ainsi que |A∩B| =|C∩D| et α = 1 pour les deux. Ce qui fait que W (A, B) = W (C, D). Cependant, l'un des chevauchements peut être moins pertinent que l'autre. Par exemple, on peut tomber sur le cas où $GC_{T_i}(A \cap B) = 1 < GC_{T_{i+1}}(A \cap B) = 9$, ce qui veut dire que la centralité (GC) de A∩B change dramatiquement entre T_i et T_{i+1}. D'autre part, le GC de C∩D reste plus stable : $GC_{T_i}(C \cap D) = GC_{T_{i+1}}(C \cap D) = 5$. Par conséquent, W (C, D) peut être plus favorisé que W (A, B) en ajoutant le paramètre de stabilité de centralité dans la fonction de pondération, mais comment ?

Nous proposons un nouveau facteur β ∈ [0, 1] qui se base sur l'indice d'amplitude de centralité (CA) qui capte la perturbation de GC entre deux point de temps successifs. Le facteur β sert à pénaliser chaque poids une fois que le chevauchement impliqué change sa centralité (GC) de T_i à T_{i+1}

$$\beta = \frac{\alpha}{\left|GC_{T_i}(A \cap B) - GC_{T_j}(A \cap B)\right| + 1}, \quad (38)$$

$$\alpha \in \{0, 1\},\ j = i+1,\ i = 1...t-1$$

En effet, pour chaque chevauchement A∩B, β dépend d'abord de la valeur prise par α. D'après la dernière équation ainsi que la **Figure 51**, le β de A∩B atteint son max = 1 si et seulement si :

- Le facteur correspondant α.= 1 ce qui assure que le GC de A∩B ne s'annule jamais entre T_i et T_{i+1}
- Le CA de A∩B est nul ($\left|GC_{T_i}(A \cap B) - GC_{T_{i+1}}(A \cap B)\right| = 0$), (**Figure 51**) ce qui signifie effectivement que le GC de A∩B est parfaitement stable entre T_i et T_{i+1}

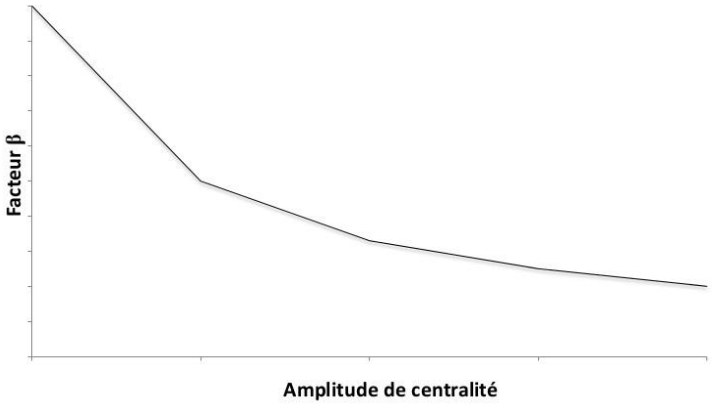


Figure 51. Relation facteur β et amplitude de centralité (CA)

Dans ce cas le poids correspondant, dans sa nouvelle formule, ne sera pénalisé:

$$W(A,B) = |A \cap B| \times \frac{GC_{T_i}(A \cap B) + GC_{T_j}(A \cap B)}{2} \times \beta.\alpha \quad , \qquad (39)$$

$$j = i+1,\ i = 1...t-1,\ \alpha \in \{0, 1\},\ \beta \in [0, 1].$$

D'autre part et d'après la **Figure 51**, plus l'amplitude de centralité est large, plus β tend vers 0. Donc, W (A, B) est plus susceptible d'être pénalisé.

La fonction de pondération W exprime finalement l'intégralité des paramètres (**Figure 52**) tout en mettant une formule bien équilibrée et en mesure d'estimer de manière judicieuse, déterminante la pertinence du chevauchement temporel impliqué dans chaque arc de TW-DAG.

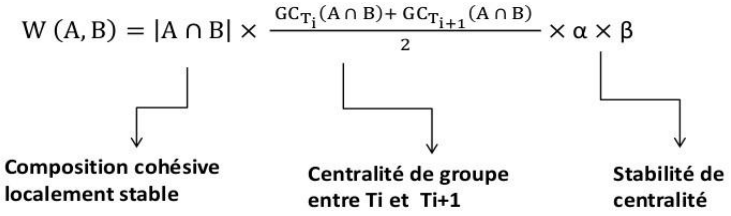


Figure 52. Composantes de la fonction de pondération, chacune exprime un paramètre (contrainte) dérivé de la phase de caractérisation

L'algorithme suivant résume comment W a été progressivement améliorée étape par étape :

Algorithme 3. Algorithme montrant comment la fonction de pondération est améliorée depuis la formule (1) jusqu'à la formule (4)

```
POUR i = 1...t-1 FAIRE
    POUR chaque chevauchement A∩B ∈ P (Ti,Ti+1) FAIRE
        SI A∩B = ø ALORS
            W (A,B) = 0
        SINON
            W (A,B) est calculé suivant la formule (1) et amélioré par la formule (2)

            SI W (A,B) = 0 ALORS
            –   // GC.Ti (A∩B) = 0 et  GC.Ti+1 (A∩B) = 0
            –   // Donc A∩B = A = B (isolé)
                α = 0.

            SINON
            –   // Ajuster W(A,B) en ajoutant plus de pramètres
                SI (GC.Ti (A∩B) ≠ 0) & (GC.Ti+1 (A∩B) ≠ 0)  ALORS
                –   // Assurer que le GC de A∩B ne s'annule pas entre Ti et Ti+1
                    α = 1

                    W (A,B) est calculé suivant la formule (3)

                –   // Ajouter le pramètre de stabilité de centralité

                    β de A∩B de se calcule selon la formule correpsodante

                    W (A,B) est calculé suivant la formule (4)
                SINON
                    α = 0
                    β = 0
                    W (A,B) = 0
```

D'après cette phase de modélisation (**Figure 49**), les sommets, les arcs et les poids présentent respectivement des groupes, des chevauchements temporels successifs et les paramètres. En d'autres termes, TW-DAG présente un processus évolutionnaire (patterns d'évolution) de groupes et des structures sous-jacentes d'un SN, étape par étape dans le temps. Un arc lourd W (A, B) dans le méta-modèle TW-DAG implique un grand chevauchement A∩B (composition localement stable), remarquablement central, avec α = 1 et β tend vers 1 (GC qui penche vers la stabilité). C'est-à-dire, une structure sous-jacente pertinente entre T_i et T_{i+1} (**Figure 53**).

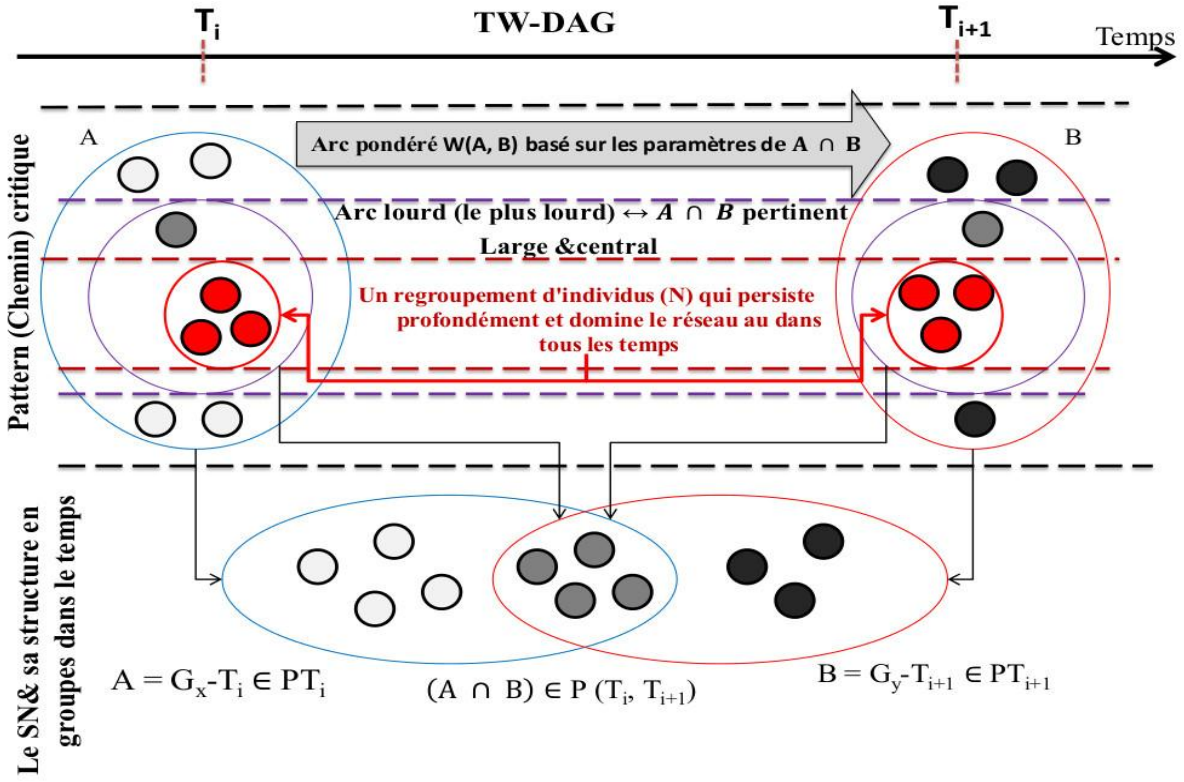


Figure 53. Aperçu sur le TW-DAG représenté dans par une architecture en couches. Il est formé par des arcs pondérés reliant des groupes (sommets) de partitions successives qui se chevauchent dans le temps. Les arcs lourds impliquent des chevauchements temporels pertinents (sous-groupe en gris) entre T_i et T_{i+1}. Le pattern critique est le chemin le plus lourd qui couvre un regroupement persistant (en rouge) à l'intérieur tout au long de la période d'observation,

6. Approche d'identification d'une identité noyau basée sur la recherche des patterns critiques dans le TW-DAG

À partir des deux phases précédentes : Caractérisation et de modélisation, un chevauchement temporel affiché par le TW-DAG comme pertinent, couvre ou représente en fait une trace d'une identité noyau dans un intervalle de temps (entre T_i et T_{i+1}). Mais, quelle structure va-t-elle incarner cette identité pertinente tout au long de la période d'observation?

Pour répondre à cette question, on met l'accent sur deux volets, la pertinence et la couverture de la durée d'observation. Si un arc lourd dans TW-DAG reflète un chevauchement temporel pertinent localement entre T_i et T_{i+1} (**Figure 53**), on s'intéresse ainsi à explorer les chemins formés par des arcs lourds. À cet égard, nous faisons appel à une méthode appelée CPM/CPA ('Critical Path Method'/ 'Critical Path Analysis') par laquelle se dessine le schéma de la phase d'identification (**Figure 49**). C'est une méthode qui fait partie des techniques de gestion et planification des activités de projets ((Hazar 2014)), mais utilisée aussi pour l'analyse des graphes acycliques ((Batagelj & Mrvar 2012)).

- L'avantage principal d'utiliser CPM/CPA dans un processus de planification consiste à aider le planificateur à développer, tester son plan et assurer sa robustesse.
- La méthode aide aussi à identifier la durée (minimale) nécessaire pour achever un projet.
- Elle aide à identifier quelles sont les étapes/ activités du projet qu'on doit accélérer pour terminer le projet dans le temps donné.

Formellement, CPM cherche des chemins critiques ((Baker 2013)) ((Armstrong-Wright & Mice 1969)) ((Kelley 1961)) en s'appliquant sur des réseaux complexes qui schématisent les processus composés de séquences d'activités (Pour commencer certaines activités, il faut que certaines d'autres soient achevées avant) : Diagramme d'activités avec des dépendances dans le temps ((Baker 2013)) ((Armstrong-Wright & Mice 1969)).

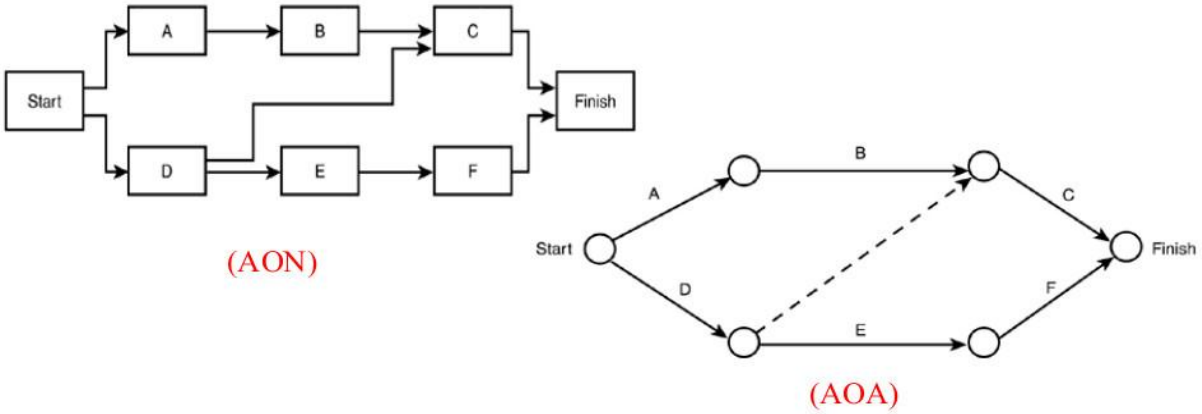


Figure 54. Exemple de diagramme d'activités en deux représentations : 'Activity-on-arrow (AOA) diagram' et 'Activity-on-node (AON) diagram' ((Francis 2009))

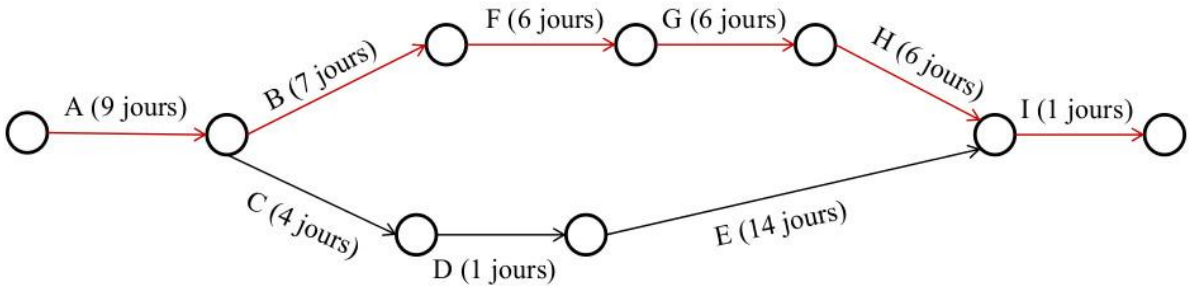


Figure 55. Exemple de 'Critical path' (chemin en rouge) dans un diagramme d'activités 'AOA' (Avec un coût, une durée) totale de 35 jours

Le tableau suivant propose d'abord une analogie entre la structure de TW-DAG et celle d'un diagramme d'activité, et prouve ainsi l'applicabilité de l'algorithme CPM sur TW-DAG.

Tableau 25. Analogie entre le méta-modèle TW-DAG et le diagramme d'activités, et l'applicabilité de l'algorithme CPM sur TW-DAG

Le méta-modèle TW-DAG	Diagramme d'activités
Pattern d'évolution, des chemins, des séquences de groupes dans le temps	Processus (projet) de séquence d'activités ordonnées dans le temps
Graphe acyclique temporel	Diagramme d'activités se présente sous forme

	de graphes acycliques temporels
Les sommets sont des groupes d'une partition PT_i à chaque' time point' T_i	Deux schémas de représentations (Réseau/ (diagramme d'activités) possibles : 'Activity-on-arrow (AOA) diagram' : L'activité se présente sur l'arc. D'où, il y a des jalons, des étapes avec des 'Endpoints' logiques (points terminaux des activités), (**Figure 54**)
	'Activity-on-node (AON) diagram' : Les nœuds représentent les activités elles-mêmes (**Figure 54**)
Les groupes d'une même partition sont en parallèle mais sans aucune relation entre eux	Il peut y avoir des activités en parallèle
Des arcs reliant des groupes qui se chevauchent successivement dans le temps. Les arcs pondérés par une combinaison de paramètres axés sur ces chevauchements. Pas de notion de temps mort	Dépendances entre les activités, étiquetées par La durée (le temps qu'il faut pour terminer une activité) et éventuellement le temps mort: « Si l'activité 'B' concerne par exemple le teste d'un panneau solaire qui nécessite le 'lever de soleil', il y aura une contrainte d'ordonnancement sur l'activité. Elle ne se lancera jamais avant le temps prévu du 'lever de soleil'. Ce qui peut introduire le temps mort ('total float')/ temps d'attente de cet événement dans l'ordonnancement des activités »
CPM sera utilisé pour chercher le chemin le plus long qui couvre toute la durée d'observation (Couverture) et le plus coûteux (Coût = poids W).	CPM calcule le chemin le plus long/ coûteux pour atteindre un 'Endpoint' donné, parfois jusqu'à la fin du projet (**Figure 55**) incluant les retards/ Total float des activités. CPM peut aussi déterminer la durée la plus courte pour terminer le projet (Coût = durée).
Séquence la plus lourde des groupes (des chevauchements) critiques (pertinents)	Détermine la série d'activités (dépendantes) critiques

Le Tableau 25 et la **Figure 55** prouvent l'applicabilité de CPM sur la structure de TW-DAG, étant un graphe temporel acyclique. ***La recherche du chemin le plus couteux, le plus long se traduit dans TW-DAG par la recherche du chemin (le pattern de groupes) le plus lourd qui couvre les t points de temps.*** Mais l'avantage est plus profond qu'une recherche dans un graphe/ diagramme simple d'activités. Les chemins critiques (Critical path : CP) dans TW-DAG seront plus significatifs. Il s'agit de patterns de groupes qui encapsulent des structures sous-jacentes (chevauchements temporels et autres) et affichent l'évolution de leurs paramètres.

Nous suivons trois critères pour qu'un chemin dans TW-DAG soit formellement critique (CP), **(Figure 56)**

- ***La couverture***: CP est une séquence (un chemin) de groupes (A_1, A_2, ..., A_t) tel que $A_i = G_x\text{-}T_i$, (i = 1 ... t). La couverture permettra d'explorer le paramètre de persistance (stabilité de composition) à l'intérieur du CP tout au long de la période d'observation (**Figure 56**).
- ***La persistance*** : Un chemin couvrant ne pourrai pas être critique à moins qu'une structure notée N persiste profondément à l'intérieur (**Figure 56**) : $N \subset A_i \cap A_j, \forall\, i = 1..t-1, j = i+1$. En d'autres termes, une structure/ un regroupement (un sous-groupe) N montré en rouge dans la **Figure 53**, doit être encapsulé dans tous les chevauchements temporels successifs $A_i \cap A_j$ couverts par CP.
- ***La lourdeur***: Le poids de CP qui sera évalué par la somme $\sum_{j=i+1}^{i=1..t-1} \sum W(Ai, Aj)$ doit être le plus lourd (**Figure 56**). Ce critère laissera CP susceptible de couvrir des arcs lourds W (A_i, A_j) qui impliquent automatiquement des chevauchements pertinents ($A_i \cap A_j$).

C'est-à-dire, une fois qu'un regroupement sous-jacent N persiste à l'intérieur d'un CP, il est susceptible d'afficher une pertinence similaire à celle des chevauchements $A_i \cap A_j$ (à valider) mais plus généralisée sur les t points de temps (**Figure 56**). Il est susceptible de se présenter comme le regroupement équilibré le plus possible: ayant une grande composition stable qui joue un rôle central et le plus stable possible durant toute la période d'observation (**Figure 56**). Ce qui signifie la durabilité et la dominance de l'identité caractérisée d'un noyau.

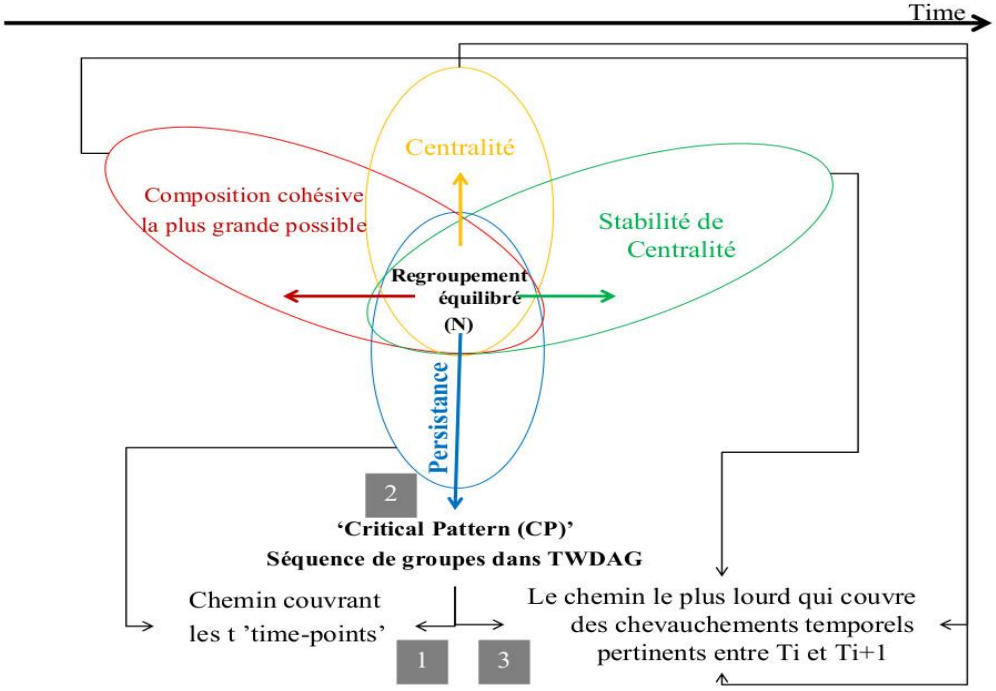


Figure 56. Les trois critères pour qu'un chemin dans TW-DAG soit un pattern critique (CP)

Ces critères de repérage ont été formalisés mais nécessiterons des épreuves de validation (à vérifier). Le chemin détecté comme CP sera le conteneur qui encapsule un regroupement ayant un comportement typique d'une structure noyau. CP est une sorte d'infrastructure de la colonne vertébrale du SN dynamique, là où on peut identifier profondément la durabilité et la dominance qui caractérisent une identité significative plus raffinée d'un noyau sous-jacent dans un SN évoluant dans le temps.

Suivant ce processus d'identification, la **Figure 53** montre comment le méta-modèle TW-DAG (d'un processus évolutionnaire) s'organise verticalement sous une architecture en couches. La verticalité elle est montré dans la **Figure 53** (et même dans la **Figure 49**) indique des couches ou des strates. Chacune représente horizontalement un pattern d'évolution dans le temps à un niveau donné. On est parti d'un ensemble d'acteurs en interactions dans un SN, en fouillant dans des patterns de groupes (CP), de chevauchements temporels vers une identité d'une structure noyau. Le tout est en évolution dans le temps.

7. Données, expérimentations et résultats

Nous proposons en premier plan des épreuves préliminaires (**Figure 57**) sur différent données de SNs/OSNs afin de bien choisir celui le plus adapté pour expérimenter l'approche proposée.

7.1.Choix de datasets et choix techniques

Tout d'abord il faudra mettre en évidence la nature, la pertinence ainsi que la richesse (données temporelles) des données sociales elles-mêmes à tester, étant un principe général de notre travail de recherche et contributions comme c'est déjà montré dans la **Figure 1.**

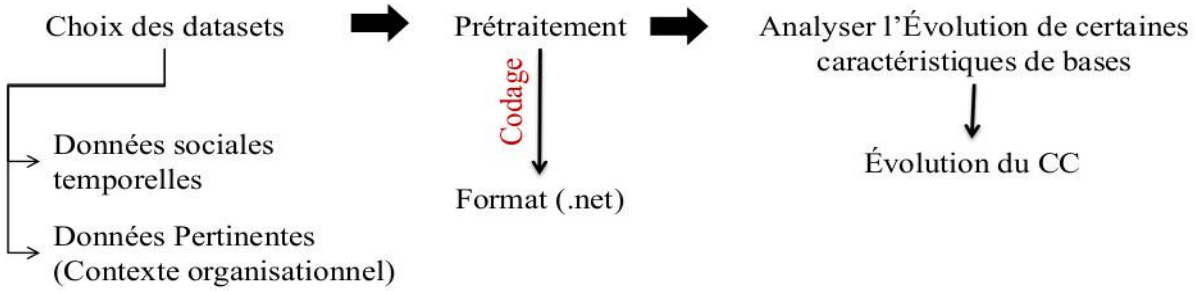


Figure 57. Choix de datasets suivant des étapes préliminaires

Par ailleurs, et dans une phase de prétraitement (**Figure 57**), nous utilisons le format .net pour coder les données sociales temporels. C'est l'un des formats puissants qui permet d'exprimer et supporter explicitement l'aspect temporel ((Beauguitte 2011)). Il nous permet de coder avec souplesse des informations considérées comme fondamentales : Durée de vie des liens et mêmes des acteurs, (dans la sous-section : Données sociales temporelles, une simple abstraction de la dynamicité temporelle) après les avoir extrait. La **Figure 58** montre le format de ces données temporelles codées dans un fichier .net. On ne met que les rubriques nécessaires notamment pour coder la période d'observation en t'time-points' et les intervalles de temps décrivant la durée de vie. L'étiquette du sommet comme d'autres rubriques (qui ne sont pas mentionnés ici) sont optionnelles.

```
*Vertices Nombre_de_sommets
1 "étiquette_de_sommet_1" ['interval_de_sa_durée_de_vie]
2 "étiquette_de_sommet_2" ['interval_de_sa_durée_de_vie]
…

*Arcs (le cas d'un graphe orienté)
numéro_de_sommet_source  numéro_de_sommet_destination  valeur_du_poid  ['interval_de_la durée_de_vie_du_lien]
…

*Edges (le cas d'un graphe non-orenté)
numéro_de_la_première_extrémité  numéro_de_la_deuxième_extrémité  valeur_du_poid  ['interval_de_la durée_de_vie_du_lien]
…
```

Figure 58. Le codage des données sociales temporelles selon le format .net

En utilisant un outil puissant comme Pajek ((Batagelj & Mrvar 2012)) ((Batagelj & Mrvar 2003a)) ((Batagelj & Mrvar 1998)) ((Beauguitte 2011)) ((Batagelj & Mrvar 2006)) (Version 3.08), le format .net est aussi bien compatible qu'adapté pour calculer quelques indices de

SNA qu'on applique dans certaines sous-étapes de notre approche. Dans ce sens, il y a aussi des exigences qui nous obligeront d'analyser l'évolution de certaines caractéristiques de base (**Figure 57**) et tirer des conclusions préliminaires (ex. sur la tendance des acteurs à se regrouper dans le temps) avant d'entamer la phase de modélisation et ensuite l'indentification.
Par conséquent, plusieurs datasets vont éventuellement subir un prétraitement et seront testés afin d'arriver à choisir le jeu de donné le plus adapté, le plus informatif qui permet de multiplier le bénéfice informationnel de cette approche.

7.1.1. Un échantillon de 'Marvel Universe Social Graph'

L'univers Marvel des bandes dessinées est un exemple de source de réseaux de collaboration sociale entre des personnages (Spider-Man, Thing, Beast, Captain America, Namor, Hulk, Ironman, etc.) dans un monde artificiel. Dans un tel univers des évènements du monde réel sont mélangés avec la science-fiction et la fantaisie ((Gleiser 2007)). Le dataset 'Marvel Universe Social Graph' complet a été collecté par Cesc Rosselló, Ricardo Alberich, et Joe Miro et contient 6486 personnages dans 12942 livres (Alberich et al 2002). Deux personnages sont liés, étant donné qu'ils apparaissent les deux dans le même livre de bande dessiné. (Alberich et al 2002) voulaient savoir si ce type de réseau formé par des nœuds qui sont des entités inventées, interconnectées par des liens crées en réalité par une équipe de rédacteur ((Gleiser 2007)), ressemble à des SNs réels ou des graphes aléatoires. Les auteurs ont montré que les caractéristiques sont comparables (Alberich et al 2002) ((Gleiser 2007)) avec celles des réseaux de collaboration dans le monde réel, comme le réseau de Hollywood, ou des réseaux de collaborations scientifiques (co-auteurs), etc. Même s'il ressemble plus à un SN réel ((Gleiser 2007)) qu'on pourrait le croire, le graphe de collaboration entre les personnages des bandes dessinées 'Marvel Comics' peut aussi être considéré comme des données synthétiques (un réseau artificiel) par rapport au sujet étudié.
Nous partons d'un échantillon de 'Marvel Univers Social Graph' (MUSG) qui évolue pendant une période partagée en 4 intervalles de temps donnés. Ce dataset incluant les données temporelles est déjà trouvé préparé sous le format .net compatible avec Pajek sur[117] (**Figure 59**)

```
*Vertices 165                    *Arcs
1 *SPIDER-MAN* [1-4]             *Edges
2 *WATSON-PARKER* [1-4]          1 2 1 [1-4]
3 *HUMANTORCH* [1-4]             3 4 1 [1-4]
4 *THING* [1-4]                  5 4 1 [1-4]
5 *MR.FANTASTIC* [1-4]           3 5 1 [1-4]
6 *INVISIBLEWOMAN* [1-4]         6 5 1 [1-4]
7 *JAMESON* [1-4]                3 6 1 [1-4]
....                             ....
140 *RORY* [2-4]                 80 141 1 [2-4]
141 *PETROVITCH* [2-4]           8 1 1 [2-4]
142 *RINTRAH* [3-4]              112 69 1 [2-4]
143 *GORGON[INHUMAN]* [3-4]      113 69 1 [3-4]
144 *KARNAK[INHUMAN]* [3-4]      16 142 1 [3-4]
....                             ....
163 *MS.MARVELII* [4-4]          165 144 1 [4-4]
164 *DELAFONTAINE* [4-4]         137 28 1 [4-4]
165 *BLACKBOLT* [4-4]            132 138 1 [4-4]
                                 80 137 1 [4-4]
```

Figure 59. Aperçu sur un échantillon du SN MUSG codé dans le format .net

Tout d'abord, nous étudions la distribution de degré dans le cas statique, là où les données temporelles ne sont prises en compte, afin de vérifier comment ce graphe présente une structure porche d'un SN réel. Pajek ne peut calculer que le vecteur des degrés, une valeur pour chaque nœud en utilisant la commande : 'Network Menu > Create Vector > Centrality >

Degree > All'. Ce qui nous oblige de faire un calcul supplémentaire basé sur ce vecteur. Il s'agit de calculer la fréquence de chaque valeur de degré: C'est le nombre des nœuds ayant la même centralité de degré (Figure 7 à gauche). Il est remarqué que le nombre d'interactions (la centralité de degré) est inversement proportionnel à sa fréquence (le degré augmente et la fréquence tend à diminuer). La distribution (comme c'est déjà vu dans la Figure 7 à gauche) suit la loi de puissance avec un seuil exponentiel qui révèle selon ((Gleiser 2007)) que le graphe n'est pas aléatoire (un graphe social). C'est aussi un réseau dominé par quelques acteurs seulement ayant beaucoup d'interactions que les autres.

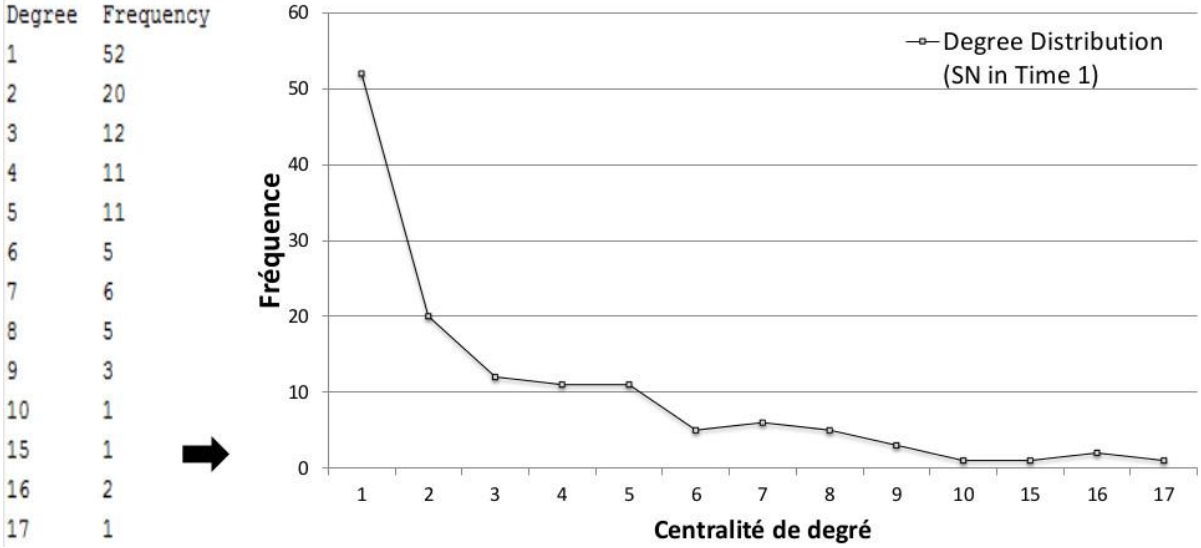


Figure 60. Distribution de degré des nœuds qui forment le premier snapshot du SN artificiel, échantillonné de MUSG

Mais la distribution des degrés statiques de ces nœuds ne donne pas une justification réaliste tant que le SN change sa structure dans le temps. D'où on applique la même procédure de calcule mais seulement sur le premier snapshot du SN (**Figure 60**). C'est un snapshot d'interactions et d'acteurs aussi car selon la **Figure 59** l'ensemble des nœuds n'est pas consistant. La **Figure 60** révèle que le SN se comporte de la même manière dès le début (dans le premier time-point) même si le nombre d'acteurs n'est pas le même.

Tableau 26. L'évolution des caractéristiques de base du SN échantillonné depuis MUSG

Time-points / caractéristique	1	2	3	4
Nombre des nœuds	130	141	148	165
Nombre de liens	220	240	260	300
Densité	0,02623733	0,02431611	0,02390145	0,02217295
Degré moyen	3,38461538	3,40425532	3,51351351	3,63636364
CC global	0,76732535	0,73014809	0,79056627	0,77718780

Dans le Tableau 26, on affiche les valeurs de certaines caractéristiques calculées sur la séquence de 4 snapshots de ce SN au fil du temps, en utilisant la commande 'Network Menu > Info > General'. Le même calcule est refait (Tableau 27) en supposant que l'ensemble de nœuds est consistant (SN-c).

Tableau 27. L'évolution de certaines caractéristiques de base du SN échantillonné depuis MUSG, avec un ensemble de nœuds consistant

Time-points / caractéristique	1	2	3	4
Nombre des nœuds	165	165	165	165
Nombre de liens	220	240	260	300

Densité	0,01626016	0,01773836	0,01921656	0,02217295
Degré moyen	2,66666667	2,90909091	3,15151515	3.63636364
CC global	0,76732535	0,73014809	0,79056627	0,77718780

D'où on crée et on analyse les courbes d'évolution de certaines de ces caractéristiques en comparant les deux cas :

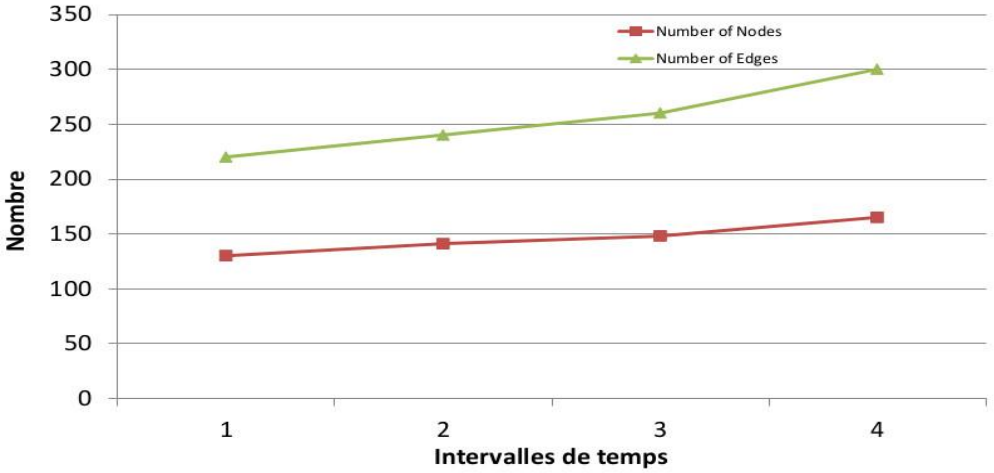


Figure 61. L'évolution du nombre des liens et des nœuds du SN échantillonné de 'MUSG' au fil du temps

À partir du Tableau 26, il est remarqué que le nombre de liens ainsi que le nombre de nœuds progressent presque en parallèle dans le temps (**Figure 61**), mais ça ne veut pas dire que le SN tend à se densifier. La **Figure 62** montre plutôt une chute dans la densité du SN (courbe en rouge). Cela révèle un déséquilibre entre l'évolution du nombre d'acteurs (n) et leurs interactions (m) (collaborations) d'après la définition de la densité dans le Tableau 4 (la fraction m / $|E^*|$). En effet, la croissance de n à chaque snapshot multiplie le nombre théorique prévus des liens $|E^*|$ = n(n-1)/2, donc creuse l'écart avec m et la densité continue à chuter (courbe en rouge). D'autre part, tant que n ne change pas (le dénominateur n(n-1)/2 est constant) dans le temps et m évolue normalement, le SN-c semble tendre à se densifier (courbe bleue). Au départ, SN-c, est beaucoup plus clairsemé en affichant une faible densité, car les nœuds qui devraient apparaitre dans les temps 2, 3 et 4 sont isolés dans le premier intervalle de temps. Dans le dernier intervalle de temps, les deux courbes se croisent ainsi que toutes les autres caractéristiques (Tableau 26 et Tableau 27). C'est le moment où l'ensemble d'acteurs dans les deux cas deviennent identiques

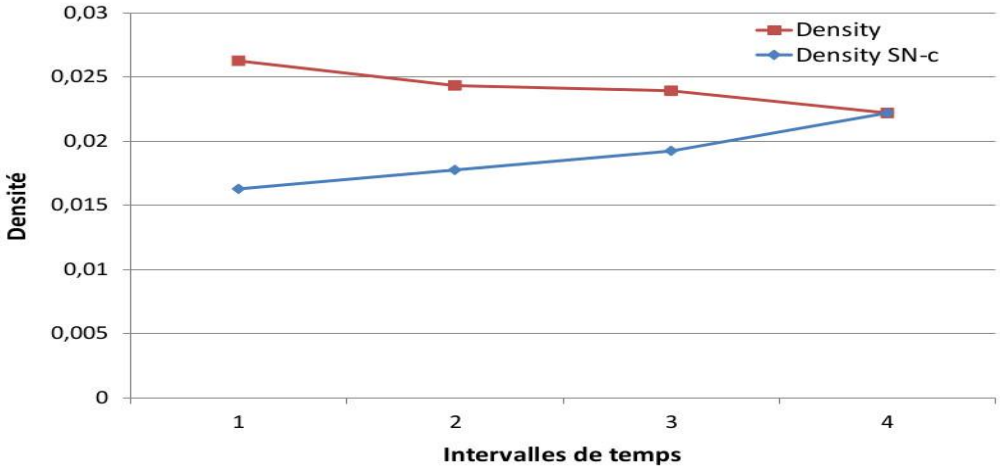


Figure 62. L'évolution de la densité dans le SN échantillonné de MUSG au fil du temps et son homologue consistant SN-c

Cependant, les valeurs du CC global dans le Tableau 26 sont les même que celles affichées dans le Tableau 27. La **Figure 63** montre l'évolution du CC dans le temps. CC ne s'annule pas, même s'il est perturbé, une décroissance dans [1-2] et une croissance dans [2-3], etc., avec des valeurs plus ou moins proches. En d'autres termes, les acteurs ont une tendance variante à former des groupes dans le temps. Mais ce comportement n'est pas influencé lorsque l'ensemble de nœuds est supposée consistant. En effet, seulement les acteurs qui apparaissent normalement à T = 1 sont en interactions. Les autres sont inactives et n'ont pas d'influence sur le CC du SN consistant, jusqu'à ce qu'ils rejoignent vraiment le réseau à T+1, T+2, etc.

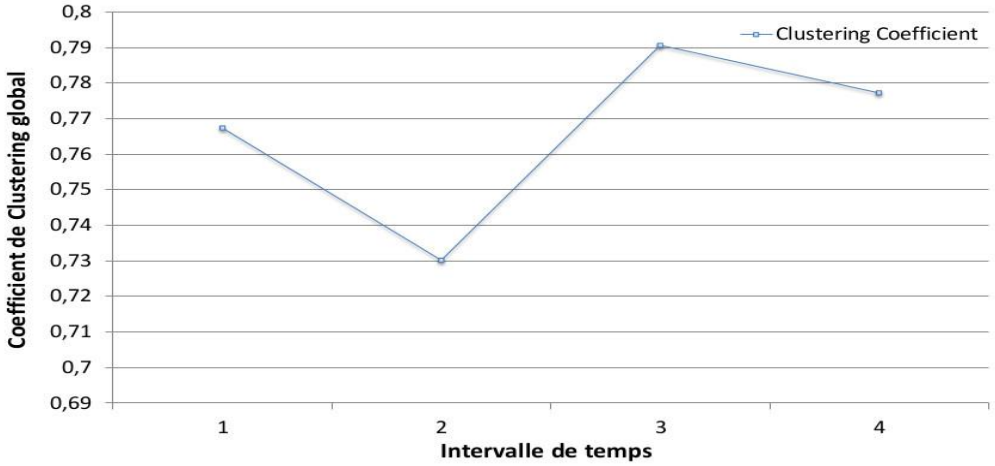


Figure 63. L'évolution du CC global du SN échantillonné de MUSG au fil du temps

7.1.2. Le dataset 'The Facebook-like Social Network'

Ce dataset est tout d'abord cité déjà dans le Tableau 17 dans la catégorie des OSNs et avant cela a été décrit dans ((Panzarasa & Opsahl 2009)), et utilisé dans les travaux de ((Opsahl et al 2008)) ((Opsahl & Panzarasa 2009)). C'est un réseau qui provient d'une communauté en ligne sur Facebook, formée par un ensemble d'utilisateurs qui sont des étudiants de l'université Irvine de la Californie ((Opsahl et al 2008)) ((Opsahl & Panzarasa 2009)). Le réseau 'The Facebook-like Social Network' (Fb-lSN) semble plus pertinent par rapport à d'autres OSNs (connaissance éphémères). Cette pertinence est acquise du fait que le réseau est constitué par un ensemble d'individus appartenant à un environnement bien connu et aussi limité géographiquement.

Le SN comprend 1899 utilisateurs (users), chacun envoie ou reçoit au moins un message. Au total, il y a 59835 messages en ligne envoyés sur 20296 liens orientés entre users. Noter bien que le dataset original inclut aussi des attributs nodaux (sexe, âge, cours assistés, etc.) mais qui ne sont pas disponibles devant les procédures d'anonymisation d'users. C'est l'une des problématiques discutées dans la partie (Pertinence et richesse des données sociales (des SNs\ OSNs)). Bien heureusement, le dataset inclut des données temporelles qui sont accessibles en mode pondéré (poids des liens est l'équivalent du nombre des messages échangés) ou binaire.

Mise à part le format UCINET sous lequel des données pondérées statiques sont organisées, les données temporelles sont simplement trouvées dans des fichiers textes (.txt) (**Figure 64**)

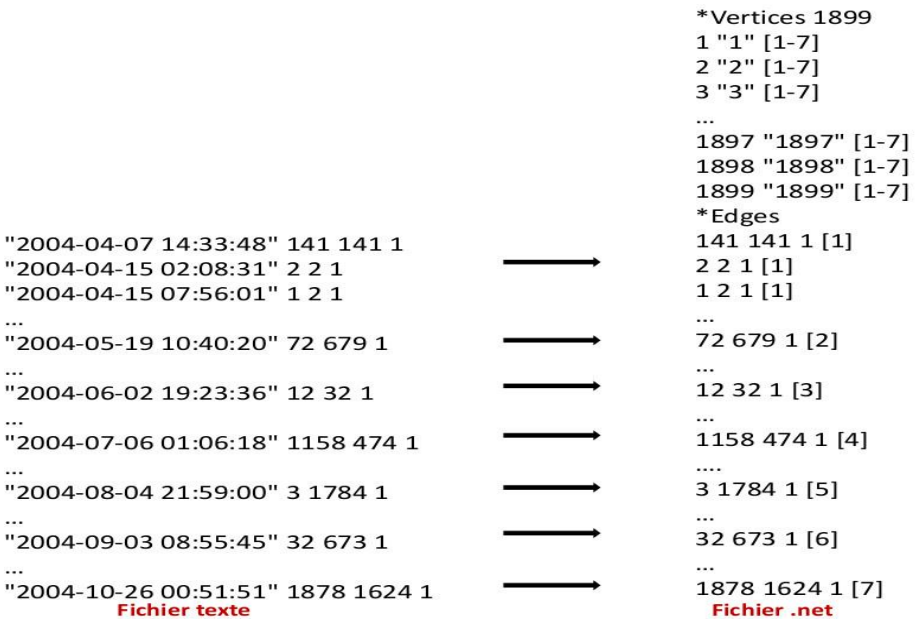
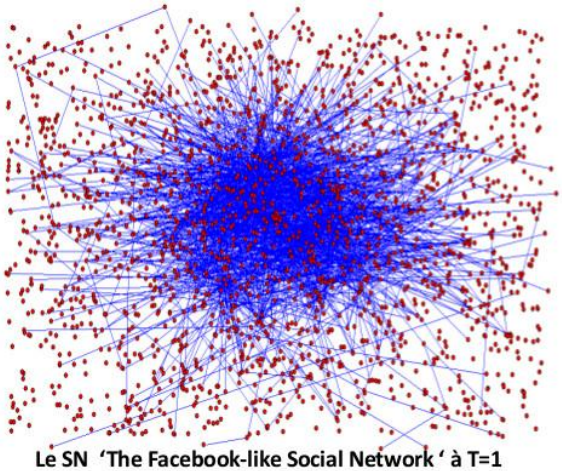


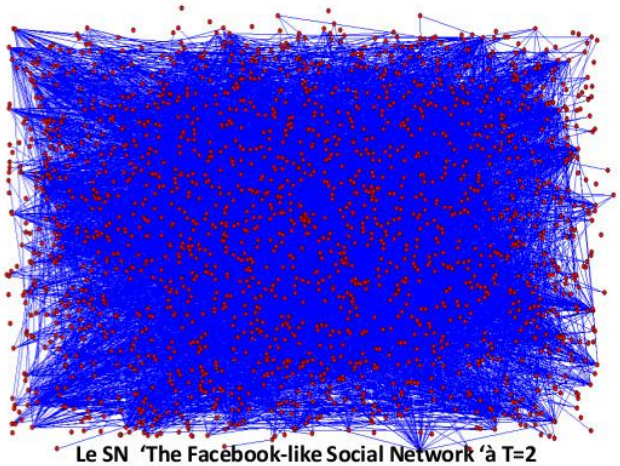
Figure 64. Données temporelles binaires du dataset 'The Facebook-like Social Network' en .txt converties au format .net

On applique un prétraitement sur le ficher texte à gauche, là où on trouve la date et l'heure de chaque message envoyé entre deux users, pour le convertir au format .net (**Figure 64** à droite) :

- D'abord, il faut définir la résolution des fenêtres de temps. Il est remarqué que les users interagissent pendant 7 mois entre le mois d'avril et octobre de la même année 2004. Nous considérons chaque mois comme un 'time-step', ce qui donne 7 time-steps.
- Avec un petit script Java, on génère d'abord la liste des 1899 sommets dont l'ensemble est supposé consistant dans le temps [1-7] (**Figure 64**).
- Pour chaque entrée (un message envoyé) dans le fichier .txt, un lien est créé dans la liste 'Edges' en mettant les identificateurs des users impliqués. La date et l'heure sont remplacées par un time-step correspondant (Avril:[1], Mai:[2],…, Octobre:[7]) en utilisant les expressions régulières. Il est possible d'avoir des relations multiples et des boucles (loops), par exemple : 2 2 [1] (**Figure 64**).



Le SN 'The Facebook-like Social Network ' à T=1
(Mois d'Avril)



Le SN 'The Facebook-like Social Network 'à T=2
(Mois de Mai)

15. Network N14 in time 1 (1899)

Number of vertices (n): 1899

	Arcs	Edges
Total number of lines	0	5862
Number of loops	0	665
Number of multiple lines	0	3448

Density [loops allowed] = 0.00306666
Average Degree = 6.17377567

The highest values of lines:

Rank	Line	Value	Line-Id
1	2-2	1.00000	2-2

16. Network N14 in time 2 (1899)

Number of vertices (n): 1899

	Arcs	Edges
Total number of lines	0	38644
Number of loops	0	928
Number of multiple lines	0	28703

Density [loops allowed] = 0.02117464
Average Degree = 40.69931543

The highest values of lines:

Rank	Line	Value	Line-Id
1	67-88	1.00000	67-88

Figure 65. Le SN Fb-lSN à T= 1 et à T=2 visualisé par Pajek, avec certaines statistiques

La **Figure 65** donne une visualisation graphique des deux premiers snapshots de ce SN, une visualisation d'une forme rectangulaire obtenue à travers 'Layout > Energy > Kamada kawai > Free' sur 'Draw Window Tools' de Pajek. Dans la partie inférieure, on trouve quelques statistiques. Comme les données sont binaires, les poids de toutes les relations qui ont eu lieu sont identiques, ce qui explique que la valeur la plus élevée d'un lien est affiché à 1 (le lien en question est sélectionné au hasard). En considérant les 'loops' et les liens multiples, le réseau à T1 est visiblement moins dense qu'à T2. La **Figure 66** confirme que le SN est inhabituellement de moins en moins dense dans les snapshots qui se succèdent dans le temps.

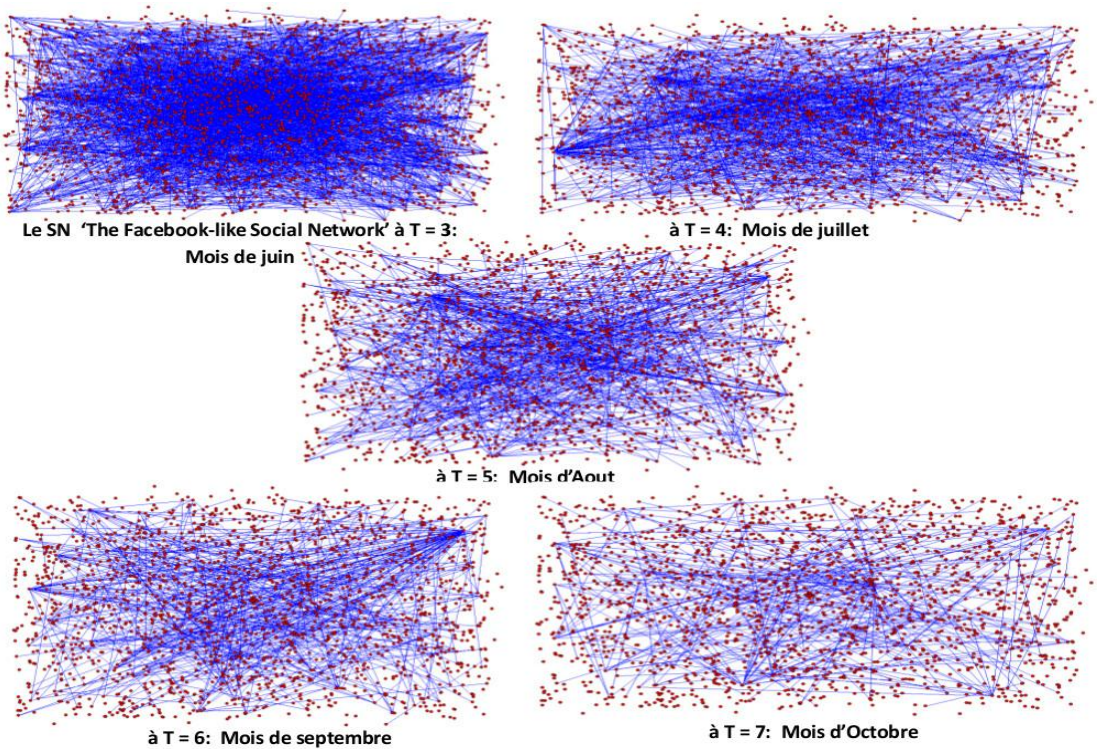


Figure 66. Le SN Fb-lSN entre T= 3 et T=7 (de Juin à Octobre) visualisé par Pajek

Devant ce constat, on étudie l'évolution des caractéristiques de base de ce SN, tout en omettant les loops et les relations multiples à l'aide de Pajek. En effet, on essaye d'ignorer toute impureté qui peut brouiller l'étude de la connectivité du SN dans le temps, pour éviter les estimations trompeuses qui seront décisives pour juger un réseau s'il est adapté ou pas à notre modélisation (par le méta-modèle TW-DAG). Par conséquent, la **Figure 67** montre la différence de statistiques par rapport à la première version du réseau à T=1 dans la **Figure 65.** Le nouveau nombre de liens = 1749 = Ancien nombre de liens (5862) – (nombre de loops (665)+ nombre de relations multiples (3448)). La densité et le degré moyen diminuent également

```
-------------------------------------------------------------------------------
58. Network N57 in time 1 (1899)
-------------------------------------------------------------------------------
Number of vertices (n): 1899
-----------------------------------------------------------
                                      Arcs          Edges
-----------------------------------------------------------
Total number of lines                    0           1749
-----------------------------------------------------------
Number of loops                          0              0
Number of multiple lines                 0              0
-----------------------------------------------------------

Density1 [loops allowed] = 0.00097000
Density2 [no loops allowed] = 0.00097051
Average Degree = 1.84202212

The highest values of lines:

      Rank          Line                      Value    Line-Id
-------------------------------------------------------------
         1           3-4                    1.00000    3-4
```

Figure 67. Statistiques du SN Fb-lSN à T= 1 sans loops et relations multiples (Pajek)

Le Tableau 28 affiche les principales caractéristiques de cette nouvelle version de réseau, calculées pendant les 7 time-steps en commençant à partir de son diamètre. Pour chaque snapshot, le diamètre est estimé par 'Network Menu > Create New Network > SubNetwork with Paths > Info on Diameter' sur Pajek.

Tableau 28. L'évolution de certaines caractéristiques du SN Fb-lSN sans loops et sans relations multiples

Time-step / Caractéris-tiques	T=1 (Avril)	T=2 (Mai)	T=3 (Juin)	T=4 (Juillet)	T=5 (Aout)	T=6 (Septembre)	T=7 (Octobre)
Diamètre	7	6	9	10	10	12	12
Les 2 acteurs impliqués	5 - 514	25–1530	147 – 352	679 – 1652	1802–1815	1038 - 1220	415 – 847
Nombre de liens	1749	8598	1841	678	480	313	179
Densité	0.0009 7051	0.004770 97	0.0010215 6	0.00037622	0.0002663 5	0.00017368	0.0000993 3
CC global (moyenne)	0.0963 1099	0.126940 67	0.0417713 2	0.05512958	0.0483596 8	0.02551893	0.0000000 0
CC global Watts & Strogatz	0.0471 0196	0.056071 23	0.0156487 5	0.01480263	0.0148215 9	0.00680658	0.0000000 0

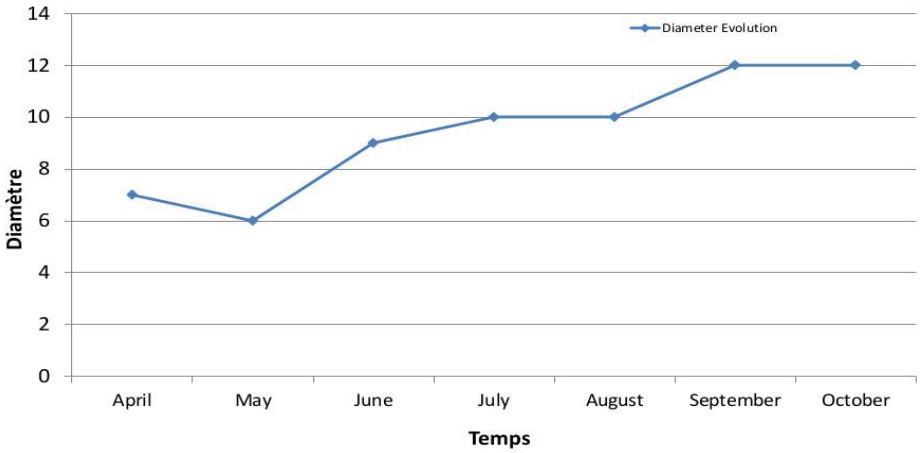


Figure 68. Évolution de diamètre du SN Fb-lSN sans loops et relations multiples

Comme il est affiché dans la **Figure 68**, la géodésique la plus longue (le diamètre) du SN change sa longueur et donne une vue grossière sur la proximité des individus pendant la période d'observation. Entre Avril et Mai le diamètre se rétrécit de 7 à 6, influencé remarquablement par l'augmentation du nombre de liens (Tableau 28) et donc par une croissance dans la densité du SN (**Figure 69**). Par la suite, le diamètre s'étend de plus en plus alors que le SN devient considérablement moins dense (le taux d'interactions diminue).

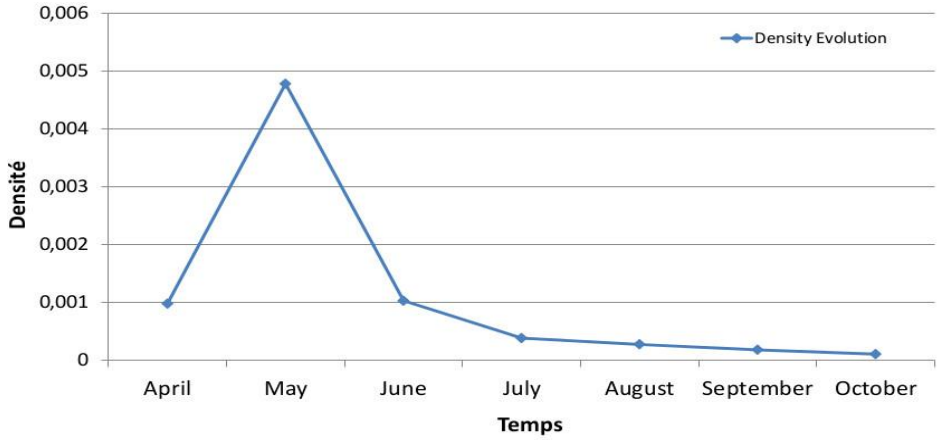


Figure 69. Évolution de la densité du SN Fb-ISN sans loops et relations multiples

En parallèle, la **Figure 70** montre l'évolution des valeurs de CC global du SN, affichées en deux versions (Tableau 28). La première est une moyenne des coefficients de clustering individuels (calculés par 'Network Menu > Create Vector > Clustering coefficients') et la deuxième est un CC généralisé de Watts et Strogatz (basé sur le concept de transitivité). Pendant que le SN se densifie entre Avril et Mai, les acteurs du SN ont une tendance croissante à se regrouper. C'est un dans un premier temps un bon signe en terme de connectivité par rapport à notre objectif de recherche. Cependant, l'effet marquant de la période qui vient après le mois de Mai est la chute du CC (avec les 2 versions) jusqu'à ce qu'il s'annule, accompagnée bien évidemment par la décroissance de la densité (**Figure 70** et **Figure 69**). Le réseau est plus clairsemé et les individus ont l'air de former plutôt une structure modulaire. Cela se confirme par les deux versions de CC tant qu'elles ont le même comportement (**Figure 70**) et ***donne un mauvais signe sur la condensabilité de cet ensemble de données pendant toute la durée d'observation par rapport à la phase de modélisation***.

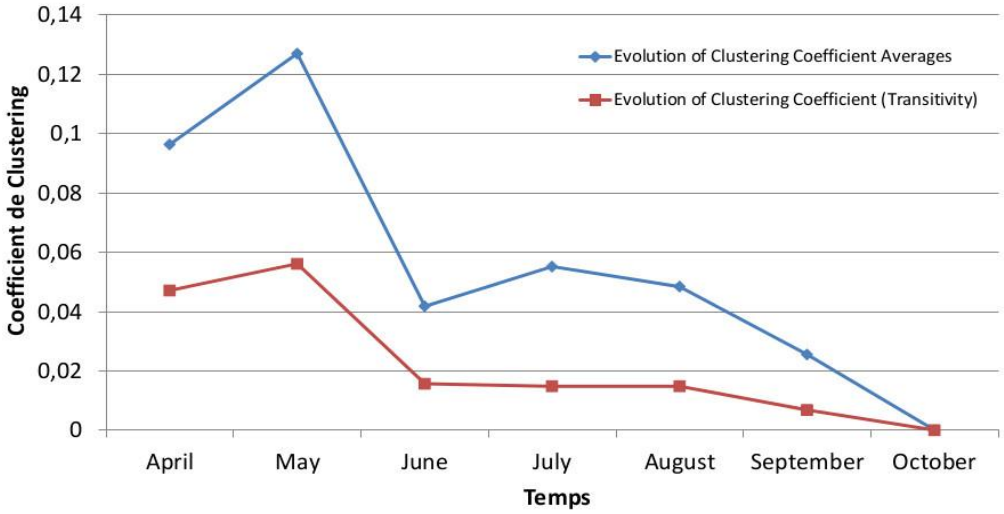


Figure 70. Évolution de CC global (en 2 versions) du SN Fb-ISN sans loops et relations multiples

7.1.3. Echantillon de dataset 'Enron email network'

Le 'Enron email network' (EEN) ((Leskovec et al 2009)) ((Klimmt & Yang 2004)) est l'un des meilleurs exemples d'OSNs/SNs pertinents convenables à étudier l'évolution du comportement des SNs dans le temps. On a vu par exemple que ce jeu de données a été abordé par (Tang et al 2010b) en étudiant les métriques de centralités individuelles temporelles devant leurs versions statiques. Comme il a été déjà cité dans le Tableau 17, le dataset est classé dans catégorie des SNs qui viennent des applications qui ne sont pas strictement d'OSN. Il est à l'origine une base de données (Data log) de communications par

emails (plus 250.000 emails échangés) entre 151 employés appartenant à la société 'Enron Energy Corporation' entre mai 1999 et Juin 2002. Enron est une société spécialisée dans le négoce et le plus grand fournisseur de gaz et services publics d'électricité aux États-Unis à la fin du XXe siècle ((von Frentz 2003)). L'échantillon de dataset que nous avons choisi est constitué d'une collection d'emails échangés entre 112 employés nœuds au cours de l'année 2000.

Des données plus pertinentes

EEN est un réseau plus pertinent notamment par rapport à notre cadre de travail car:

- Le réseau est formé par des entités sociales (des employés) qui évoluent au sein d'un environnement organisationnel et les relations sont beaucoup donc plus orientées, plus pertinentes par rapport aux liens communément déclarés sur les OSNs explicites. Le réseau parait plus adapté pour notre étude car il s'agit d'un SN émergent au sein d'une société là où les obligations institutionnelles et sociales se croisent. Ainsi, des aspects intéressants dont on a besoin peuvent se manifester. Tels acteurs sont susceptibles d'être étroitement liés à cause des interactions régulières ce qui peut produire des régions cohésives et durables. En outre, les besoins de partage d'information pour coordonner ou collaborer peuvent facilement rendre la communication dominée.

- Avant quelques années, les données de ce réseau ont été collectées par les autorités du gouvernement américain (Commission fédérale de la réglementation de l'énergie 'Federal Energy Regulatory Commission' [FERC]) et faisaient l'objet d'une enquête sur le scandale de comptabilité d'Enron. Au début de l'automne de 2001, des rumeurs commençaient à circuler sur les difficultés financières du géant de l'énergie Enron à Texas ((Noble & Weiss 2004)). La société n'était pas sur le point de déclarer la faillite seulement mais aussi coupable de méfaits de comptabilité et éthiques, ce qui a poussé le congrès américain de soulever nombreuse accusations et demandes de renseignements contre l'entreprise ((Noble & Weiss 2004)). En décembre 2001 Enron a déclaré une faillite record ((von Frentz 2003)) en découvrant des manipulations frauduleuses de comptabilité ayant dissimulé des billions de dollars de dettes ((von Frentz 2003)) (Tang et al 2010b). Dans ce cas, la détection d'un noyau est prometteuse quand il s'agit d'enquêter sur un réseau illégal dissimulé qui évolue et bénéfice d'une structure légale.
- Ces données sont aussi pertinentes dans le sens où la dimension économique et politique se croisent. On sait que les classes élites détectées comme un noyau d'un SN et qui évoluent dans un contexte économique pèsent la sur la scène politique. Selon le rapport annuaire Gcr 'Global corruption report' ((Noble & Weiss 2004)), Enron n'est que l'illustration d'une société qui a fait jouer la politique en sa faveur. Son scandale financier fut aussi une bombe politique selon les rapports du centre des politiques responsables à Washington (Center for Responsive Politics [CRP]) ((Noble & Weiss 2004)). D'habitude, le CRP met le public en conscience de l'argent et son rôle pour gagner les élections, quels intérêts ayant plus d'influence entre politiciens et quelles propositions législatives qui ont les meilleures chances de passer ((Noble & Weiss 2004)). CRP veut montrer aussi que les grands donateurs exploitent leur accès aux hommes politiciens pour établir des relations avec les élus, en cherchant plus d'influence aux plus hauts niveaux du pouvoir. Selon les statistiques de CRP, des employés d'Enron avaient fait don de près de 6 millions $US aux candidats pour le Congrès, président, et des partis politiques (républicains ou démocratiques) pendant les 13 années qui ont précédé ce scandale ((Noble & Weiss 2004)). Deux anciens P.-

D.G Ken Lay et Jeffrey Skilling étaient parmi les donateurs les plus généreux. Une histoire qui faisait la une de plusieurs grands journaux aux Etats unis et dans le monde entier. Les liens politiques avec ce scandale ont amené à se demander si Washington avait fermé les yeux sur ces transgressions ((Noble & Weiss 2004)). Sensibiliser le public de qui finance qui dans les élections américaines est la meilleure façon pour CRP pour lutter contre tels abus dans le futur (Larry Noble and Steven Weiss [CRP, United States] ((Noble & Weiss 2004)).

Prétraitement

Les données du réseau EEN ainsi que l'échantillon étudié comprend des données étiquetant des changements temporels. Sur un outil de visualisation des réseaux dynamique comme 'Commetrix', le dataset peut être visualisé dans 2 tableaux (**Figure 71**, **Figure 72**), l'un pour les nœuds et l'autre pour les liens. Le début des interactions ('Network start') date le 4 janvier 2000 à 18 :00 AM, et la fin des interactions ('Network end'), date le 30 décembre 2000 à 05 :05 AM. Donc, ce SN est observé durant la quasi-totalité de l'année 2000. Entre Janvier et décembre, on divise la période d'observation en t = 12 'time points', chacun correspond à un mois. C'est la résolution de la taille des fenêtres de temps.

Index	Group	Database-ID	Keywordlist	Email	Name	Function	Information	test	Linkevents sent	Linkevents recieved	References on linkevents recieved	Number of Linkevents	First Participation
0	0	1214		andrea.ring...	Andrea Ring	N/A	null		9	13	0	22	2000-01-31 03:47:00.0
1	0	1203		andy.zipper...	Andy Zipper	Vice President	Enron Online		11	64	0	75	2000-04-14 07:34:00.0
2	0	1163		barry.tycholi...	Barry Tycholiz	Vice President	null		10	17	0	27	2000-05-02 09:01:00.0
3	0	1156		benjamin.ro...	Benjamin Ro...	N/A	null		1	1	0	2	2000-05-16 08:46:00.0
4	0	1260		bill.williams...	Bill Williams	N/A	null		0	11	0	11	2000-09-08 10:09:00.0
5	0	1172		brad.mckay...	Brad Mckay	N/A	null		0	20	0	20	2000-06-19 09:05:00.0
6	0	1211		cara.semper...	Cara Sempe...	Employee	Senior Analy...		20	15	0	35	2000-10-31 05:47:00.0
7	0	1271		carol.clair@...	Carol Clair	House Lawyer	null		508	395	0	903	2000-01-04 07:25:00.0
8	0	1242		chris.dorland	Chris Dorland	Employee	null		12	3	0	15	2000-01-11 07:16:00.0

Figure 71. Un extrait de la table des nœuds (employés) du SN étudié depuis la BDD originale d'EEN

Dans le tableau des nœuds (**Figure 71**), on trouve essentiellement:

- L'index de chaque nœud dans le SN échantillon et son identificateur ID dans la base des données BBD originale. Chaque nœud représente l'adresse email d'un employé.
- Le nombre de son 'link-events sent' et 'link-events received' qui signifie respectivement le nombre de mails envoyés, et reçus, ainsi que le nombre total 'Number of link events'.
- La date et l'heure de la première et la dernière participation (échange)

Même si cette dernière donnée signifie qu'un acteur ne participe pas durant toute la période d'observation, nous considérons que l'ensemble des nœuds est consistant, car en effet il s'agit des employés, affiliés à la société durant toute cette période (12 'time-points'). Nous programmons un premier script Java capable de prendre en entrée le tableau des nœuds (**Figure 71**) sous forme d'une feuille CSV afin de créer en sortie la liste correspondante des nœuds sous le format (.net). Chaque ID dans le tableau sera l'étiquette d'un nœud dans la liste, associée à l'intervalle de temps [1-12] (Consistance). L'index des nœuds dans la liste commence de 1 à 112 (**Figure 73**).

Node1 ID	Node2 ID	Number of contacts	Number of linkevents	First linkevent	Last linkevent	Se
1174	1261	161	161	2000-01-04 02:18:00.0	2000-12-27 03:10:00.0	30
1142	1174	54	54	2000-01-04 04:54:00.0	2000-12-08 06:28:00.0	1
1174	1265	87	87	2000-01-04 04:54:00.0	2000-12-08 07:48:00.0	78
1221	1239	54	54	2000-01-04 05:58:00.0	2000-12-08 08:02:00.0	9
1192	1271	420	420	2000-01-04 07:25:00.0	2000-12-28 00:31:00.0	148
1130	1192	62	62	2000-01-04 07:25:00.0	2000-12-05 08:14:00.0	9
1152	1173	153	153	2000-01-04 10:15:00.0	2000-12-06 10:33:00.0	2
1139	1173	265	265	2000-01-04 10:15:00.0	2000-12-13 01:23:00.0	13
1157	1261	68	68	2000-01-05 00:30:00.0	2000-12-27 03:10:00.0	31
1173	1261	1	1	2000-01-05 00:30:00.0	2000-01-05 00:30:00.0	0
1206	1239	114	114	2000-01-05 02:08:00.0	2000-12-27 06:43:00.0	12

Figure 72. Un extrait de la table des relations du SN étudié, depuis la BDD originale d'EEN

Dans le tableau des relations (**Figure 72**), chaque ligne correspond à une relation entre 2 employés et implique :

- ID des 2 nœuds (extrémités) : Node1 ID et Node2 ID. Deux nœuds (adresses) 'u' et 'v' seront reliés, si et seulement si 'u' envoie au moins un email à 'v' ou inversement.
- Nombre de 'linke-events' qui sont les mails échangés dans les 2 sens et qui animent la relation. Ce nombre peut être un poids ou une fréquence de la relation. Toutes les relations représentent des échanges bilatéraux, ce qui permet de considérer le réseau comme non-orienté
- La date et l'heure de création de la relation. C'est la date du premier échange 'First link-event'. Ensuite la date du dernier échange 'Last link-event'. Ce sont des données permettant d'extraire la durée de vie de la relation.

Pour mettre ces liens dans le format (.net), un autre script Java est implémenté. Il prend en entrée une feuille CSV là où conserve dans chaque ligne une relation avec l'ID des 2 nœuds, la date du premier et le dernier échange (First/ Last Link-event) entre les deux (**Figure 73**). Dans chaque ligne, ce script java extrait d'abord les ID des nœuds, cherche les indexes correspondants dans la liste (.net). Ensuite, il extrait le mois (time-point) par exemple 2 à partir de la date de First Link-event : 2000-02-04 et 11 à partir de la date de First Last-event 2000-11-29, ce qui donne l'intervalle [2-11] comme durée de vie du lien (**Figure 73**).

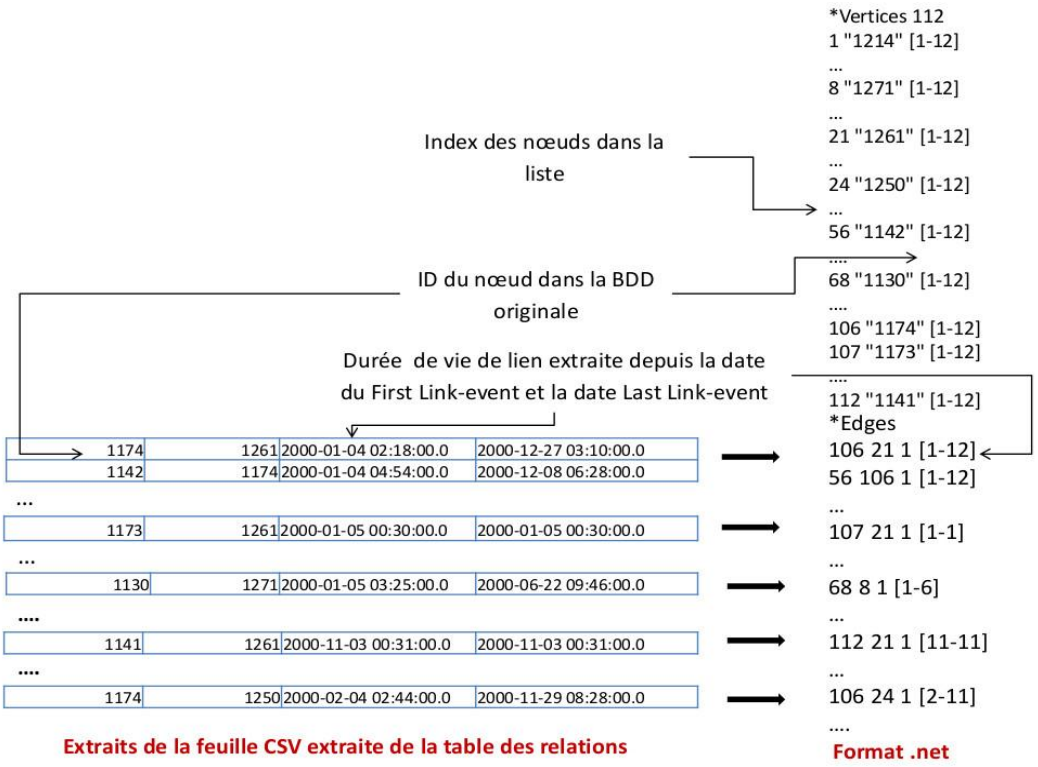


Figure 73. L'échantillon du SN d'EEN converti en format .net

Analyse de certaines caractéristiques de base

La **Figure 74** montre d'abord les 12 snapshots du SN d'EEN agrégés dans un graphe statique visualisé à travers 'Draw Window Tools' de Pajek. Sa distribution de degré a été étudiée pour vérifier que la présente structure est porche d'un SN réel. Sous Pajek, on utilise la commande : 'Network Menu > Create Vector > Centrality > Degree > All' pour calculer d'abord le vecteur des degrés des nœuds, et puis on calcule la fréquence de chaque valeur (**Figure 74**). Avec la forme statique de ce réseau, la distribution des degrés statiques suit visiblement la loi de puissance d'un SN réel (**Figure 74**). En général, plus le degré (nombre d'interactions) augmente, plus le nombre d'employés ayant ce degré diminue (**Figure 74**). Même s'il y a un certain cas particulier, quand le degré égale à 2, sa fréquence chute rapidement, puis elle reprend sa lente décroissance (**Figure 74**). Ces premières interprétations donnent un premier signe que réseau peut être dominé par un sous-ensemble d'employés ayant tout simplement beaucoup d'interactions que les autres. C'est en effet un signe sur la centralisation du réseau.

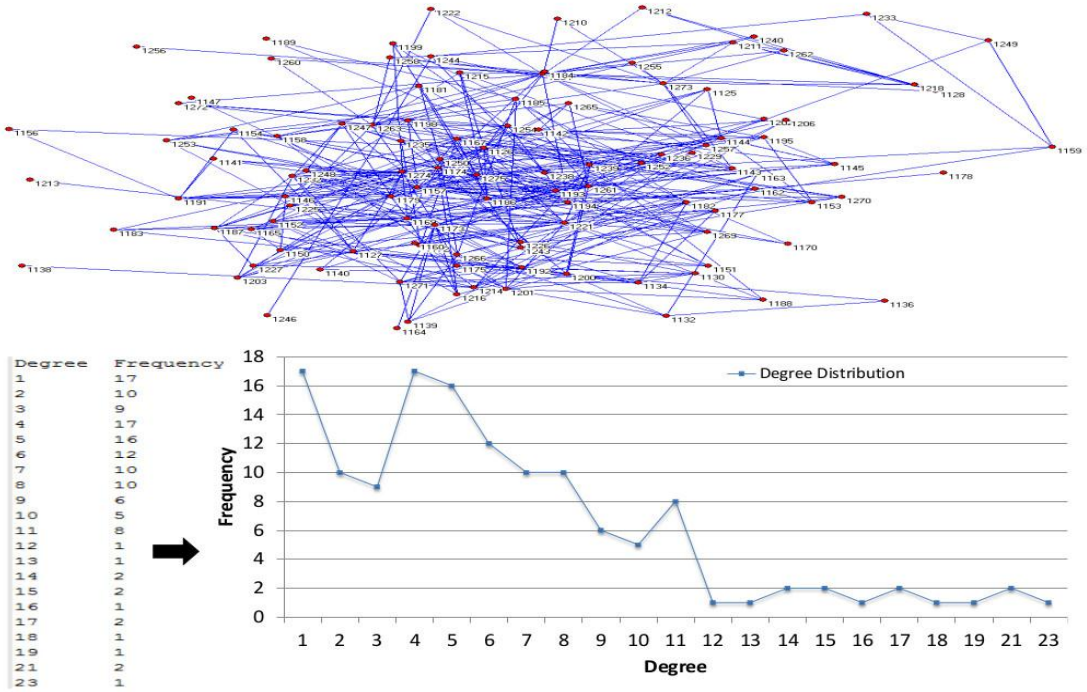


Figure 74. Distribution de degré dans le SN d'EEN, les 12 snapshots sont agrégés dans une représentation statique

On reprend la version dynamique temporelle de ce SN. Le Tableau 29 affiche le changement de certaines de ses caractéristiques (à travers les 12 snapshots), calculées par l'intermédiaire des commandes de Pajek qu'on a utilisé avec les datasets précédents.

Tableau 29. Des caractéristiques de base du SN d'EEN évalués pendant 12 time-points

Caractéristiques / Time-points	Nombre de liens	Densité	Degré moyen	Coefficient de Clustering
T1-Jan	40	0,00643501	0,71428571	0,44126984
T2-Feb	59	0,00949163	1,05357143	0,43812698
T3-Mar	74	0,01190476	1,32142857	0,55388056
T4-Apr	92	0,01480051	1,64285714	0,39361194
T5-May	106	0,01705277	1,89285714	0,36789486
T6-Jun	119	0,01914414	2,125	0,36904207
T7-Jul	140	0,02252252	2,5	0,36298473
T8-Aug	197	0,03169241	3,51785714	0,50629859
T9-Sep	194	0,03120978	3,46428571	0,52129008
T10-Oct	208	0,03346203	3,71428571	0,50252462
T11-Nov	223	0,03587516	3,98214286	0,46147382
T12-Dec	186	0,02992278	3,32142857	0,42327984

La **Figure 75** montre le premier impact de la consistance de l'ensemble d'acteurs, là où la densité du réseau est directement proportionnelle au nombre d'interactions. Ils ont quasiment la même courbe d'évolution en remarquant également le degré moyen (Tableau 29). Généralement, ce SN tend à se densifier même s'il y a une petite décroissance entre le mois de novembre et décembre.

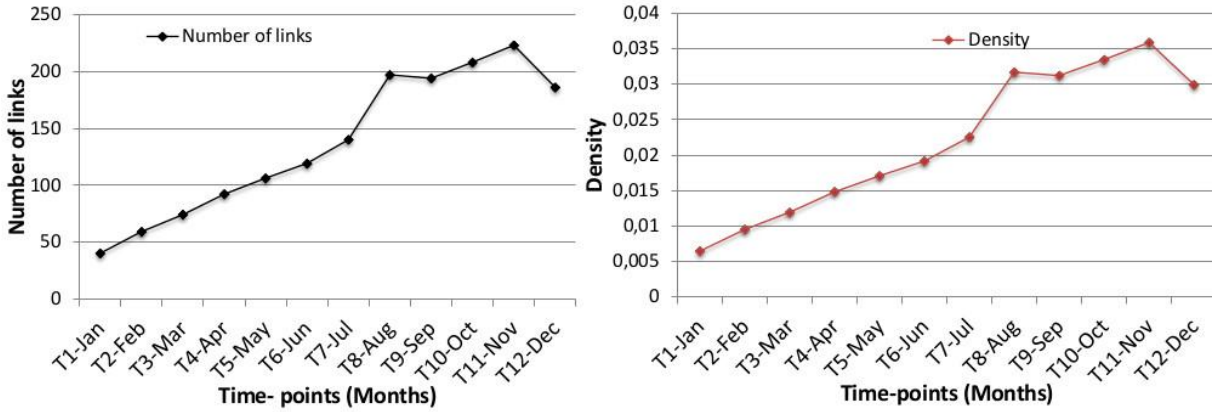


Figure 75. Nombre de liens et densité du SN d'EEN en évolution dans le temps

D'autre part, et comme on constate dans la **Figure 76**, les valeurs du CC global du SN qui sont affichées dans le Tableau 29 varient dans le temps. Chaque valeur est le résultat de calcul d'une moyenne de coefficients de clustering individuels. Le CC prend généralement des valeurs dans un intervalle estimé de 0,3 à 0,5, et connait des périodes de stabilité, par exemple entre Avril et Juillet ou Aout et Octobre (**Figure 76**). Ainsi, la tendance des acteurs à se regrouper de manière générale varie mais qui ne s'annule surtout pas durant la période d'observation.

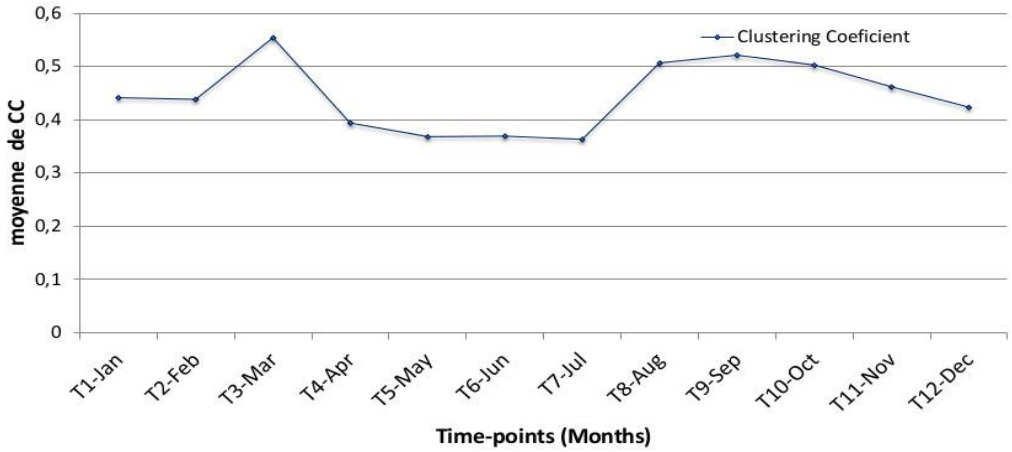


Figure 76. Variation de la moyenne de CC du SN d'EEN dans le temps

7.1.4. Bilan sur les datasets testés et lequel le plus adapté

Partant de la pertinence et la richesse des données sociales à tester, passant par son prétraitement, le choix de dataset compte sur l'analyse de certaines caractéristiques de bases (Tableau 30). D'abord, si la richesse en termes de données temporelles est assurée dans les 3 datasets, le réseau d'EEN est plus pertinent notamment quand il s'agit de détecter sa structure noyau. Une structure sous-jacente qui doit afficher des caractères spécifiques de cohésion, de résistance et de dominance, et qui seront présentés par un regroupement ayant un comportement typique en termes de stabilité, de centralité, etc. Par rapport à cet objectif de recherche, l'analyse de l'évolution des caractéristiques comme la densité, le CC, etc., est une étape préalable qui donne des signes sur la connectivité du réseau et sa centralisation. Même si le réseau extrait de MUSG est proche d'un NS réel, on le considère comme un petit SN artificiel dont le CC est plus au moins stable dans le temps. Sa densité ne cesse de décroitre qu'après avoir supposé que l'ensemble de nœuds consistant. D'autre part, le SN de Fb-lSN, présente une taille acceptable mais il est de moins en moins dense dans le temps, ce qui diminue la probabilité de le voir centralisé. Le coefficient de clustering n'est pas stable et s'annule à la fin de la période d'observation. Cependant, le SN extrait d'EEN tend à se

densifie et son CC ne s'annule pas. Il est ainsi susceptible d'être plus centralisé. En plus, la distribution degré montre qu'il y a un sous-ensemble d'employés qui domine les interactions. Sur le plan structurel, ce dataset montre encore une fois qu'il est plus adapté pour expérimenter notre approche qui cherche à détecter une identité noyau significative dans un SN évoluant dans un environnement organisationnel où plusieurs enjeux se croisent.

Tableau 30. Comparatif de datasets selon quelques critères

Critères / Dataset testés	Pertinence	Données temporel-les	Prétraitement vers (Format .net)	Densité	Evolution de CC
Echantillon de MUSG	SN artificiel	Oui	N'était pas nécessaire. Ensemble de nœuds inconsistant. Supposer une version consistante	Décroissante avec un ensemble de nœuds inconsistant	CC perturbé et n'est pas influencé par la consistance
Fb-lSN	OSN limité géographiquement et formé par des acteurs étudiants évoluant dans un environnement universitaire bien défini	Oui	Prétraitement nécessaire. Supposer l'ensemble consistant. Omettre les loops et les relations multiples	Un SN de moins en moins dense	CC tend vers 0
Echantillon d'EEN	OSN venant des Data log qui a fait l'objet des investigations, formé par des employés d'une société. Contexte et l'enjeu économique et politique se croisent	Oui	Prétraitement nécessaire	Le SN tend à se densifier	CC qui ne s'annule pas

7.2.Etudes empiriques et résultats

L'approche que nous avons proposée pour caractériser, modéliser et identifier une identité significative d'une structure noyau sera expérimentée sur le SN extrait de 'Enron email communication network'. On effectue un ensemble d'études empiriques au cours des phases de modélisation et d'identification. Après avoir franchi les étapes préliminaires abordant l'évolution basique de sa connectivité et menant ainsi à choisir et préparer ce dataset, le SN d'EEN sera représenté par le méta-modèle TW-DAG. Durant sa période d'observation, on extrait les groupes, leurs chevauchements temporels successifs et les paramètres de chacun de ces chevauchements afin d'instancier respectivement les sommets, les arcs et les poids des arcs de TW-DAG. On verra que certaines hypothèses sur la dynamique seront prouvées à un niveau plus profond à partir le TW-DAG (liaison basé sur des chevauchements temporels). Les arcs seront pondérés et les poids seront améliorés suivant le schéma de pondération. Dans ce sens, nous étudions différentes distributions de chevauchements qui vont montrer et justifier l'avantage conféré par les paramètres et les facteurs, ajoutés à la fonction de pondération par rapport à des versions antérieures. Après avoir appliqué le processus d'identification, nous verrons si un CP couvre vraiment des structures sous-jacentes,

chevauchements pertinents en cherchant une structure noyau plus raffinée, encapsulé à l'intérieur. Il faudra déterminer dans quelle mesure les paramètres localement affichés par ces chevauchements sont corrélés avec ceux d'un regroupement (N) qui persiste profondément à l'intérieur, qualifié pour présenter cette structure noyau. Par ailleurs, nous proposons des tests supplémentaires afin de valider aussi bien la phase d'identification que le calibre d'une telle structure détectée. Il s'agit d'étudier la sensibilité du réseau envers ce noyau par rapport à d'autres régions persistantes.

Découverte des zones cohésives par partitionnement en groupes dans le temps.

Il a été montré que les acteurs du SN ont une tendance à former des régions cohésives (des groupes) dans le temps, mais suivant un comportement spécifique. À cet égard, nous exportons chaque snapshot du SN de Pajek vers un outil de visualisation VOSviewer ((VOSviewer 2013)) ((Van Eck & Waltman 2010)) ((Van Eck & Waltman 2011)) plus efficace et élaboré.

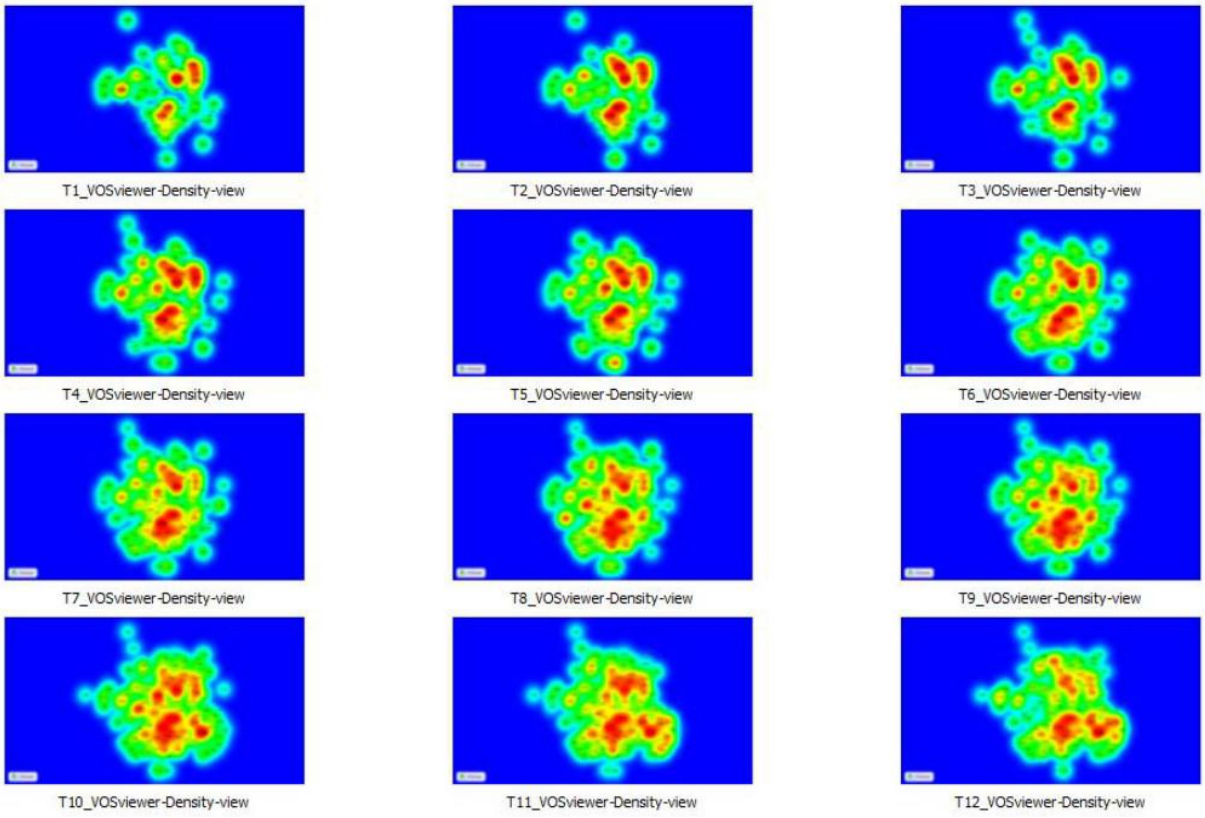


Figure 77. Vue sur les régions denses (cohésives) du SN d'EEN dans durant les 12 time-points (VOSviewer)

La **Figure 77** donne un aperçu rapide sur la densité ('Density View' ((Van Eck & Waltman 2011))) des interactions du SN dans le temps. Chaque acteur est étiqueté par son 'ID' et possède une couleur (allant du bleu vers le rouge) qui dépend de la densité de ses interactions. Plus le nombre d'interactions est important plus la couleur est proche du rouge (zone la plus dense). Inversement lorsque l'acteur est moins actif, la couleur tend vers le bleu (la plus faible densité). De T_1 jusqu'à T_8, T_9, T_{10}, T_{11} et T_{12}, des zones (comprenant les acteurs 1174, 1232, etc.) plus denses et cohésives émergent.

Les groupes seront le cadre formel pour capter ces régions cohésives. On découvre la structure en groupes de chaque snapshot du réseau à travers la méthode itérative de détection de communautés 'Louvain Method' (Nettleton 2013) (Blondel et al 2008) mieux que 'VOS

Clustering' sous Pajek (3.08). La méthode offre une décomposition dont la qualité compte sur la maximisation d'une fonction de modularité standard. Il y a la possibilité d'ajuster quelques paramètres mais on garde simplement les valeurs par défaut.
La partition se produit dans un fichier (.clu) qui a le format suivant (**Figure 78**)

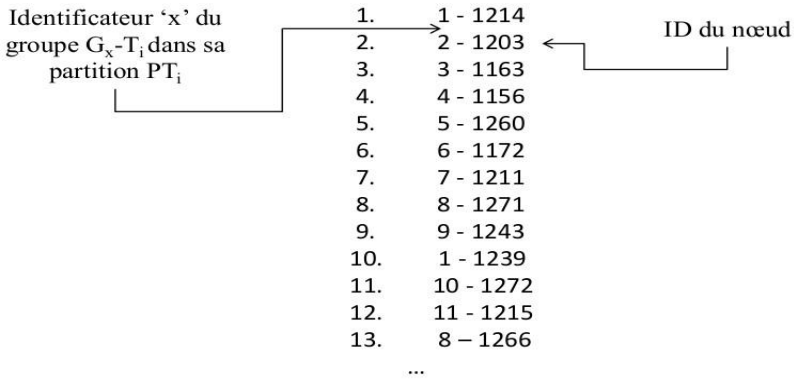


Figure 78. Format .clu d'une partition PT_1 produite par 'Louvain Method' sous Pajek

On trouve dans la **Figure 79** le bilan correspondant à ce premier partitionnement à T_1. Cette structure communautaire est exportée à VOSviewer tel que les zones qui apparaissent denses dans 'Density view' se présentent sous forme groupes (des régions cohésives) à T_1 dans 'Label View'. Avec ce dernier mode de visualisation, les acteurs se représentent par des cercles. La taille et la police de l'étiquette et du cercle dépend de l'ampleur d'interactions. L'acteur 1173 semble avoir le degré le plus élevé (**Figure 79**). D'autre part, la couleur du cercle détermine à quel cluster (groupe) l'acteur est affecté.

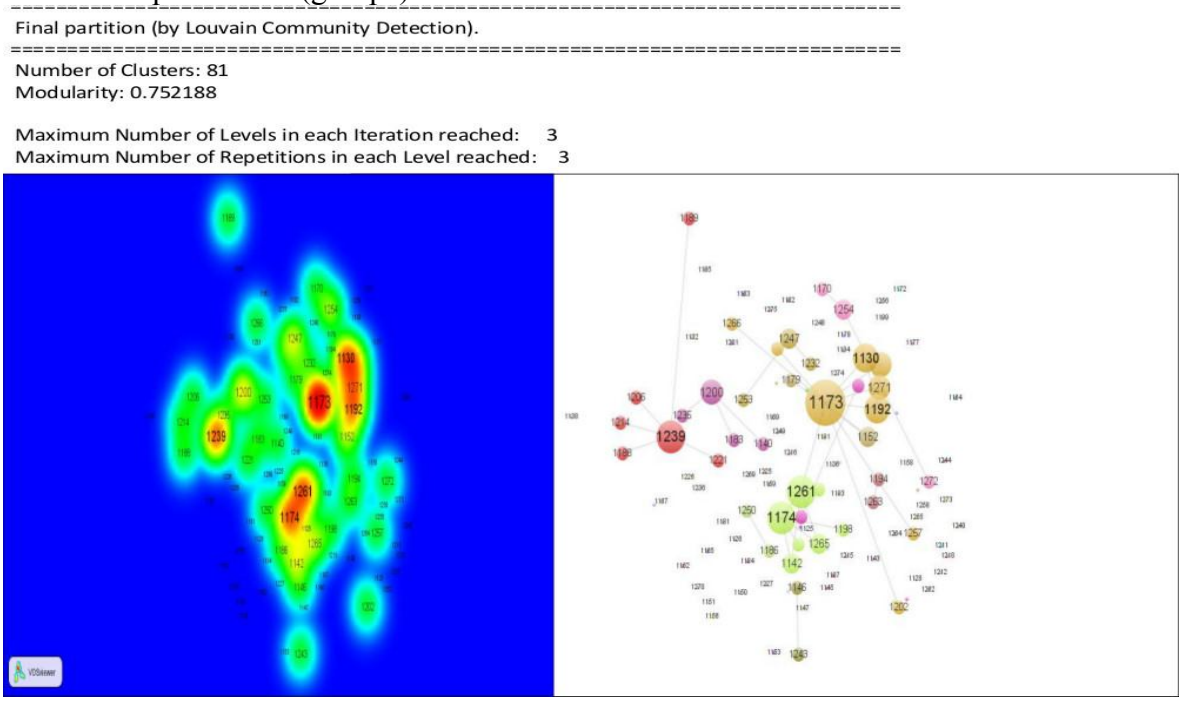
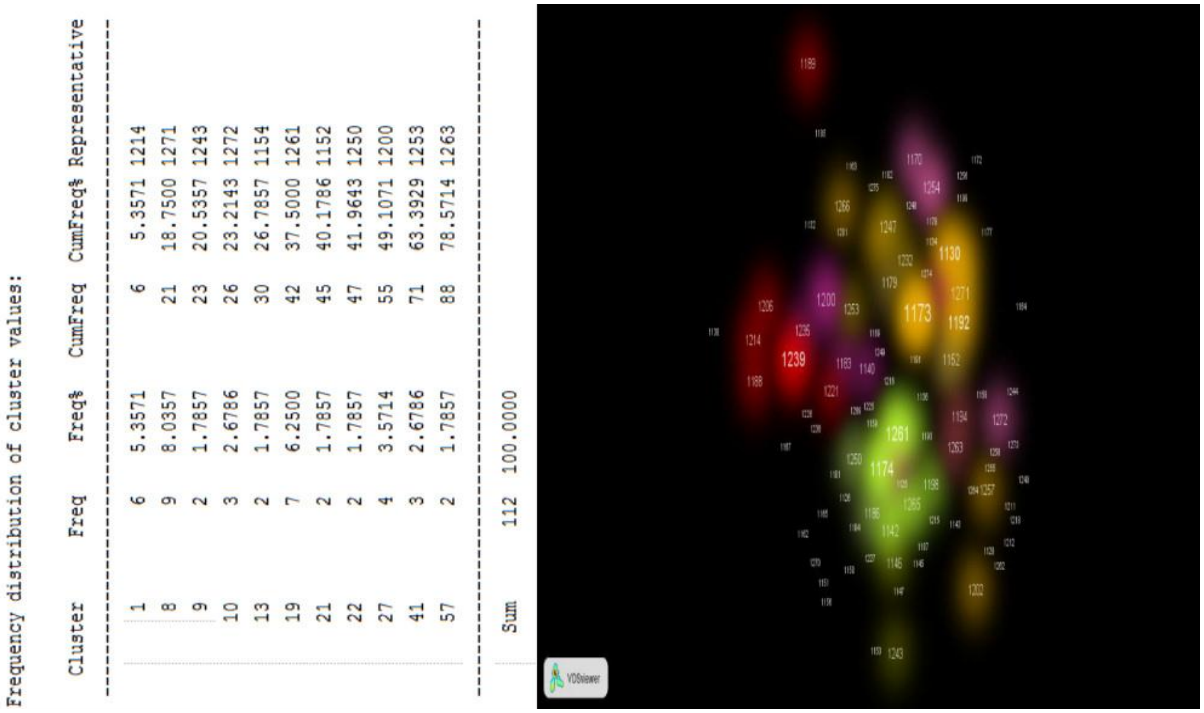


Figure 79. Le SN d'EEN à T_1 (Density View), partitionné en groupes (Label View-VOSviewer)



Frequency distribution of cluster values:

Cluster	Freq	Freq%	CumFreq	CumFreq%	Representative
1	6	5.3571	6	5.3571	1214
8	9	8.0357	21	18.7500	1271
9	2	1.7857	23	20.5357	1243
10	3	2.6786	26	23.2143	1272
13	2	1.7857	30	26.7857	1154
19	7	6.2500	42	37.5000	1261
21	2	1.7857	45	40.1786	1152
22	2	1.7857	47	41.9643	1250
27	4	3.5714	55	49.1071	1200
41	3	2.6786	71	63.3929	1253
57	2	1.7857	88	78.5714	1263
Sum	112	100.0000			

Figure 80. Clusters (groupes) du SN d'EEN à T_1 et le Cluster Density View - VOSviewer

Toujours avec le premier snapshot, la **Figure 80** donne un aperçu sur la densité de chaque cluster (groupe) à T_1 (Clsuter Density View ((Van Eck & Waltman 2010))). Il est similaire à la 'Density view' sauf que les groupes sont séparément affichés. En effet, seulement les liens intra-groupes de chaque cluster sont considérés. Les couleurs des acteurs sont nuancées et ceux qui sont proches (ayant la même couleur) ont plus de liens entre eux (plus de cohésion). Ainsi, on peut distinguer des zones plus ou moins importantes dans le cluster lui-même. La **Figure 80** montre aussi comment le nombre des acteurs est distribué sur ces clusters en omettant les singletons.

Tableau 31. Nombre de clusters et la modularité maximale affichée après le partitionnement du SN d'EEN à chaque 'time-point'

Partition PT_i à un time-point T_i	**Nombre de Clusters (Groupes)**	**Modularité (max)**
PT_1-Jan	81	0.752188
PT_2-Feb	73	0.662022
PT_3-Mar	62	0.685811
PT_4-Apr	58	0.667533
PT_5-May	51	0.665228
PT_6-Jun	43	0.674458
PT_7-Jul	39	0.648393
PT_8-Aug	30	0.626994
PT_9-Sept	28	0.633649
PT_{10}-Oct	18	0.659336
PT_{11}-Nov	20	0.648153
PT_{12}-Dec	22	0.700066

Le bilan complet du partitionnement se trouve dans le Tableau 31 qui donne le nombre de groupes G_x-T_i, pour chaque partition PT_i (i =1..12). Etant un indice de qualité en termes de cohésion des groupes détectés, la meilleure modularité est montrée quasiment stable

précisément entre T_2 T_{11} (Tableau 31). Par ailleurs, l'évolution du nombre de groupes trouvé est analysée dans la **Figure 81.**

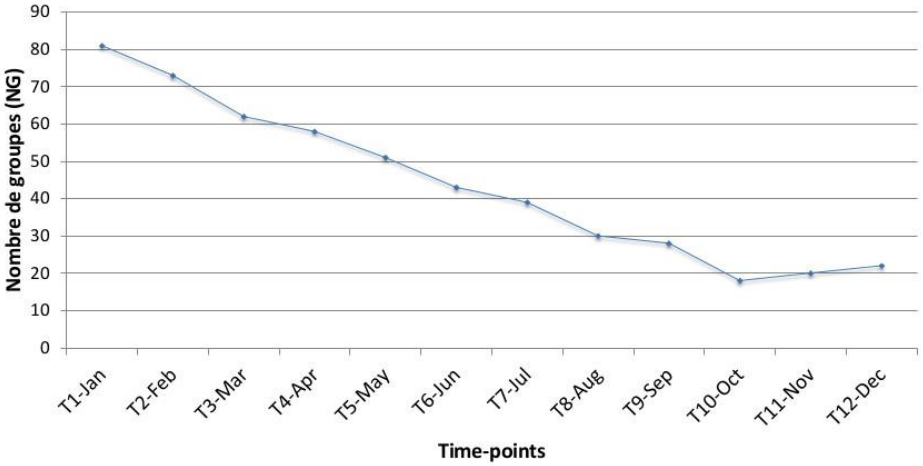


Figure 81. Variation du nombre de groupes du SN d'EEN dans le temps. 525 groupes au total avec une moyenne de 43 et 44 groupes par T_i

Au départ, le réseau est moins dense (**Figure 75**), c'est la première raison derrière le nombre important de groupes (NG dans la **Figure 81**). En effet, il y a des acteurs isolés (**Figure 79** et **Figure 80**) qui sont automatiquement classés par n'importe quelle méthode de décomposition comme des groupes singletons. Après, on sait que le réseau se développe en devenant constamment de plus en plus dense. Ainsi, les acteurs se rapprochent, devenant fortement liés, ce qui permet de réunir et fusionne des groupes, devenant plus large et plus dense et assure également le paramètre de cohésion. Par conséquent, le NG diminue au moment où le réseau tend à se densifier jusqu'à les derniers mois (**Figure 81**). En observant la densité (Cluster Density View) de ces groupes dans la **Figure 82**, on voit que le nombre de couleurs augmente dans le temps. Ce n'est pas parce que le NG progresse mais plutôt les petits groupes et les groupes singletons qui deviennent plus denses apparaissent avec des couleurs plus saillantes. NG est tellement sensible à la densité du SN qu'il augmente légèrement dans le dernier mois quand la densité affiche une petite chute (**Figure 75** et **Figure 81**).

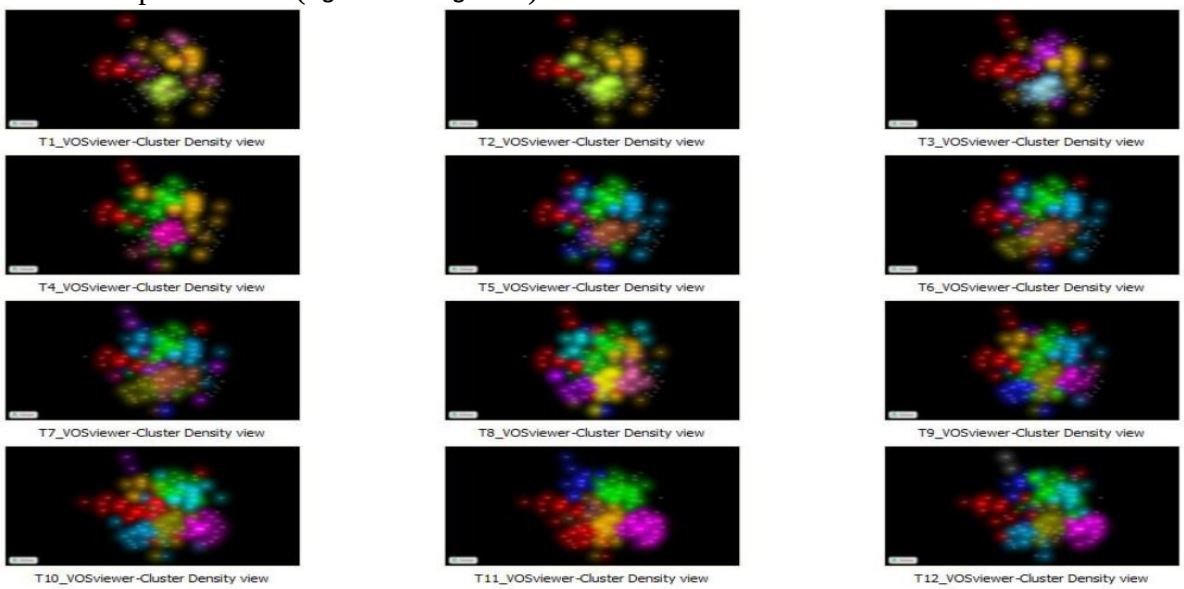


Figure 82. Cluster Density View sur les groupes du SN d'EEN dans le temps (VOSviewer)

Représenter le SN d'EEN par le méta-modèle TW-DAG standard

À chaque time-point, le SN se décrit maintenant par une partition donnée. C'est un ensemble de groupes, l'équivalent d'une partie de sommets dans le TW-DAG. On commence par créer une version standard de TW-DAG.

Ses arcs seront au départ crées et pondérés simplement par la taille des chevauchements (intersections) temporels, une simple similarité entre deux groupes de deux partitions différentes. Les valeurs de similarité seront fournies par une matrice appelée R (t-1 × t-1). R est proposé comme une matrice de ressemblance qui offre la taille de chaque chevauchement qui a lieu entre chaque groupe G_x-T_i et G_y-T_i. Pour calculer ces valeurs, on fait appel d'abord à la commande 'Partitions Menu > Info' sous Pajek qui prend en entrée 2 partitions données (2 fichiers .clu) et renvoie une table de contingence (.txt). D'où, si on applique la fonction sur 2 partitions d'un réseau dans le temps PT_i et PT_j, la table de contingence résultante (Tableau 32) nous offre la taille d'intersection (de chevauchement temporel) entre chaque G_x-$T_i \in PT_i$ et G_y-$T_j \in PT_j$. $1 \leq i < j \leq 12$, i=1..t-1, j=2..t, (t = 12). La table se transforme en une feuille CSV (Tableau 32).

Tableau 32. Aperçu sur la table contingence (81×73) entre les groupes de PT_1 et ceux de PT_2

y de G_y-PT_2 / x de G_x-PT_1	1	2	3	4	5	6	7	8	9	10
1	6	0	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0
6	0	0	0	0	0	1	0	0	0	0
7	0	0	0	0	0	0	1	0	0	0
8	0	0	0	0	0	0	0	9	0	0
9	0	0	0	0	0	0	0	0	2	0
10	0	0	0	0	0	0	0	0	1	1
11	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0

On constate que le Tableau 32 inclut un type de lignes ou des colonnes qui contient que des 0 sauf une cellule à 1. Cette ligne ou colonne correspond à un groupe singleton qui donne un chevauchement pauvre. Il peut rester singleton dans le prochain 'time-points', sinon il ne donnera pas un chevauchement intéressant, soit il reste isolée ou affichant une très faible centralité. Ce type de groupe ne sera pas ainsi considéré dans ces tables, ce qui donne des tables de contingences ajustées. Ensuite pour chaque PT_i, on construit un tableau de contingence, le résultat de concaténation horizontale de ses 12 – i tables de contingence avec les partitions suivantes (Tableau 33). Voici un aperçu sur le tableau de contingence des groupes de PT_1 avec les groupes non-singletons des partitions suivantes dans le temps, partant de PT_2 jusqu'à PT_{12} (Tableau 33).

Tableau 33. Aperçu sur le tableau de contingence de PT_i résultant de la concaténation des 12 – i tables de contingence

PT_2 > / PT_1	1	8	9	19	20	21	22	PT_3 >	1	…
1	6	0	0	0	0	0	0		6	…
8	0	9	0	0	0	0	0		0	…
9	0	0	2	0	0	0	0		0	…
10	0	0	1	0	0	0	0		0	…
13	0	0	0	0	0	0	0		0	…
19	0	0	0	7	0	0	0		0	…
21	0	0	1	0	0	1	0		0	…
22	0	0	0	0	0	0	2		0	…
27	0	0	0	4	0	0	0		4	…
41	0	0	3	0	0	0	0		0	…
57	0	0	0	0	0	2	0		0	…
…	…	…	…	…	…	…	…		…	…

Les 11 tableaux de contingences (de PT_1 jusqu'à PT_{11}) forme chacun un bloc de lignes dans la matrice R, une matrice asymétrique, d'adjacence entre les groupes dans le temps (Tableau 34). L'entête de chaque ligne est l'identificateur d'un groupe (non-singleton) d'une partition PT_i (i = 1..11). L'entête de chaque colonne est un identificateur de groupe (non-singletons) appartenant à une partition PT_j. (j = 2..12, i < j). Dans chaque croisement entre un bloc de lignes PT_i et un bloc de colonnes PT_j, on trouve les valeurs d'une table de contingence (de similarité) $PT_i \times PT_j$ (Tableau 34).

Donc, R [i, j] est un bloc, une sous-matrice $PT_i \times PT_j$, tel que i (ainsi que j) présente des sous-lignes (sous-colonnes) indexées par les groupes G_x-$T_i \in PT_i$ (G_y-$T_i \in PT_i$), $1 \leq i < j \leq t$, i = 1..t-1, j = 2..t, (t =12). Les étapes de construction de la matrice R est résumée par le pseudo-code suivant

Algorithme 4. Tables de contingence entre PT_i et PT_j formant les blocs de la matrice de ressemblance R

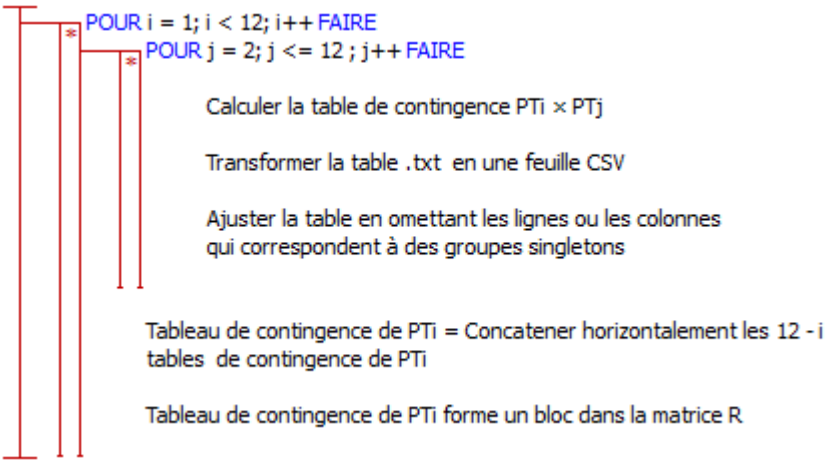


Tableau 34. Matrice de ressemblance R

PT2 > / PT1 ↓	1	8	9	19	20	21	22	PT3 >	1	8	9	18	23	PT4 >	1	2	8	9	13	17	20	41	PT5 >	1	2	3	7	11	15	PT6 >	1	…
1	6	0	0	0	0	0	0		6	0	0	0	0		6	0	0	0	0	0	0	0		5	0	0	0	0	0		5	
8	0	9	0	0	0	0	0		0	9	0	0	0		0	0	9	0	0	0	0	0		0	0	0	9	0	0		0	
9	0	0	2	0	0	0	0		0	0	2	0	0		0	0	0	2	0	0	0	0		0	0	2	0	0	0		0	
10	0	0	1	0	0	0	0		0	0	0	0	2		0	2	0	0	0	0	0	0		0	2	0	0	0	0		0	
13	0	0	0	0	0	0	0		0	0	0	0	0		0	0	1	0	0	0	0	0		0	0	0	0	0	0		0	
19	0	0	0	7	0	0	0		0	0	0	7	0		0	0	0	0	7	0	0	0		0	0	0	0	7	0		0	
21	0	0	1	0	0	1	0		0	1	0	0	1		0	1	1	0	0	0	0	0		0	1	0	1	0	0		0	
22	0	0	0	0	0	0	2		0	0	0	2	0		0	0	0	0	0	2	0	0		0	0	0	0	0	2		0	
27	0	0	0	4	0	0	0		4	0	0	0	0		0	0	0	0	0	4	0	0		0	0	0	0	1	3		0	
41	0	0	3	0	0	0	0		0	0	0	0	3		0	3	0	0	0	0	0	0		0	3	0	0	0	0		0	
57	0	0	0	0	0	2	0		0	2	0	0	0		0	0	0	0	0	0	0	2		0	0	0	0	2	0		0	
PT2 ↓								**PT3 >**	**1**	**8**	**9**	**18**	**23**	**PT4 >**	**1**	**2**	**8**	**9**	**13**	**17**	**20**	**41**	**PT5 >**	**1**	**2**	**3**	**7**	**11**	**15**	**PT6 >**	**1**	**…**
1									6	0	0	0	0		6	0	0	0	0	0	0	0		5	0	0	0	0	0		5	
8									0	9	0	0	0		0	0	9	0	0	0	0	0		0	0	0	9	0	0		0	
9									0	0	3	0	8		0	8	0	3	0	0	0	0		0	8	3	0	0	0		1	
19									4	0	0	10	0		0	0	0	0	10	4	0	0		0	0	0	0	11	3		0	
21									0	3	0	0	0		0	0	1	0	0	0	0	2		0	0	0	1	2	0		0	
22									0	0	0	2	0		0	0	0	0	0	2	0	0		0	0	0	0	0	2		0	
PT3 ↓														**PT4 >**	**1**	**2**	**8**	**9**	**13**	**17**	**20**	**41**	**PT5 >**	**1**	**2**	**3**	**7**	**11**	**15**	**PT6 >**	**1**	**…**
1															7	0	0	0	0	4	0	0		6	0	0	0	1	3		6	
8															0	0	11	0	0	0	0	2		0	0	0	11	2	0		0	
9															0	0	0	3	0	0	0	0		0	0	3	0	0	0		0	
18															0	0	0	0	11	2	0	0		0	0	0	0	11	2		0	
23															0	12	0	0	0	0	0	0		0	12	0	0	0	0		1	
PT4 ↓																							**PT5 >**	**1**	**2**	**3**	**7**	**11**	**15**	**PT6 >**	**1**	**…**
1																								6	0	0	0	0	0		6	
2																								0	15	0	0	0	0		2	
8																								0	0	0	13	0	0		0	
9																								0	0	3	0	0	0		0	
13																								0	0	0	0	12	0		0	
17																								0	0	0	0	1	6		0	
20																								0	0	2	0	0	0		0	
41																								0	0	0	0	2	0		0	
PT5 ↓																														**PT6 >**	**1**	**….**
1																															6	
2																															2	
3																															0	
7																															0	
11																															0	
…								…..						…….									…….							….	…	

R est la matrice de poids standards qui nous permettra de construire à travers un programme java la première version du méta-modèle TW-DAG (Algorithme 5). Ce dernier est généré sous forme d'un réseau .net en deux étapes :

- Générer une liste V des 525 nœuds qui représentent les 525 groupes trouvés dans les 12 time-points. Chaque nœud est étiqueté sous la forme suivante : G_x-T_i tel que le x est l'indice du groupe dans sa partition PT_i.
- Tant que TW-DAG se base sur les chevauchements temporels successifs dans la formations des arcs, le programme java se focalise particulièrement sur les blocs du diagonal (i, j = i+1) de R (Tableau 34). Pour accéder à la matrice en java, R est transformé en un fichier .xls qui sera accessible en lignes et colonnes en utilisant l'API POI (version 3.9 – 2012-12-03), (Algorithme 5). R est parcouru ligne par ligne et cellule par cellule. Dans la première case de chaque ligne, on récupère l'indice x du groupe G_x-T_i dans sa partition PT_i, et le remplace par son indice correspondant (étant un nœud) de la liste V dans la liste des arcs : une première extrémité d'un arc 'a'

(**Figure 83**). Ensuite, il faut parcourir les cases suivantes, et chercher celles avec des valeurs ≥ 1 (un seuil) qui sont poids d'intersection/ taille de chevauchements de G_x-T_i avec les autres groupes de partitions suivantes. On récupère l'indice du groupe G_y-T_{i+1} qui se trouve dans l'entête de colonne de chacune de ces cases, et le remplace par son indice correspondant (étant un nœud) depuis la liste V dans la liste des arcs : la deuxième extrémité de 'a', succédé par son poids (**Figure 83**).

Algorithme 5. Pseudo-code pour créer les arcs de TW-DAG avec des poids standards à partir de la matrice R

```
- Initialisation des variables qui concernent un groupe d'une partition courante PTi
- Indices de groupes dans leur partiton PTi
int groupe_courant_PTi = 0;
int groupe_prec_PTi = 0;
-
int groupe_courant_PTi_real_index = 1;
Ti = 1;
- Initialisation des variables qui concernent un groupe d'une partition PTi+1
- Indices de groupes dans leur partiton PTi+1
int groupe_PTi_+_1 = 0;
int groupe_prec_PTi_+_1 = 0;
- Indices du groupe de PTi étant un noeud dans la liste des noeuds
int groupe_PTi_+_1_real_index = 82;
int  Ti_+_1 = 2;

int numLigne = 0;
POUR  (Iterator rowIt = sheet.rowIterator(); rowIt.hasNext();) FAIRE
      numLigne++;
-     Parcourir une ligne
      row = (HSSFRow) rowIt.next();
      POUR (Iterator cellIt = row.cellIterator(); cellIt.hasNext();) FAIRE
      -     Parcourir les ceullules d'une ligne
            cell = (HSSFCell) cellIt.next();

            SI cell.getColumnIndex() == 0 ALORS
            -     Si on est dans la première colonne qui contient les indices des groupes de PTi
                  groupe_prec_PTi  = groupe_courant_PTi;

            -     Extraire l'indice du groupe dans PTi
                  groupe_courant_PTi = (int) cell.getNumericCellValue();

            -     Si on entame une nouvelle partition
                  SI groupe_courant_PTi < groupe_prec_PTi     ALORS
                  -     Si on est  passé verticalementà une nouvelle partition PTi
                        Ti++;
                  -     Opération pour trouver groupe_courant_PTi_real_index
                        groupe_prec_PTi  = groupe_courant_PTi;
            SINON
            -     Si on n'est pas dans la première colonne
                  SI  (numLigne > 1)   ALORS
                        groupe_prec_PTi_+_1 =  groupe_PTi_+_1
                  -     Extraire l'indice du groupe dans PTi+1
                        groupe_PTi_+_1 = (int) row_partition_suivante.getCell(cell.getColumnIndex()).getNumericCellValue();
                  -     L'indice de ce groupe dans la liste des noeuds

                        SI (groupe_PTi_+_1 < groupe_prec_PTi_+_1) & (Ti_+_1 <= 12) ALORS
                        -     Si on est  passé horizontalement à une nouvelle partition PTi+1
                              Ti_+_1 ++;
                        -     Opération pour trouver groupe_PTi_+_ 1_real_index
                              groupe_prec_PTi_+_1 =  groupe_PTi_+_1;
                        int taille_de_chevauchement  = (int) cell.getNumericCellValue();

                        SI  (taille_de_chevauchement > 0 )   ALORS
                        -     Créer l'arc en exprimant les indices des groupes (noeuds) impliqués
                              int Gx_Ti  =groupe_courant_PTi_real_index + groupe_courant_PTi-1;
                              int Gy_Ti_+_1 = groupe_PTi_+_ 1_real_index+  groupe_PTi_+_1; - 1;
                              System.out.println(Gx_Ti +" "+ Gy_Ti_+_1   +" "+taille_de_chevauchement);
```

*Vertices 525	*Arcs
1 "G1-T1"	1 82 6
2 "G2-T1"	1 155 6
...	...
522 "G19-T12"	491 512 14
523 "G20-T12"	492 514 4
524 "G21-T12"	494 504 4
525 "G22-T12"	494 505 2

Figure 83. Format .net du méta-modèle TW-DAG standard

La figure suivante (**Figure 84**) donne un premier aperçu sur une version complète du graphe (élargi) généré par l'Algorithme 5, si on considère aussi les liens possibles entre les groupes de PT_i et tous les autres groupes de PT_j ($i < j$). Comme chaque nœud dans ce graphe, G_9-T_{11} représente un groupe (G_9) dans une partition (PT_{11}) et possède selon la **Figure 85** les propriétés suivantes :

- G_9-T_{11} a au plus une seule relation avec les autres groupes, tandis qu'il n'a aucune relation les groupes de la même partition.
- Si G_9-T_{11} est la cible d'un arc donné, cela implique que la source est un groupe de PT_i précédente tel que $i < 11$, par exemple G_8-$T_4 > G_{11}$-T_9, G_7-$T_5 > G_{11}$-T9, etc.
- Si G_9-T_{11} constitue est source d'un arc donné, cela implique que la cible est un groupe d'une partition PT_j suivante dans le temps ($11< j$), par exemple G_{11}-$T_9 > G_{10}$-T_{10}, G_{11}-$T_9 > G_9$-T_{11}, etc.

Évidemment, chaque nœud comme G_3-T_{12} qui représente un groupe (G_3) appartenant à la dernière partition PT_{12} ne possède que des arcs entrants. Après avoir vérifié « l'acyclicité » de ce graphe sur Pajek (**Figure 86**), ces exemples illustratifs montre encore une fois qu'il s'agit d'un graphe acyclique, orienté dans le temps (DAG).

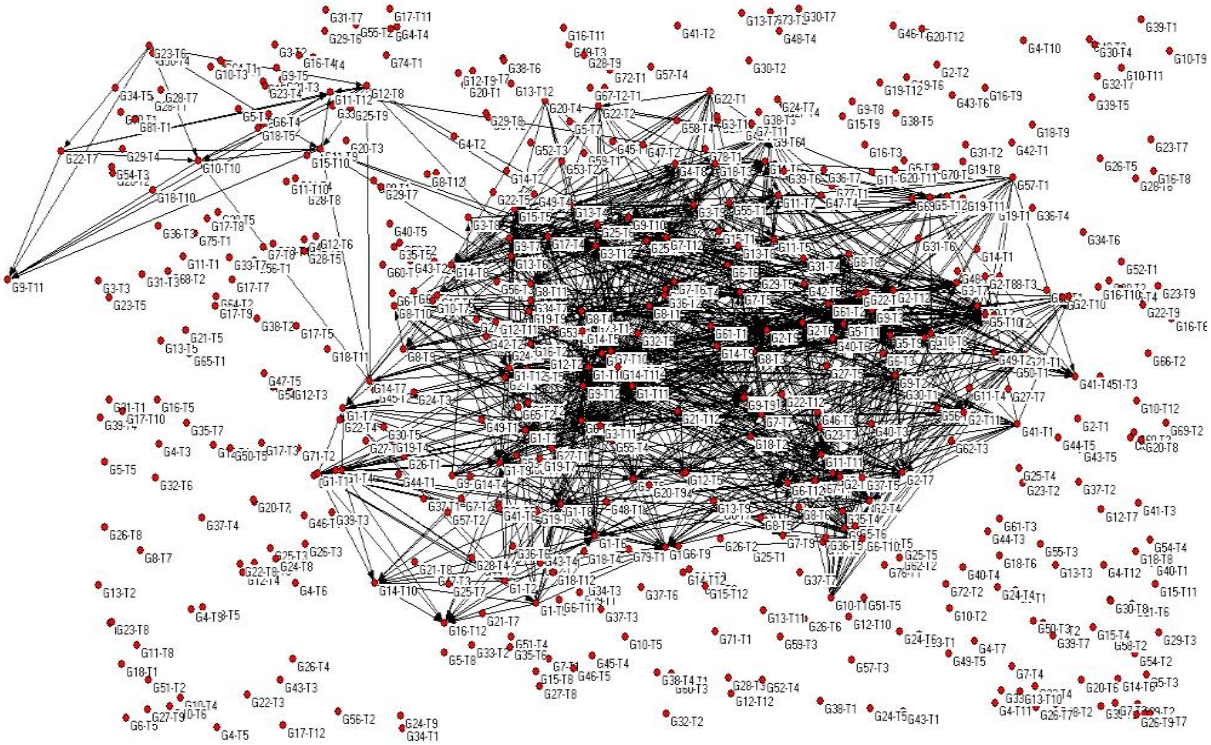


Figure 84. Graphe élargi de TW-DAG standard visualisé par Pajek

```
224.448          val=1.0000  / G8-T4.G11-T9
281.448          val=1.0000  / G7-T5.G11-T9
332.448          val=1.0000  / G7-T6.G11-T9
348.448          val=2.0000  / G23-T6.G11-T9
382.448          val=2.0000  / G14-T7.G11-T9
390.448          val=2.0000  / G22-T7.G11-T9
419.448          val=5.0000  / G12-T8.G11-T9
448.475          val=5.0000  / G11-T9.G10-T10
448.492          val=4.0000  / G11-T9.G9-T11
448.514          val=4.0000  / G11-T9.G11-T12
```

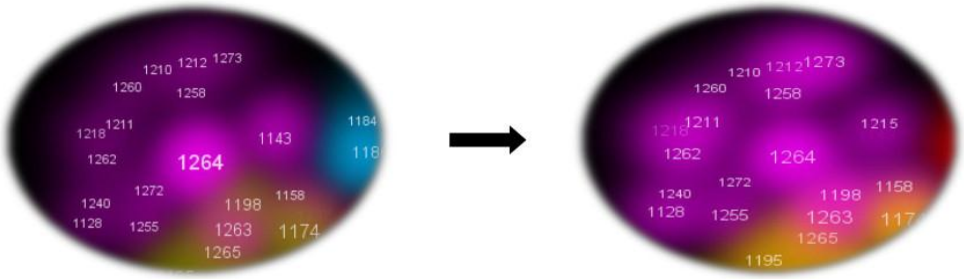
Arcs | Poids | Arcs avec des extrémités étiquetées

Figure 85. Arcs entrant et sortant du groupe G_9-T_{11} avec tous les groupes possibles des autres partitions

```
Number of vertices (n): 525
-----------------------------------------------------------
                                    Arcs          Edges
-----------------------------------------------------------
Number of lines with value=1         358              0
Number of lines with value#1         679              0
-----------------------------------------------------------
Total number of lines               1037              0
-----------------------------------------------------------
Number of loops                        0              0   Maximum Input Degree:     37
Number of multiple lines               0              0   Maximum Output Degree:    33
-----------------------------------------------------------

Density1 [loops allowed]    = 0.00376236
Density2 [no loops allowed] = 0.00376954
Average Degree = 3.95047619

The highest values of lines:

      Rank          Line                    Value    Line-Id
-------------------------------------------------------------
         1       470.488                 19.00000    G5-T10.G5-T11
```

Figure 86. Quelques statistiques fournies par Pajek sur la version élargie de TW-DAG généré par Algorithme 5

L'arc le plus lourd (**Figure 86**) qui représente le plus grand chevauchement (de 19 acteurs) est de G_5-T_{10} vers G_5-T_{11}. En effet, G_5-T_{10} se compose de 19 acteurs et évolue vers G_5-T_{11} qui contient 21 acteurs et apparait plus dense (**Figure 87**). En d'autre termes G_5-$T_{10} \subset G_5$-T_{11} et persiste localement dans le temps.



G_5-T_{10} G_5-T_{11}

Figure 87. Arc le plus lourd dans TW-DAG qui représente l'évolution du groupe G_5-T_{10} vers G_5-T_{11} (Cluster Density View-VOSviewer)

Cet arc, le plus lourd, est entre 2 point de temps successifs. C'est un point avantageux pour nos hypothèses concernant le potentiel des chevauchements temporels successifs sur lesquels se base le méta-modèle TW-DAG. La **Figure 88** propose un mode de visualisation plus avancé, qui s'adapte parfaitement avec une telle représentation d'un processus évolutionnaire de groupes dans le temps. Cette visualisation donne une vue générationnelle des groupes et leurs liaisons pendant les 12 'time-points'. Elle est d'abord obtenue sous Pajek par la commande suivante 'Network Menu>Acyclic Network > Depth Partition> Generational', étant l'une des

opérations spécifiques aux réseaux acycliques. Cette opération produit une partition (.clu) de nœuds dans des couches. Dans notre cas chaque couche est en effet un ensemble de groupes qui se distinguent par une couleur unique à T_i, une sorte de généalogie (**Figure 88**). Le résultat de l'opération, comme il est montré dans la **Figure 88** est exporté à VOSviewer. Enfin, on visualise le graphe élargi de TW-DAG (autrement que dans la **Figure 84**) ainsi que le vrai méta-modèle TW-DAG (Des groupes non-singletons successivement liés dans le temps) avec des poids standards (**Figure 88**). Un nœud (groupe) est proportionnellement dimensionnée en fonction du nombre de ses arcs (entrants/ sortants) et de leur poids. Un arc relie deux groupes qui se chevauchent entre T_i et T_{i+1}

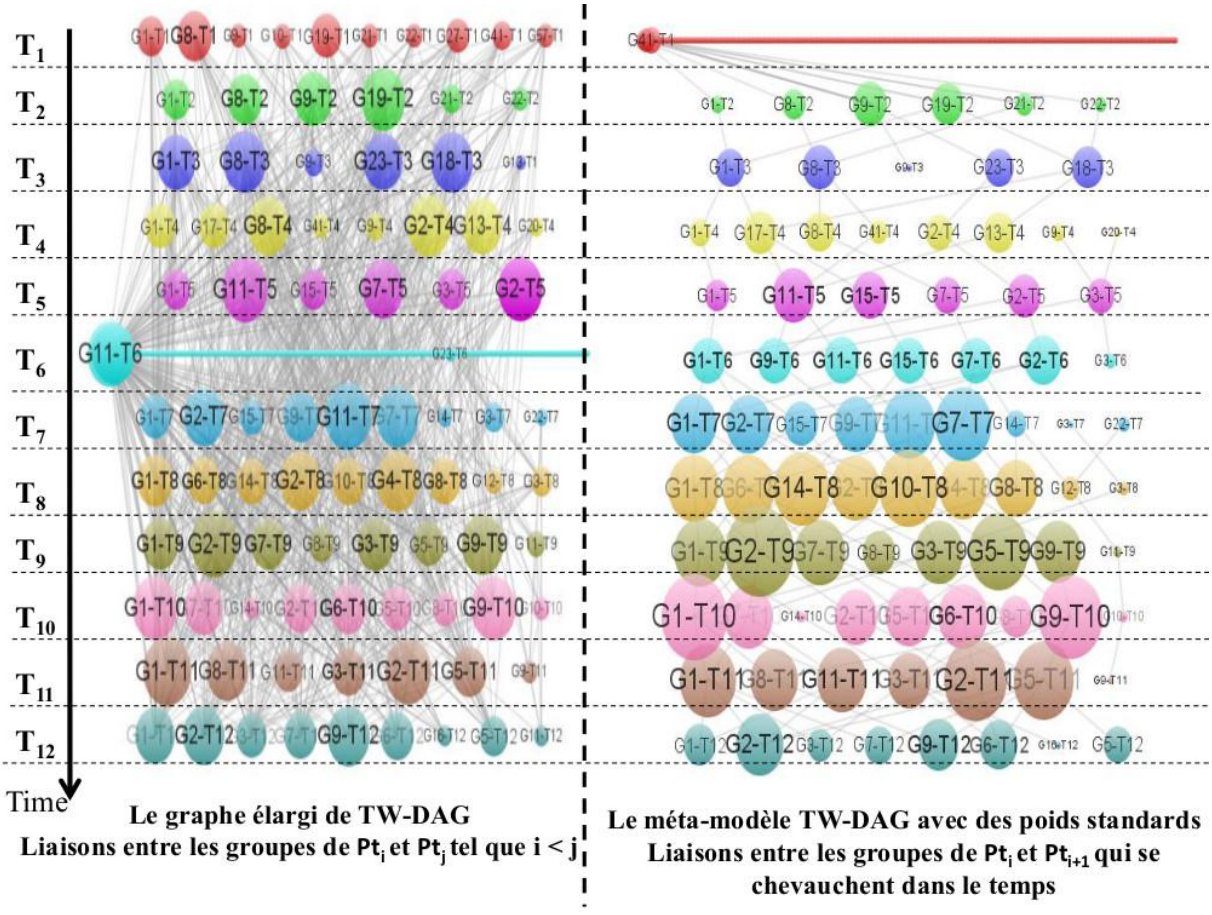


Figure 88.Graphe élargi et le vrai méta-modèle proposé TW-DAG qui représente le réseau d'EEN avec des poids standards (Generational/ layer view- VOSviewer)

Dans les premiers 'time-points', les arcs sont plus facilement à distinguer (**Figure 88** à droite). Leur nombre qui est inférieur ou égale à Z, le nombre des chevauchements qui ont eu lieu entre PT_i et PT_{i+1}. Ensuite, si le nombre des groupes diminue avec le temps (**Figure 81**), la **Figure 88** confirme aussi qu'ils ont tendance à devenir plus grand. D'où, ils sont susceptibles de se chevaucher plus entre T_i et T_{i+1} et ainsi il y a plus d'arcs entre eux dans les time-points suivants (**Figure 88**). Par conséquent, Z est en croissance comme il est montré dans la **Figure 89**. Mais, Z qui est présenté en pourcentage par rapport le nombre théorique prévu des chevauchements X × Y, n'atteint jamais X × Y (**Figure 89**). Cette statistique prouve pratiquement nos hypothèses précédentes sur la dynamique des groupes et leurs chevauchements temporels dans tels contexte réduit, dans la partie : 'Des arcs impliquant des chevauchements temporels'. Un groupe G_x-T_i se chevauche significativement avec peu de groupes de PT_{i+1} et parfois, toute sa composition est conservée dans un seul groupe G_y-T_{i+1}.

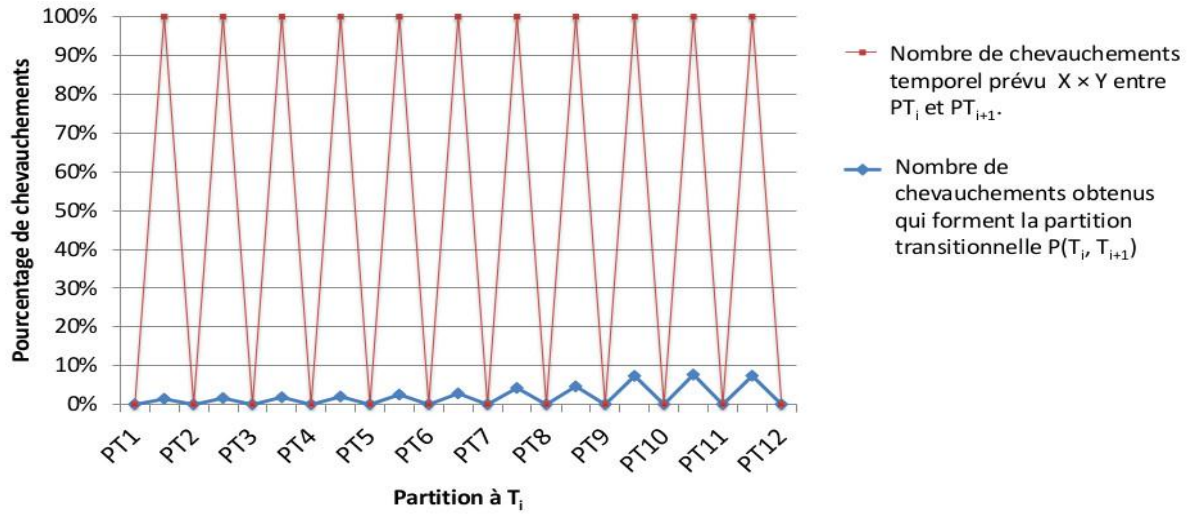


Figure 89.Pourcentage du nombre de chevauchements temporels obtenus par rapport le nombre théorique prévu entre PT_i et PT_{i+1}

Ces chevauchements temporels successifs ne sont pas le seul moyen qui montre la dépendance entre ces partions dans le temps. Il y a aussi à des mesures statistiques bi-variées comme : Cramers'V un indice de dépendance, 'Adjuster Rand Index' ARI un indice de similarité, 'Spearman correlation coefficient', 'Rajski index' un indice de dépendance ((Tab. univ. & biv. 2013)) ((De Nooy et al 2004)). Ce sont des mesures fournies par la commande Partitions menu> Info sous Pajek, et qui donnent des estimations sur la corrélation et la dépendance de deux partitions données. Dans notre cas, on peut les appliquer sur chaque paire de partitions successives dans le temps. Le Tableau 35 montre par exemple que le rapport de similarité ARI entre chaque partition PT_i et PT_{i+1} (valeurs diagonales) est souvent plus important par rapport PT_{i+2},…, PT_{12}.

Tableau 35. Rapport de similarité calculé ARI entre partitions PT_i et PT_j (i <j)

	PT1	PT2	PT3	PT4	PT5	PT6	PT7	PT8	PT9	PT10	PT11	PT12
PT1		0,57079161	0,40330716	0,41411	0,29519438	0,27051043	0,21563248	0,1780436	0,14186396	0,11185873	0,07731121	0,10876861
PT2			0,50737899	0,51022048	0,42739657	0,37541803	0,33168374	0,28301183	0,25261916	0,18260644	0,10663213	0,16575858
PT3				0,64446516	0,52229635	0,43613908	0,39651151	0,30927533	0,31251286	0,21387414	0,15649167	0,20375816
PT4					0,78329744	0,6251886	0,51982103	0,40967504	0,41748875	0,25693656	0,17986911	0,24511212
PT5						0,72038439	0,61617491	0,42184029	0,47193099	0,29851799	0,1770332	0,24529964
PT6							0,75945622	0,5663571	0,54088871	0,41007249	0,26252343	0,36400391
PT7								0,47213789	0,52014271	0,40841908	0,29805195	0,34888959
PT8									0,68781268	0,4360705	0,35644321	0,34733706
PT9										0,48248023	0,36874762	0,45336476
PT10											0,61765669	0,64227222
PT11												0,56075444

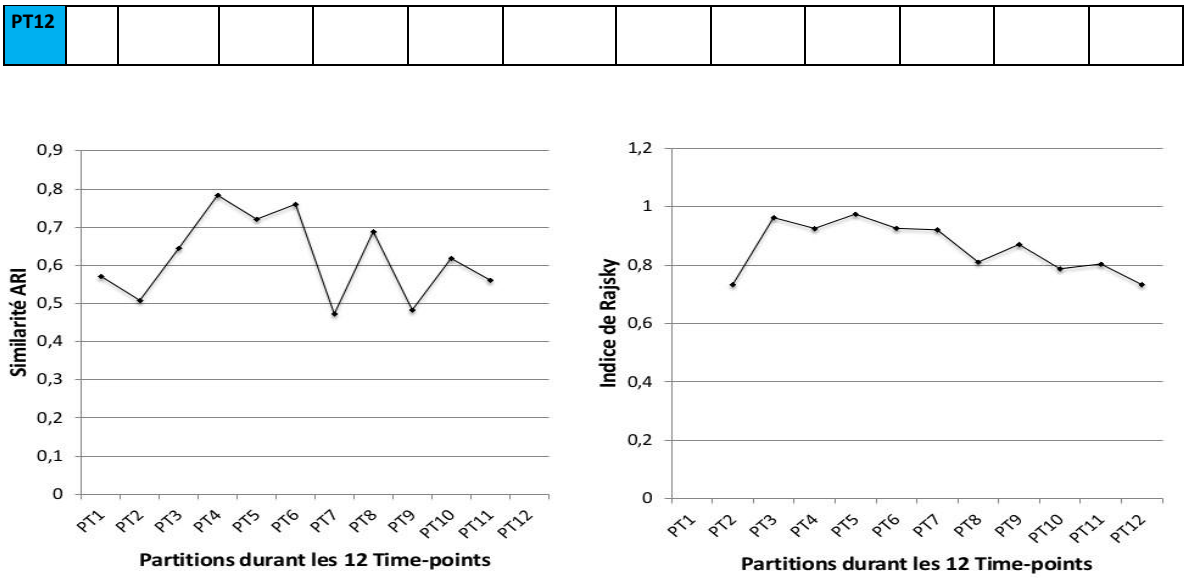


Figure 90. Rapport de similarité ARI et indice de Rajski entre chaque parie de partitions PT_i et PT_{i+1}

La **Figure 90** (à gauche) montre que les valeurs de ARI change dans le temps en remarquant ainsi des hauts et des bas dans la dépendance globale entre PT_i et $PT_{i+1.}$ L'autre courbe à droite (**Figure 90**) représente le changement de valeurs de l'une des versions de l'indice de Rajski qu'on utilise pour estimer le taux de préservation de l'information dans le temps. Un taux de préservation d'une information qui concerne ici la distribution des acteurs dans les groupes d'une partition PT_i par rapport PT_{i-1}. Il est remarquable que ce taux diminue lentement dans le temps. Ça veut dire que le les structures de groupes changent progressivement étape par étape dans le temps, ***sous une sorte de gouvernance qui ne laisse pas le SN se disloquer aléatoirement***.
Donc la liaison proposée dans TW-DAG exclusivement entre groupes qui se succèdent dans le temps nous permettra d'identifier et de suivre aussi des structures sous-jacentes pertinentes étape par étape dans le temps.

Identification d'une large composition qui persiste dans un chemin critique CP

Sur ce méta-modèle qui se présente par un réseau orienté avec des poids standards, on applique la méthode CPM via la commande 'Network Menu>Acyclic Network > Critical Path Method (CPM)', étant une opération spécifique aux graphes acycliques. Mais avant cela, on cherche une version moins stricte que CP. La **Figure 91** montre d'abord un simple chemin de T_1 vers T_{12} en 2×4 voies possibles, détecté par Search Path Count (SPC) qui s'applique souvent sur les réseaux de citations (Source to Sink). Ce sont juste des chemins couvrants. Cependant, le CP formé par t groupes (A_1, A_2, ..., A_t) et détecté par CPM n'est pas seulement couvrant mais probablement le plus lourd (**Figure 91**). Juste à côté, il y a la matrice des poids des arcs qui forment ce CP entre chaque T_i et T_{i+1}. Le poids global de ce CP est à 144.

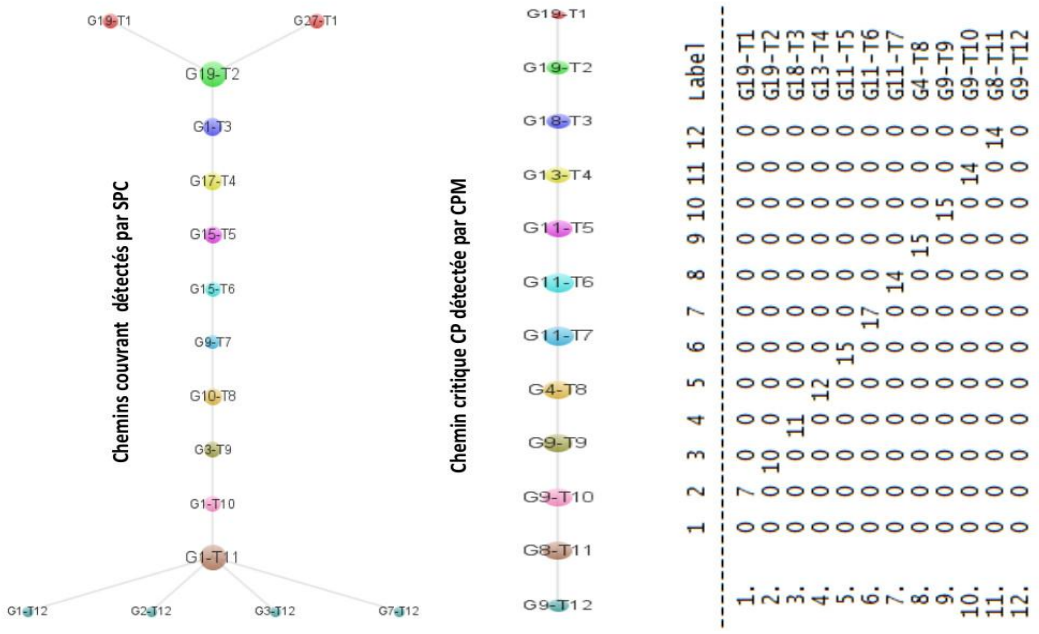


Figure 91. Chemins couvrants et chemin critique CP

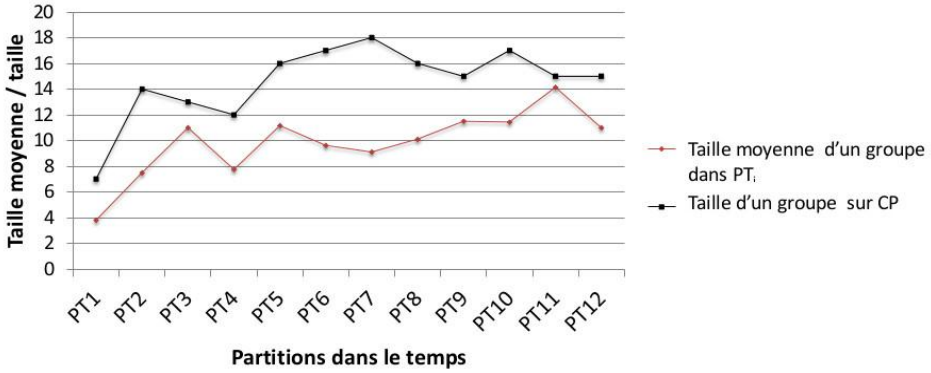


Figure 92. Taille de groupe couvert par CP chaque T_i devant la taille moyenne des groupes dans PT_i et devant

Il faut noter d'abord en observant la **Figure 92** que le CP couvre à chaque T_i un groupe dont la taille est toujours au-dessus de la taille moyenne des groupes de chaque PT_i. Par ailleurs, la **Figure 93** prouve que CP est un vrai chemin critique. La **Figure 93** affiche comment ces poids varient dans le temps et montrent qu'ils sont très proches des poids les plus lourds du TW-DAG standard entre chaque T_i et T_{i+1}. Sachant que l'arc le plus lourd reflète le plus grand chevauchement qui implique les 2 groupes les plus similaires, il est remarquable que la séquence CP couvre 6 fois les 2 groupes les plus similaires entre 2 partitions successives. Cependant, les poids sur un autre chemin couvrant (dont le poids global est à 114) semblent moins inférieures et notamment à partir de T_5 où l'écart se creuse considérablement.

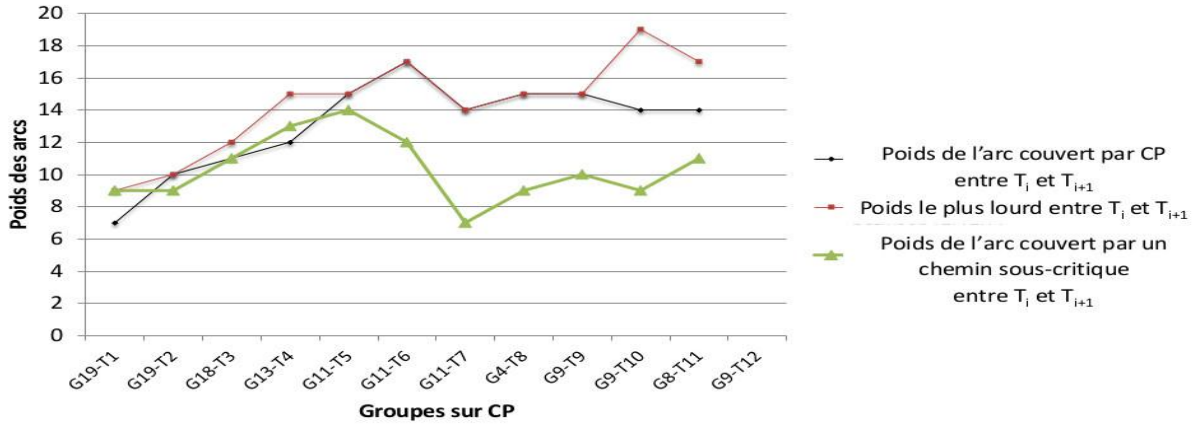


Figure 93. Variation des poids standards des arcs couverts par CP (en noir) par rapport les arcs les plus lourds (le plafond en rouge) entre T_i et T_{i+1} Un autre chemin (en vert) est ajouté à la comparaison

Avec ces poids standards, nous vérifions d'abord le paramètre de persistance s'il est assuré ou pas dans cette séquence CP. Ce pattern ne couvre pas seulement une séquence de grands groupes et grands chevauchements. Il encapsule aussi un regroupement stable (de 6 acteurs : 1261, 1142, 1144, 1157, 1265, 1174) qui persiste à l'intérieur de tous ces groupes et profondément dans leurs chevauchements temporels (**Figure 94**). Pour trouver cette structure, une séquence d'intersections (chevauchements) binaires hiérarchisés ont été effectuées (**Figure 94**), vue que Pajek n'autorise qu'une opération d'intersection à 2 paramètres (2 partitions) seulement (voir plus loin). En résultat, on obtient une composition inchangée, dont l'évolution se présente comme une couche profonde et stable au fil du temps (**Figure 95**). Elle incarne ***un premier caractère de résistance d'une structure noyau***. Cependant, les compositions (groupes et chevauchements) dans les couches supérieures sont plus dynamiques (**Figure 95**), considérés ainsi comme des éléments périphériques. La taille de ces chevauchements dans la **Figure 95** correspond aux poids affichés en noir dans la **Figure 93.**

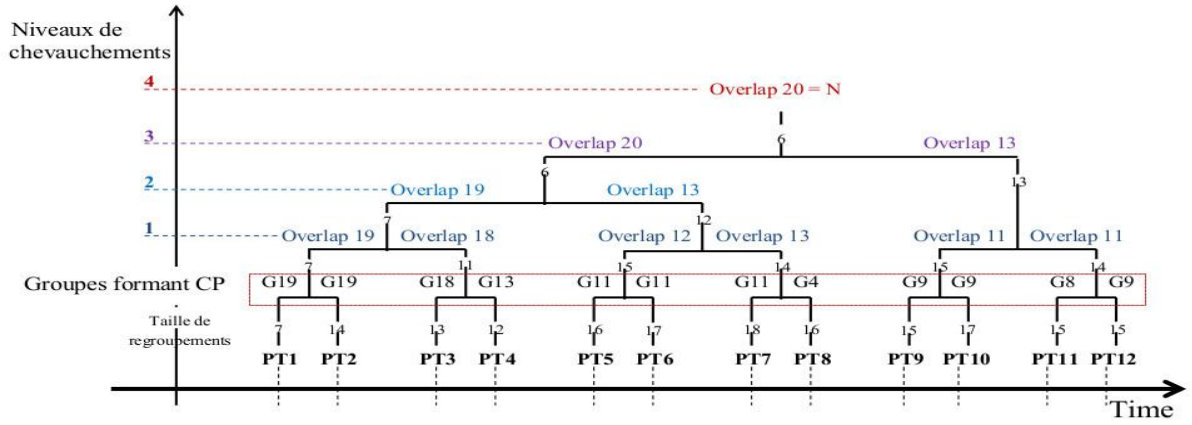


Figure 94. Schéma de chevauchements temporels couverts par le CP détecté, menant à révéler une structure persistante N pendant les 12 'time-points'

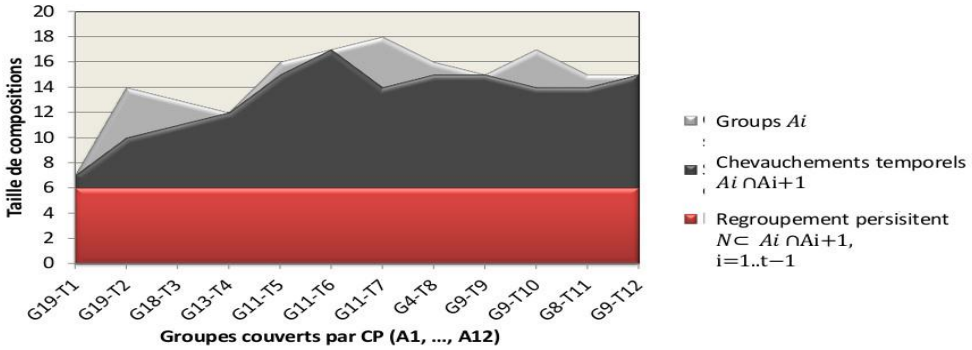


Figure 95. Taille des groupes et structures sous-jacentes couvertes par le CP dans le temps, visualisées en couches. La couche grise supérieure présente les tailles de groupe. La deuxième couche plus sombre présente les tailles de leurs chevauchements. La couche profonde, rouge présente la taille d'un regroupement stable qui persiste dans CP.

En appliquant notre schéma de pondération, les poids des arcs seront améliorés. Certains d'entre eux seront annulés, pénalisés ou suffisamment équilibrés pour représenter plus significativement la pertinence de ces structures sous-jacentes (chevauchements temporels, etc.).

Amélioration des poids de TW-DAG

Les poids (dans la matrice R) de TW-DAG sont affectés d'abord suivant la formule (1) de pondération qui évoque le calcul de centralité (GC) des chevauchements. Donc une nouvelle matrice de poids, appelée M1, contient une moyenne de centralité pour chaque chevauchement temporel qui a eu lieu entre T_i et T_{i+1}

Étant un sous-groupe qui persiste localement à T_i et à T_{i+1}, un chevauchement aura un GC qui se calcule en réduisant le SN. Le modèle réduit (Shrinking) du SN est tout d'abord bien adapté à nos hypothèses et utilisé pour le calcul de GC des groupes d'une partition PT_i. Sous Pajek, cette réduction se fait par la commande 'Operations>Network + Partition> Shrink Network' suivant une partition PT_i. Chaque groupe G_x-T_i est remplacé par un sommet, un 'super vertex' SV_{Gx} . Un lien entre 2 'supper vertices' (SV_{G1}, SV_{G2}) est pondéré par le nombre d'interactions $|\{(individu_1, individu_2) / individu_1 \in G_1\text{-}T_i, individu_2 \in G_2\text{-}T_i\}|$, c'est-à-dire le nombre d'interactions entre les individus de la frontière de G_1-T_i et celle de G_2-T_i. Par la suite, la centralité (GDC) de G_x-T_i sera équivalente à la centralité pondérée (Weighted Degree) de degré de SV_{Gi}.

Algorithme 6. Calculer le GC des groupes dans un SN réduit suivant une partition PT_i

```
POUR i = 1; i <= 12; i++ FAIRE
    Réduire le SN suivant PTi ← Operations > Network + Partition > Shrink Network
    Shrunk-Network à Ti with deleted loops
    POUR supper vertex SVGx qui représente Gx-Ti de PTi FAIRE
        GDC de Gx-Ti ← Centralité du supper vertex SVGx dans Shrunk-Network
```

L'opération donne techniquement une évaluation correcte dans le cas de GDC ou GCC. Mais ce n'est pas toujours le cas avec le GBC, tant que la cohésion interne du groupe (les liens intra-groupe) sont réduits en un seul point SV_{Gx}. Avec le même principe on calcule le GDC de chaque chevauchement obtenu, étant un groupe dans une partition transitionnelle P (T_i, T_{i+1}) à T_i et à T_{i+1}. Chaque P (T_i, T_{i+1}) est constituée en croisant PT_i avec PT_{i+1} sous Pajek, via la commande 'Partitions> Intersection of Partitions' (**Figure 96**). Les chevauchements sont numérotés (comme des groupes) dans P (T_i, T_{i+1}). En observant la matrice R, le résultat de

G_{41}-T_1 ∩ G_9-T_2 donne par exemple le chevauchement numéro 60: $O_{60}(T_1, T_2)$ dans le rapport d'intersection P (T_1, T_2), (**Figure 96**). Après avoir réduit le réseau suivant les regroupements P (T_1, T_2) et calculé leurs GDC à T_1 et à T_2 on obtient ainsi une moyenne de centralité pour chacun. La procédure se réitère pour tous les autres chevauchements (Algorithme 7). Cette moyenne remplacera les poids précédents de R, notamment les blocs en diagonal ce qui donne la matrice M1. D'où, le programme java (Algorithme 5) va mettre à jour les poids de TW-DAG

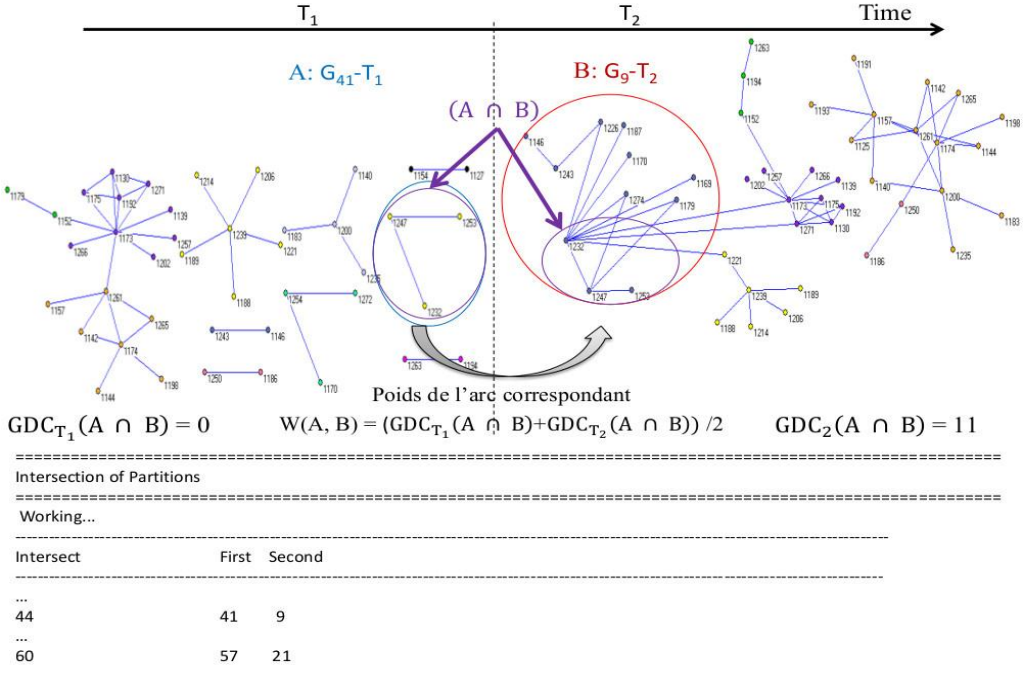


Figure 96. Poids (selon Formule (1)) d'un arc qui implique le GDC d'un chevauchement étant un groupe transitionnel d'une partition transitionnelle P(T_1, T_2)

Algorithme 7. Calculer la moyenne de GC des chevauchements dans un SN réduit suivant une partition P (T_i, T_{i+1})

```
POUR i = 1; i < 12; i++ FAIRE
    POUR Chaque cheveauchement temporel Oz–(Ti, Ti+1) étant un groupe dans P (Ti, Ti+1) FAIRE
        Shrunk-Network à Ti suivant la partition P (Ti, Ti+1) ← Operations >
        Network + Partition > Shrink Network

        Calculer le GC de Oz–(Ti, Ti+1) à Ti

        Shrunk-Network à Ti+1 suivant la partition P (Ti, Ti+1) ← Operations >
        Network + Partition > Shrink Network

        Calculer le GC de Oz–(Ti, Ti+1) à Ti+1

        Calculer la moyenne de centralité de Oz entre Ti et Ti+1

        Mise à jour du poids correspondant dans R qui donne la matrice M1
```

Dans la **Figure 97** nos chevauchements temporels successifs ont été distribués entre 2 axes. L'axe horizontal et vertical présente respectivement la taille du chevauchement (poids précédent) et son poids correspondant basé sur la Formule (1). La distribution prouve des cas particuliers dégagés de cette version de pondération, discutés précédemment. Par exemple certains regroupements formés par 3 individus ou plus, comme |G_9-T_3∩G_9-T_4|, ont un poids, une centralité moyenne nul, c'est-à-dire un GDC nul à T_i et T_{i+1} (**Figure 97**), ce qui correspond au cas '*Composition stable mais isolée*'. En outre, il y a des *chevauchements singletons*

(comme G_{15}-$T_5 \cap G_9$-T_6) qui jouent un rôle plus stratégiques que d'autres qui ne sont pas (ex. formé par 12 individus), (**Figure 97**), etc. Par conséquent, ces poids expriment ce paramètre de centralité, ils ne sont pas suffisamment équilibrés pour afficher la pertinence réelle de ces chevauchements. Donc le CP risque de louper des groupes et des structures sous-jacentes pertinents.

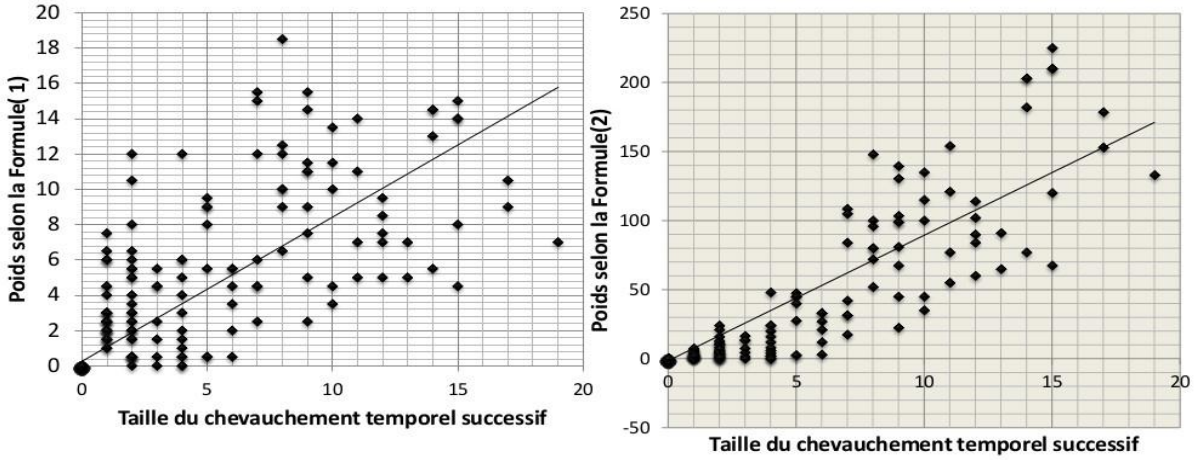


Figure 97. Chevauchements temporels successifs distribués en nuage de points selon les poids correspondant de la Formule (1) et 2 par rapport à leur taille

La Formule (2) de pondération est la première tentative pour régler ces cas. Les arcs de TW-DAG sont pondérés en se basant sur une nouvelle matrice de poids M2. C'est la mise à jour de M1 et R (blocs en diagonal) qui combine, à travers un script Java, la moyenne de centralité et la taille d'un chevauchement (G_x-$T_i \cap G_y$-T_{i+1}) entre T_i et T_{i+1} suivant la Formule (2), (Algorithme 8). Les matrices en entrée et sortie sont en .xls.

Algorithme 8. M2 une mise à jour des poids suivant la Formule (2) de pondération

```
POUR i = 1; i < 12; i++ FAIRE
    POUR Chaque cheveauchement temporel Gx-Ti ∩ Gy-Ti+1 de P (Ti, Ti+1) FAIRE
        W (Gx-Ti, Gy-Ti+1 ) = M2 (Gx-Ti, Gy-Ti+1) = R (Gx-Ti, Gy-Ti+1) × M1 (Gx-Ti, Gy-Ti+1)
```

Par ce moyen, il est remarquable que la distribution (un modèle de régression linéaire) des chevauchements dans la **Figure 97** (à droite) se rétrécie autour de la droite de régression. En intégrant le facteur taille, les poids donnent une estimation plus précise, particulièrement par rapport les petits chevauchements voir les singletons. Entre chaque T_i et T_{i+1}, la détection d'un chemin CP va chercher donc l'arc le plus lourd qui implique normalement chevauchement grand et central au même temps. Bien qu'ils semblent être plus équilibrés, ces nouveaux poids ont encore une lacune au niveau de la moyenne de centralité. Comme il est montré dans la **Figure 98**, par exemple un chevauchement comme G_{41}-$T_1 \cap G_9$-T_2 (**Figure 96**) affiche un GDC = 0 à T_1 alors qu'il est plus actif à T_2. Donc, entre T_i et T_{i+1}, CP peut couvrir un arc qualifié comme lourd alors qu'il implique ce type chevauchement temporel isolé à T_i ou à T_{i+1}. À ce moment-là, le regroupement persistant à l'intérieur de CP notamment dans ce chevauchement serait automatiquement isolé, et non plus pertinent pour présenter une identité d'un noyau.

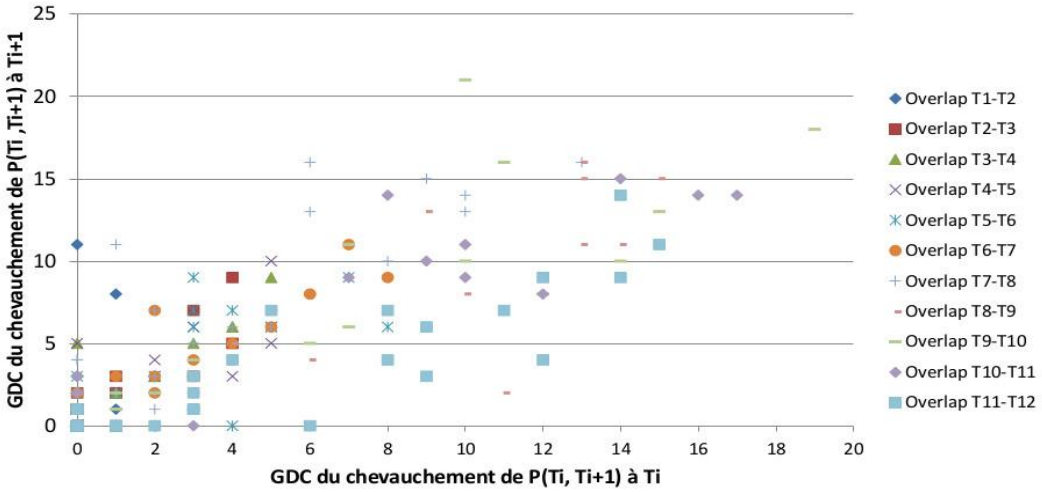


Figure 98. Chevauchements temporels successifs distingués par intervalle de temps et distribués en nuage de points entre leur GDC à T_i et à T_{i+1}

Néanmoins, en introduisant le facteur α, la Formule (3) de pondération permet d'annuler ces poids trompeurs impliquant ce type de chevauchement. D'où, une nouvelle matrice de poids, appelée M3, est déduit à partir M2 et Mα. Mα se construit d'abord par un programme Java qui s'exécute itérativement, en attribuant des valeurs binaires de α à chaque chevauchement selon son GDC entre T_i et T_{i+1} (Algorithme 9). Il prend en entrée les valeurs de M1 et deux vecteurs de centralité GDC pour ces chevauchements l'un à T_i et l'autre à T_{i+1} : 'VecteurGDCT$_i$ et VecteurGDCT$_{i+1}$' (en feuilles .xls). Ensuite M3 se construit (Algorithme 10) pour mettre à jour de la même manière les poids des arcs de TW-DAG, ce qui permet de couvrir des arcs lourds, plus significatifs par CP. La manipulation de toutes ces matrices se fait comme avant, au niveau des blocs en diagonal.

Algorithme 9. Déduire les valeurs binaires de α dans une matrice Mα

```
– Entrées:
– Matrice M1
– VecteurGDCTi qui contient le GDC de chaque chevauchements de P(Ti, Ti+1) à Ti
– VecteurGDCTi+1 qui contient le GDC de chaque chevauchements de P(Ti, Ti+1) à Ti+1
POUR i = 1; i < 12; i++ FAIRE
    POUR Chaque cheveauchement temporel Gx-Ti ∩ Gy-Ti+1 de P (Ti, Ti+1) FAIRE
        SI M1 (Gx-Ti, Gy-Ti+1) = 0 ALORS
            Mα(Gx-Ti, Gy-Ti+1) = 0
        SINON
            SI (VecteurGDCTi (Gx-Ti ∩ Gy-Ti+1) ≠ 0 )& (VecteurGDCTi+1 (Gx-Ti ∩ Gy-Ti+1) ≠ 0 ) ALORS
                Mα(Gx-Ti, Gy-Ti+1) = 1
            SINON
                Mα(Gx-Ti, Gy-Ti+1) = 0
```

Algorithme 10. M3 une mise à jour des poids suivant la Formule (3) de pondération

```
POUR i = 1; i < 12; i++ FAIRE
    POUR Chaque cheveauchement temporel Gx-Ti ∩ Gy-Ti+1 de P (Ti, Ti+1) FAIRE
        W (Gx-Ti, Gy-Ti+1 ) = M3 (Gx-Ti, Gy-Ti+1) = M2 (Gx-Ti, Gy-Ti+1) × Mα (Gx-Ti, Gy-Ti+1)
```

Selon la **Figure 99**, seulement les poids faibles ont été affectés (annulés) par le facteur α, car ils impliquent des chevauchements à faible centralité. Cela se confirme aussi dans la **Figure 98** où un chevauchement qui a un GDC = 0 à T_i affiche généralement un faible GDC à T_{i+1}, ou vice-versa. Il pratiquement est difficile de constater un écart considérable entre 2 points de temps

successifs. Dans ce cas, les poids annulés ne sont pas déjà assez lourd pour qu'ils soient couverts par CP. Cependant, avec un autre type de centralité comme GBC, ces poids trompeurs peuvent être plus lourds et α aura ainsi plus d'influence. Par exemple si GDC = 0 → GBC = 0 à Ti, mais après un faible GDC à T_{i+1} n'implique pas une faible GBC, à ce moment-là le chevauchement en question peut jouer un rôle d'intermédiaire plus significatif ('structural hole'/hollow').

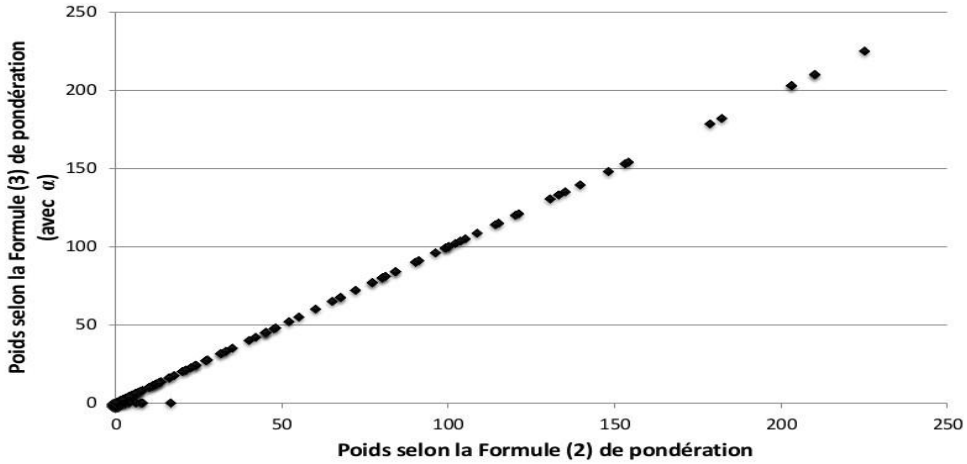


Figure 99. Poids de TW-DAG entre la Formule de pondération (2) et (3)

En tous les cas, CP assure la couverture des structures sous-jacentes plus pertinentes dont la centralité ne s'annule jamais à n'importe quel moment. Enfin, le paramètre de stabilité de centralité rentre en teste avec l'ultime amélioration de pondération (Formule(4)). Certains des poids de TW-DAG qui restent sont modifiés en intégrant le facteur β. Pour chacun des chevauchements précédents, le poids correspondant en M3 (en diagonal) est multiplié (pénalisé) par une valeur correspondante de Mβ ce qui nous donne une nouvelle matrice poids M4 (Algorithme 11). En effet, Mβ fournit un taux de stabilité de centralité β pour chaque chevauchement G_x-T_i ∩ G_y-T_{i+1} ayant α = 1. Un autre script Java est mis en place (Algorithme 12) pour calculer les valeurs de Mβ en prenant en entrée les valeurs de Mα ainsi que 'VecteurGDCT$_i$' et 'VecteurGDCT$_{i+1}$' (feuilles .xls).

Algorithme 11. M4 une mise à jour des poids suivant la Formule (4) de pondération

```
POUR i = 1; i < 12; i++ FAIRE
    POUR Chaque cheveauchement temporel Gx-Ti ∩ Gy-Ti+1 de P (Ti, Ti+1) FAIRE
        W (Gx-Ti, Gy-Ti+1 ) = M4 (Gx-Ti, Gy-Ti+1) = M3 (Gx-Ti, Gy-Ti+1) × Mβ (Gx-Ti, Gy-Ti+1)
```

Algorithme 12. Calculer les taux de stabilité de centralité de β dans une matrice Mβ

```
– Entrées:
– Matrice M1
– VecteurGDCTi qui contient le GDC de chaque chevauchements de P(Ti, Ti+1) à Ti
– VecteurGDCTi+1 qui contient le GDC de chaque chevauchements de P(Ti, Ti+1) à Ti+1
POUR i = 1; i < 12; i++ FAIRE
    POUR Chaque cheveauchement temporel Gx-Ti ∩ Gy-Ti+1 de P (Ti, Ti+1) FAIRE
        SI Mα(Gx-Ti, Gy-Ti+1) = 1 ALORS
            Mβ (Gx-Ti, Gy-Ti+1) =
            Mα(Gx-Ti, Gy-Ti+1) / (|VecteurGDCTi (Gx-Ti ∩ Gy-Ti+1) -VecteurGDCTi+1 (Gx-Ti ∩ Gy-Ti+1)|+1)
        SINON
            Mβ (Gx-Ti, Gy-Ti+1) = 0
```

La **Figure 100** présente une distribution uniforme des chevauchements temporels successifs entre le GDC à T_i et à T_{i+1}. Si le GDC est équilibré entre T_i et T_{i+1} alors l'amplitude ce centralité

CA du chevauchement G_x-T_i ∩ G_y-T_{i+1} sera nul, ce qui veut dire une centralité parfaitement stable. Donc, G_x-T_i ∩ G_y-T_{i+1} se place sur l'axe de symétrie (**Figure 100**). Par conséquent, le β correspondant est égal à 1(**Figure 101**) et le poids W (G_x-T_i, G_y-T_{i+1}) ne sera pas ainsi pénalisé et se positionne sur la ligne de régression rouge (**Figure 102**). Lorsque GDC_{Ti+1} > GDC_{Ti}, le cas par exemple du chevauchement G_{19}-T_1 ∩ G_{19}-T_2 (GDC_{T2} (G_{19}-T_1 ∩ G_{19}-T_2) = 8 > GDC_{T1} (G_{19}-T_1 ∩ G_{19}-T_2) = 1), G_{19}-T_1 ∩ G_{19}-T_2 flotte au-dessus de l'axe de symétrie (**Figure 100**) et dans le cas contraire (GDC_{Ti} > GDC_{Ti+1}), il se place au-dessous de cet axe. Autrement dit, son CA s'élargie, son β tend vers 0 (**Figure 101**) et le poids correspondant est en conséquence pénalisé en se positionnant sous la droite (**Figure 102**). On remarque que l'amplitude le plus large (CA = 11) correspond au point le plus loin de l'axe de symétrie dans la **Figure 100** (GDC_{Ti} = 10 et GDC_{Ti+1} = 21). Entre T_5 et T_6 le β ne dépasse 0.5 (**Figure 101**) mais avec le temps, il y a plus de chevauchements avec (β tend vers 1) des centralités plus stables.

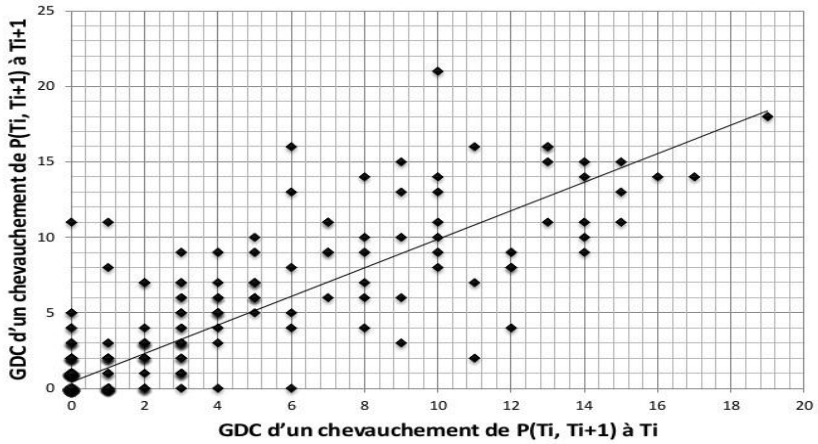


Figure 100. Distribution uniforme des chevauchements temporels successifs entre leur GDC à T_i et à T_{i+1}

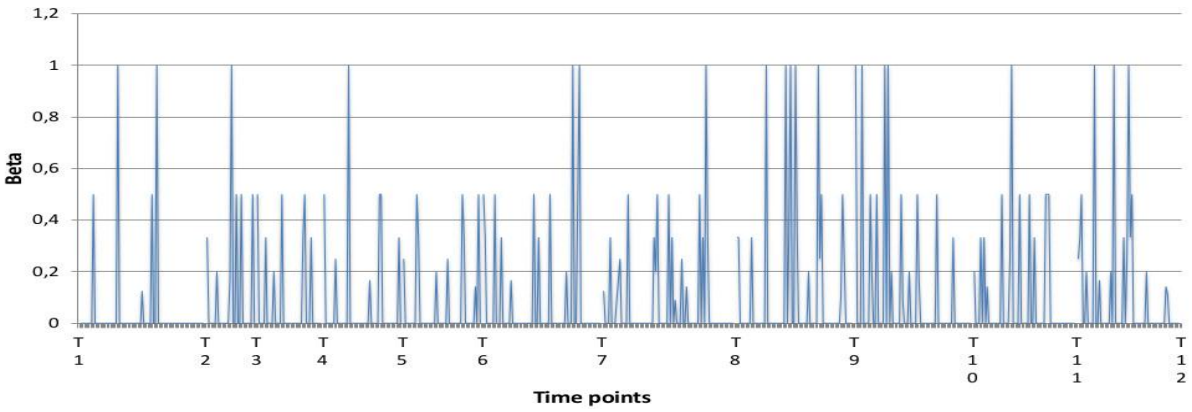


Figure 101. Taux de stabilité de centralité β des chevauchements entre T_i et T_{i+1}

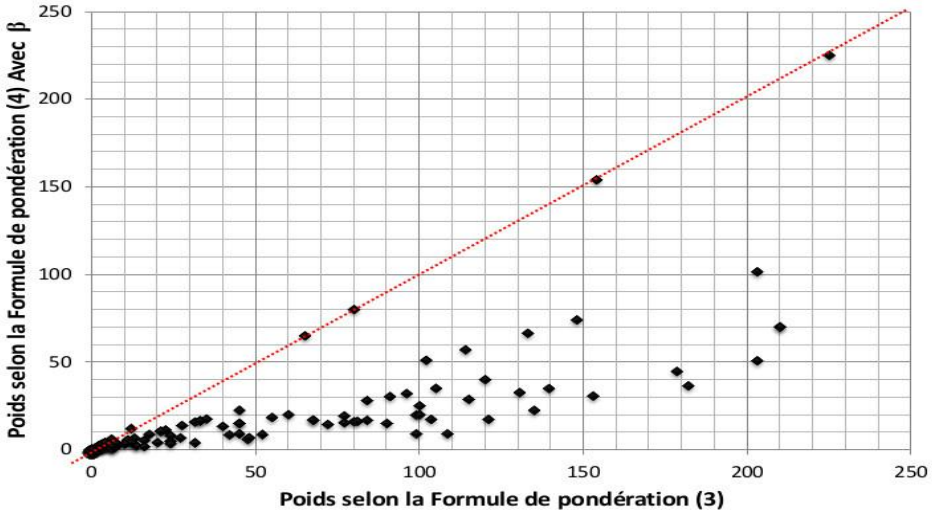


Figure 102. Poids de TW-DAG entre la Formule de pondération (3) et (4)

Beaucoup de poids (arcs) ont été annulés ou pénalisés (affaiblis), ce qui nous donne la dernière version du méta-modèle présentée dans la **Figure 103.** Les sommets (groupes) sont redimensionnés. Ceux qui sont impliqués dans les poids qui ont résisté le plus aux changements sont devenus plus grands et les autres sont plus petits (**Figure 103**). W (G_4-T_8, G_9-T_9) = 225 est finalement l'arc le plus lourd qui implique les deux plus grands sommets de TW-DAG : G_4-T_8 et G_9-T_9 (**Figure 103**). En plus, il n'a pas été pénalisé car G_4-T_8∩ G_9-T_9 a un GDC =15 stable à T_8 et T_9. C'est dans cet intervalle de temps qu'on trouve le plus grand nombre de chevauchements ayant des centralités plus stables (**Figure 101**).

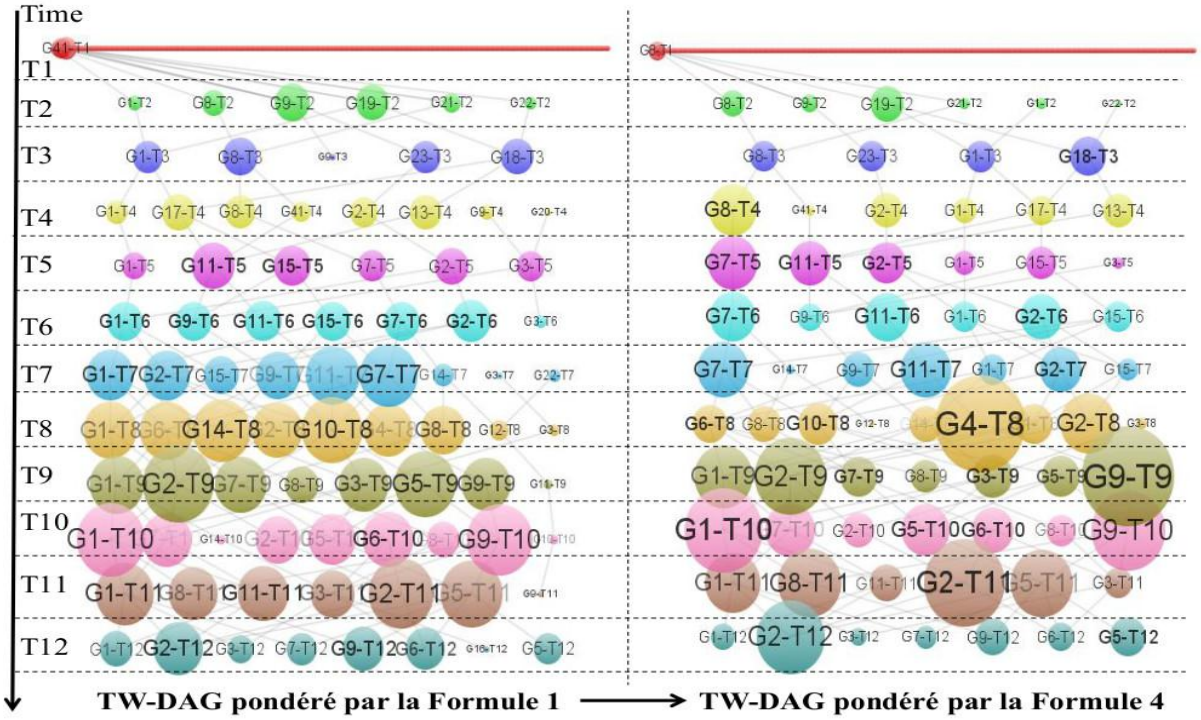


Figure 103. Le méta-modèle TW-DAG amélioré : de la Formule (1) de pondération 1 vers la Formule 4 (Generational/ layer view- VOSviewer)

Tableau 36. Bilan sur l'amélioration des poids du méta-modèle TW-DAG

Méta-modèle TW-DAG	Nombre d'arcs	Paramètres exprimés dans les poids (Localement entre Ti et Ti+1)	L'arc le plus lourd	Poids global de CP	Taux de corrélation entre les poids des arcs les plus lourds et ceux couverts par CP
TW-DAG Standard	1037 (143 entre PTi et PTi+1)	Stabilité de composition (taille du chevauchement)	W (G_5-T_{10}, G_5-T_{11}) = 19	144	0.8253
TW-DAG-Formula (1)	143	Centralité d'une composition stable locale (Moins de dépendance avec sa taille)	W (G_2-T_9, G_1-T_{10}) = 18.5	109.5	0.9484
TW-DAG-Formula (2)	143	Plus de dépendance et équilibre entre les paramètres	W (G_4-T_8, G_9-T_9) = 225	1517.5	0.9965
TW-DAG-Formula (3)	125	Stabilité de composition et centralité non-nulle: Des poids (faibles) sont annulés	W (G_4-T_8, G_9-T_9) = 225	1517.5	0.9965
TW-DAG-Formula (4)	125	Stabilité de composition et stabilité de centralité : Des poids pénalisés par β (Amplitude centralité)	W (G_4-T_8, G_9-T_9) = 225	614.02	0.8396

Paramètres affichés par les structures sous-jacentes dans CP

En se basant sur la combinaison de tous les paramètres, ces derniers poids de TW-DAG deviennent plus fins, permettant ainsi de détecter plus précisément des chevauchements temporels plus pertinents dans CP (A_1, A_2, ..., A_t). En effet, quelque-soit la formule de pondération, on a trouvé que CP (étant le chemin le plus lourd) couvre de 82% à 99% des arcs les plus lourds pendant toute la période d'observation (Tableau 36). La **Figure 104** confirme ce constat, notamment quand on remarque que l'arc le plus lourd W (G_4-T_8, G_9-T_9) est inclus dans cette séquence.

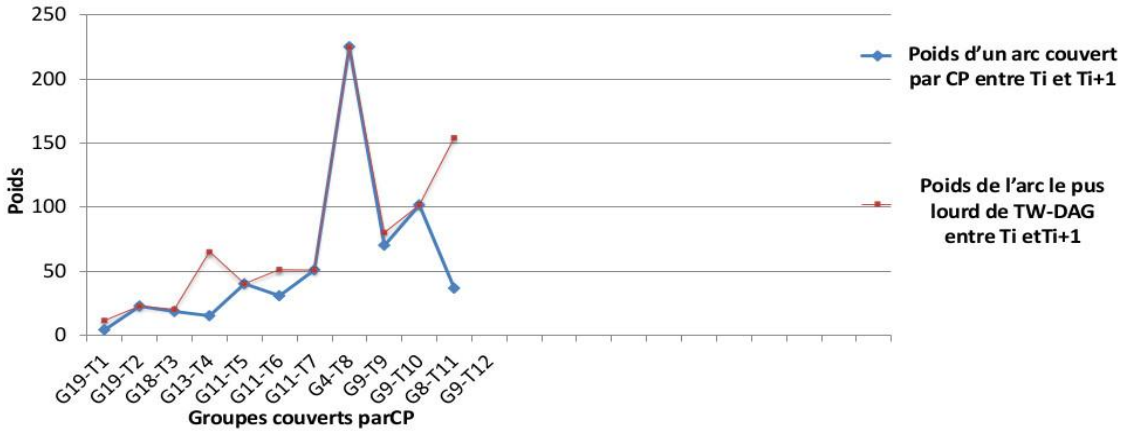


Figure 104. Poids des arcs les plus lourds (en rouge) entre chaque 2 points de temps successifs et ceux qui sont couverts par CP (en bleu). CP passe par 82% - 99% des arcs les plus lourds entre Ti et Ti + 1.

Dans ce cas, couvrir des arcs lourds révèle que les chevauchements ($A_i \cap A_j, \forall i = 1..11, j = i + 1$) et le regroupement qui persiste à l'intérieur sont susceptibles d'être grands, centraux et les plus stables possibles en termes de centralité. La **Figure 105** expose d'abord le paramètre de centralité affiché par les 3 couches dessinées dans la **Figure 95** (groupes, chevauchements,

regroupement persistant) dans le temps, par rapport le groupe le plus central à chaque T_i. La **Figure 105** montre que CP couvre effectivement des groupes centraux (A_1, A_2, ..., A_t). Leurs chevauchements entre T_i et T_{i+1} sont également centraux (avec des valeurs très proches). Plus particulièrement, G_4-T_8 ∩ G_9-T_9 est impliqué dans les pics de centralité entre T_8 et T_9. Il est à noter à la fin que le regroupement persistant à l'intérieur $N \subset A_i \cap A_j, \forall i = 1..11, j = i + 1$ affiche aussi dans la plupart du temps des scores plus élevés de centralité qui varient corrélativement avec ceux des couches supérieurs (Tableau 37). Selon un aperçu microscopique, la **Figure 106** confirme que la zone qu'il occupe le regroupement N et sa périphérie est quasiment la plus saillante et dense (rouge), au centre du SN à T_8 : Une région cohésive qui domine la communication dans le réseau.

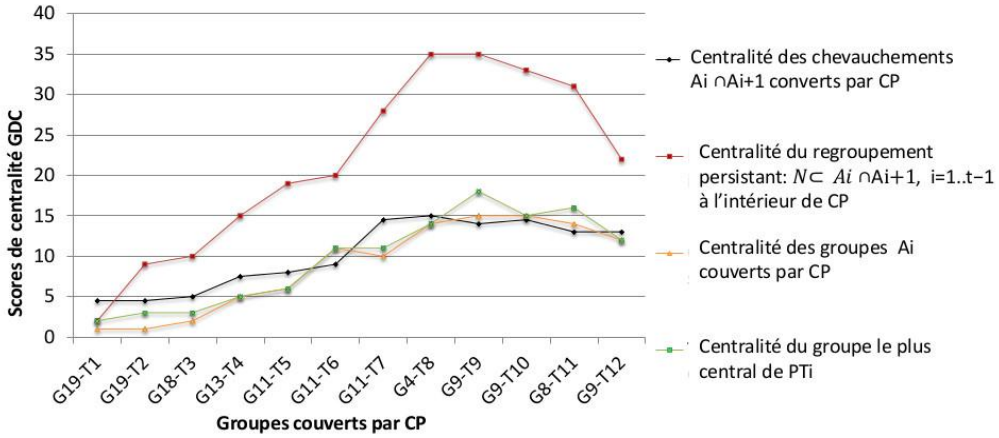


Figure 105. Évolution des scores de centralité des groupes et des structures sous-jacentes couverts par CP dans le temps. Courbe verte, orange, noire et rouge présente respectivement la centralité du groupe le plus central à chaque T_i, la centralité des groupes couverts par CP à T_i, la centralité du chevauchement temporel couvert par CP entre T_i et T_{i+1}, et la centralité du regroupement persistant N à T_i.

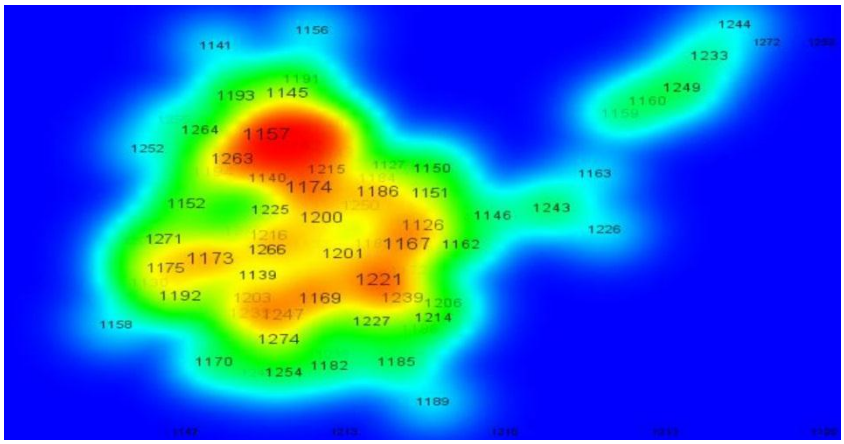


Figure 106. Vue sur le SN d'EEN à T_8 qui montre le regroupement N en tant qu'une région cohésive saillante plus active (Density View -VOSviewer)

Par la suite, la **Figure 107** montre à quel point la centralité de ces structures notamment le regroupement (N) qui persiste dans CP est stable. Pour cela, le β de chaque recouvrement inclus dans CP et le β du regroupement persistant à l'intérieur sont comparés par rapport au β le plus élevé entre chaque T_i et T_{i+1} (**Figure 107**). De nombreux chevauchements de CP ont une centralité plus ou moins stable. Pour certains le β = 1 (centralité parfaitement stable). C'est le cas par exemple de G_4-T_8 ∩ G_9-T_9, impliqué déjà dans l'arc le plus lourd. C'est un grand chevauchement avec un rôle central parfaitement stable localement entre T_8 et T_9. Pour

d'autres 0,2 <= β <= 0,5 (**Figure 107**). D'autre part, les valeurs β de $N \subset A_i \cap A_j, \forall i = 1..11, j = i + 1$ suivent quasiment le même rythme (avec un taux de plus de 95% (Tableau 37)).

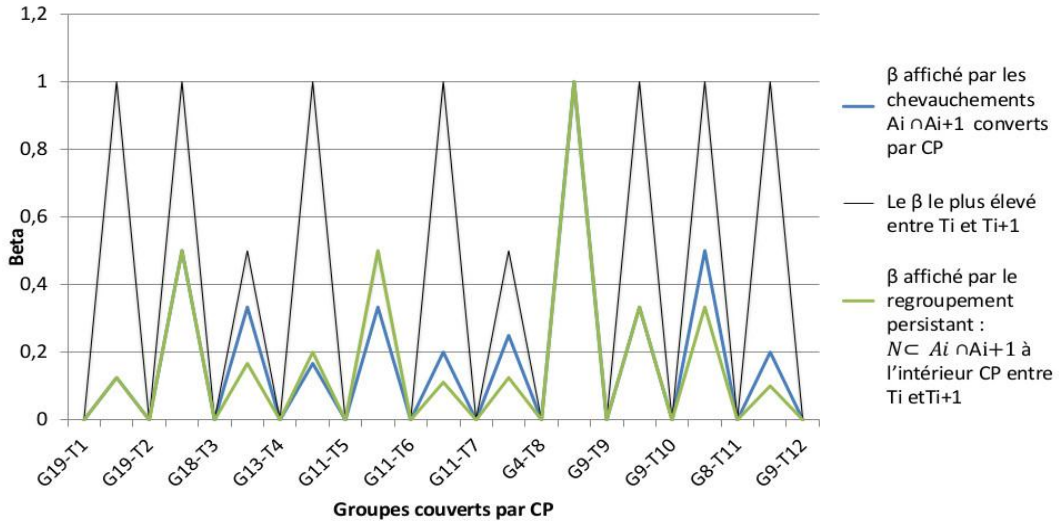


Figure 107. Valeurs β affichées par les chevauchements (en bleu) couverts par CP et celles du regroupement persistant (en vert) inclus à l'intérieur. Le β le plus élevé (en noir) entre Ti et Ti + 1 se réfère au chevauchement le plus stable en termes de centralité

Nous pouvons déduire que le pattern CP dans la dernière version de TW-DAG nous mène à détecter des chevauchements plus pertinents (grands et centraux) avec une centralité qui n'est pas toujours parfaitement stables. Cela confirme d'une part qu'il n'est pas évident de jouer un rôle central et le plus stable au même temps. D'autre part, les poids finaux ne sont basés sur un seul paramètre, mais une combinaison d'un ensemble de paramètres et contraintes. ***Ces structures transitionnelles pertinentes couvertes par CP, incarnent localement dans un intervalle de temps une trace d'une identité noyau***. Les résultats obtenus ainsi que les taux de corrélation (Tableau 37) entre les paramètres de ces structures et les paramètres affichés par le regroupement (N) qui persiste profondément à l'intérieur confirment que ce dernier semble avoir une cohésion similaire, une domination et une résistance, mais surtout, tout au long de la période d'observation. C'est la plus grande composition qui persiste et joue un rôle central, le plus stable possible dans le temps, faisant référence à une structure noyau, plus significative, plus raffinée.

Tableau 37. Taux de corrélation entre les paramètres affichés par les chevauchements couverts par CP et ceux affichés par le regroupement qui persiste à l'intérieur

N⊂ Ai∩Aj / Ai∩Aj de CP	**Taille et stabilité de composition de N**	**Centralité**	**Stabilité de composition & centralité**	**Centralité non-nulle (α de N)**	**Stabilité de centralité (β de N)**
TW-DAG Standard	|N| = 6	/	/	/	/
TW-DAG-Formula (1)	|N| = 6	0.9508	/	/	/
TW-DAG-Formula (2)	|N| = 6	0.9508	0.9785	/	/
TW-DAG-Formula (3)	|N| = 6	0.9508	0.9785	$N \subset Ai \cap Ai+1 \subset CP$→corresponding α = 1	/
TW-DAG-Formula (4)	|N| = 6	0.9508	0.9785	$N \subset Ai \cap Ai+1 \subset CP$→corresponding α = 1	0.9589

Sensibilité du SN par rapport à une structure noyau

La dernière étape pour valider le calibre de cette identité significative de structure noyau, consiste à tester à quel point le SN est sensible à un tel regroupement (N) par rapport à d'autres régions persistantes dans le temps. À cette fin, nous essayons d'abord d'identifier des structures différentes (en termes de composition) ayant des prospérités semblables, notamment en termes de persistance. Tout d'abord, on enlève le pattern CP, un sous-ensemble sommets (groupes) de TW-DAG (**Figure 108**). Les conséquences sont affichées dans le Tableau 38. En appliquant une recherche de chemins critiques sur ce TW-DAG affecté (**Figure 108**)), nous obtenons un autre pattern appelé CP(1), couvrant et détecté comme le plus lourd, tel que W (CP(1)) < W(CP) (Tableau 38). Toutefois, on a trouvé que le CP(1) comprend un regroupement noté N1 formé par 5 individus, qui persiste jusqu'à T_{11} seulement, après sa trace est perdue dans le dernier recouvrement G_2-$T_{11} \cap G_2$-T_{12} (**Figure 108**). Autrement dit, même s'il est couvrant et lourd, CP(1) ne vérifie pas le critère de durabilité (Persistance). De ce fait, CP(1) est modifié en remplaçant (déviation) W (G_2-T_{11}, G_2-T_{12}) par W (G_2-T_{11}, G_6-T_{12}), (**Figure 108**)., ce qui nous donne le chemin P tel que W(P) = 282.13 (Tableau 38). Il est vrai que W (P) < W (CP (1)), mais P assure que la structure N1 encapsulée à l'intérieure, persiste tout au long des 12 points de temps. Suite à cette altération, un autre pattern critique noté CP(2) est détecté, là où on trouve un seul individu (étant une structure N2) qui persiste à l'intérieur (Tableau 38).

Tableau 38. TW-DAG impacté en supprimant ou en affectant des chemins critiques pour trouver autres structures persistantes

	TW-DAG Non affecté	**Après la suppression de CP**	**En évitant W (G_2-T_{11}, G_2-T_{12})**	
Nombre des arcs	125	102	101	
Arc le plus lourd	W (G_4-T_8, G_9-T_9) = 225	W (G_2-T_{11}, G_2-T_{12}) = 154	W (G_1-T_9, G_1-T_{10}) = 80	
Poids des chemins (CP) détectés	W (CP) = 614.02	W (CP(1)) = 418,88	W (CP(2)) = 329.2	CP(1) est converti à P W(P)=282.13
Taille de la structure qui persiste dans le chemin	|N| = 6	0	|N2| = 1	|N1|=5

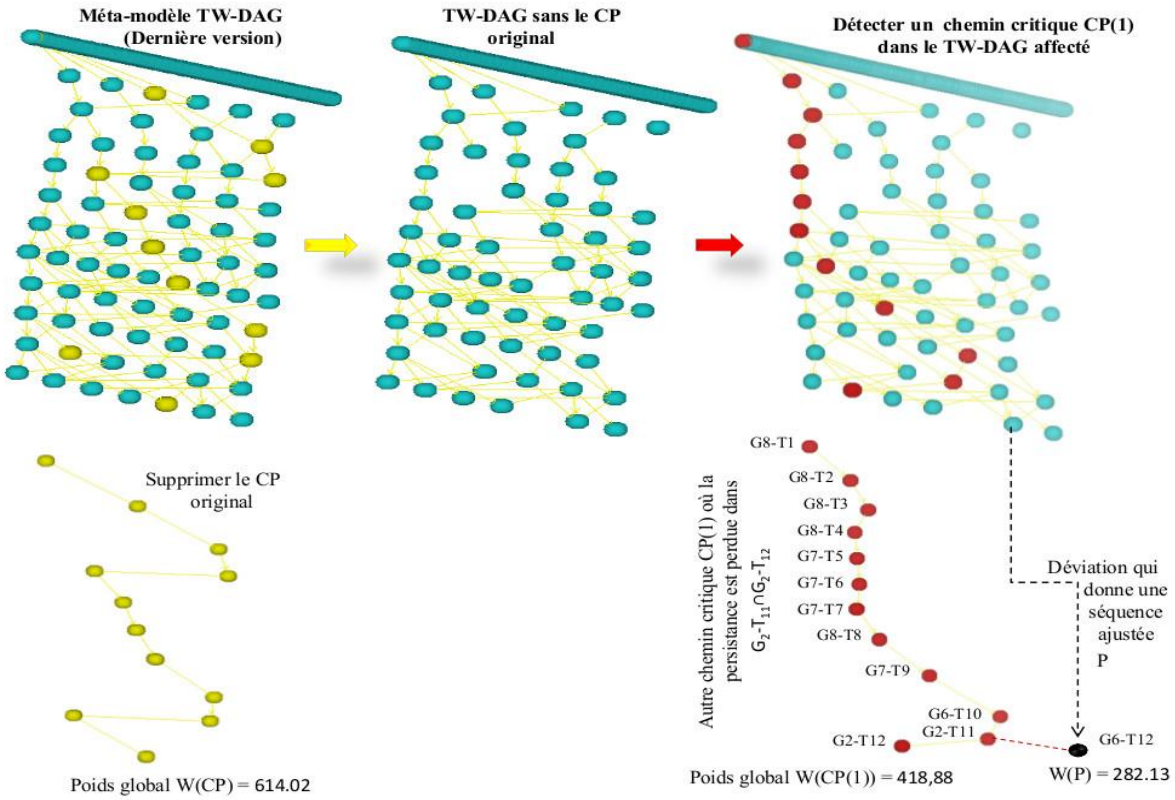


Figure 108. TW-DAG sans le CP original et la détection d'autres patterns CP(1) et P

Par conséquent, CP, P et CP(2) nous donnent trois structures persistantes différentes N, N1 et N2, respectivement. La **Figure 109** prouve que N2 est l'illustration d'une petite structure (regroupement singleton) qui a une faible centralité quasi-stable, alors que N1 a plus d'individus et affiche une centralité beaucoup plus supérieure mais plus ou moins stable dans le temps (**Figure 109**). D'autre part, pour une structure élite susceptible de présenter une identité noyau du SN, N et N1 sont plus grandes plus cohésives. Cependant, le regroupement N encapsulé dans CP joue souvent le rôle le plus central (**Figure 109**) qui ne se stabilise pas pour longtemps mais a plus d'influence sur la centralisation globale du SN, comme le montre la **Figure 110.** Nous avons mis le SN dans deux situations différentes et indépendantes en comparant le changement de ses caractéristiques : nombre de liens, densité, CC, notamment sa centralisation globale qui est calculée via la commande 'Network > Create Vector > Centrality > Betweenness' sous Pajek à chaque T_i. Dans le cas normal, la centralisation d'intermédiarité du SN évolue normalement au même temps où le réseau se développe dans le temps (courbe verte, **Figure 110**). Elle continue à évoluer et affiche parfois des scores plus élevés bien que la structure N1 est retirée du réseau (courbe bleue, **Figure 110**). Cependant, quand le SN évolue sans la structure critique, noyau N, sa centralisation commence à chuter remarquablement à partir T_5 et atteint ses valeurs les plus faibles entre T_8 et T_9. Cela prouve que cette structure persistante révélée comme une identité noyau joue le rôle le plus central en tant qu'un groupe, sa connectivité externe tend à dominer progressivement le réseau dans le temps. Par ailleurs, sa composition cohésive la plus grande qui résiste est d'autre part cruciale pour dominer et préserver sa dominance pendant la période d'observation.

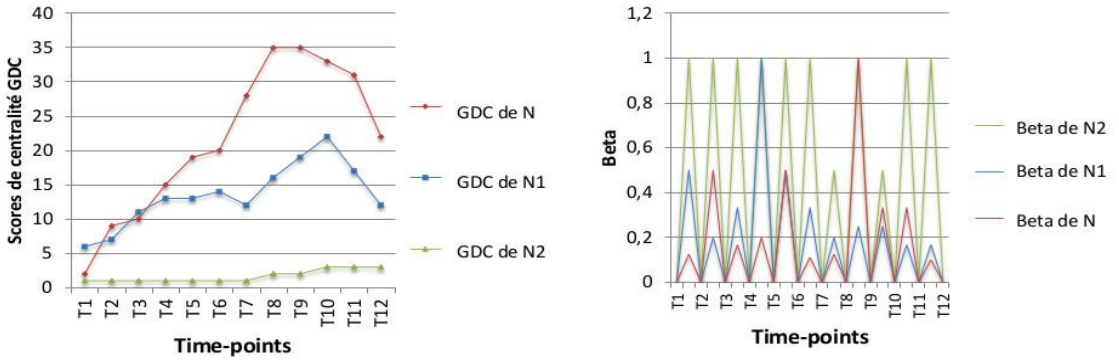


Figure 109. Centralités et stabilité de centralité des structures qui persistent dans CP, P et CP(2)

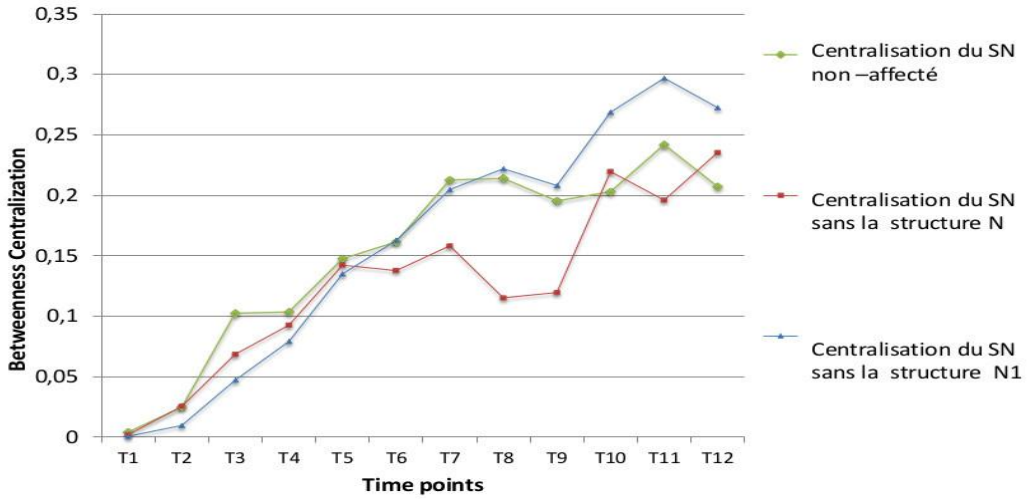


Figure 110. L'influence de la structure critique N (noyau) sur la centralisation d'intermédiarité du SN dans le temps par rapport à une autre structure persistante N1

Discussion

Les expérimentations ont montré que le méta-modèle TW-DAG, en particulier sa fonction de pondération ainsi que ses patterns critiques CP sont bien adaptés pour représenter respectivement le SN étudié d'EEN et explorer ses structures sous-jacentes vers une identité noyau plus significative, évoluant dans le temps. TW-DAG possède quelques points en commun avec le modèle proposé en (Berger-Wolf & Saia 2006) dans l'utilisation des groupes et l'exploitation des liens entre eux dans le temps. Mais la différence est flagrante dès la formation de ces groupes qui sont des groupes de femmes collectés par évènement dans (Berger-Wolf & Saia 2006) où l'ordre de ces groupes est l'ordre des évènements (pas de partitions dans le temps). La différence se trouve aussi dans la liaison que nous avons proposé qui se basent sur des chevauchements temporels successifs, dans les poids, etc.

Après avoir découvert des régions cohésives émergeantes sous forme des groupes en partitions PT_i, nous sommes arrivés à 1037 chevauchements entre PT_i et PT_j tel que $i < j$. Mais seulement, les chevauchements temporels successifs entre PT_i et PT_{i+1} sont exclusivement mis en évidence pour la création et la pondération des arcs de TW-DAG (Tableau 36). Nous avons constaté que G_5-$T_{10} \cap G_5$-T_{11} est le plus grand chevauchement (19 individus), bien qu'il ne soit pas le plus central. Son poids selon la formule $W_{Formule(1)}$ (G_5-T_{10}, G_5-T_{11}) = 7. En améliorant les poids, G_2-$T_9 \cap G_1$-T_{10} semble être le chevauchement le plus central, $W_{Formule(1)}$ (G_2-T_9, G_1-T_{10}) = 18.5, mais pas le plus grand (8 individus). Cependant, les figures : **Figure 95**, **Figure 104**, **Figure 105**, **Figure 107**, ont montré que G_4-$T_8 \cap G_9$-T_9 est le plus

pertinent, le plus équilibré, impliqué dans l'arc le plus lourd qui n'a pas été annulé, ni pénalisé $W_{\text{Formule (2/3/4)}}$ (G_4-T_8, G_9-T_9) = 255, et qui est en plus couvert par CP. Entre, T_8 et T_9, les paramètres de centralité, de stabilité de centralité β atteignent les valeurs les plus élevées.
CP a été montré comme une infrastructure dorsale du SN dans le temps. Ce pattern passe à chaque fois par la plupart (82%-99%) des arcs les plus lourds (Tableau 36), les chevauchements les plus pertinents entre T_i et T_{i+1}, mais il ne sera pas critique à moins qu'un regroupement N persiste à l'intérieur tout au long de la période d'observation. De ce fait, les tests de convergence, plus profonde, basée sur une architecture en couches expliquent pourquoi les paramètres de la structure N sont corrélativement similaires aux paramètres de ces recouvrements qui l'encapsulent profondément dans CP (**Figure 111**). Par exemple, la corrélation au niveau des valeurs de centralité et même de stabilité centralité, atteint 95% - 97% (Tableau 37). En d'autres termes, dans une succession de chevauchements temporels grands et centraux, le regroupement qui résiste à l'intérieur montre qu'il est le plus équilibré en tant qu'une grande composition qui joue un rôle central, le plus stable possible au fil du temps (**Figure 111**). Donc, il rassemble les concepts de cohérence, de domination et de résistance qui caractérisent une identité de structure noyau (**Figure 111**). De l'autre côté, un tel noyau aurait probablement causé ces groupes qui se chevauchent, ces chevauchements pertinents et ainsi ce CP, de telle sorte que les couches supérieures présentent une périphérie significative qui évolue autour de ce noyau dans le temps (**Figure 111**). Pour déterminer sa véritable envergure, nous avons constaté aussi que le réseau tend à être de plus en plus sensible à cette structure critique N dans le temps, par rapport à d'autres régions persistantes. Dans l'absence de N, le réseau est montré beaucoup plus impacté, notamment entre T_8 et T_9 (**Figure 110**). C'est exactement le moment où la structure N, est plus performante en termes de centralité et de stabilité de centralité (**Figure 105**, **Figure 107**).
Finalement, étant une somme des poids des arcs qu'il couvre, le poids global d'un chemin critique peut être aussi amélioré et renforcé pour éviter éventuellement l'ambiguïté qui se procure avec la détection de plusieurs patterns critiques et classes élites similaires. Le poids peut inclure par exemple les paramètres globaux de chaque regroupement qui persiste à l'intérieur, notamment sa stabilité de centralité basée sur l'amplitude globale de centralité (OCA). Par exemple, $W'(\text{CP}) = \frac{\frac{\sum_{i=1}^{t} GDC_{T_i}(N)}{t} \times |\text{N}| \times \text{W(CP)}}{OCA(N)+1}$, une version généralisée du poids. Par conséquent, W(P) aurait été plus important que W(CP (2)), (Tableau 38), alors que CP l'original reste toujours le plus lourd. Cela confirme une autre fois que la recherche à base de patterns critique dans TW-DAG conduit efficacement à identifier une telle classe élite.

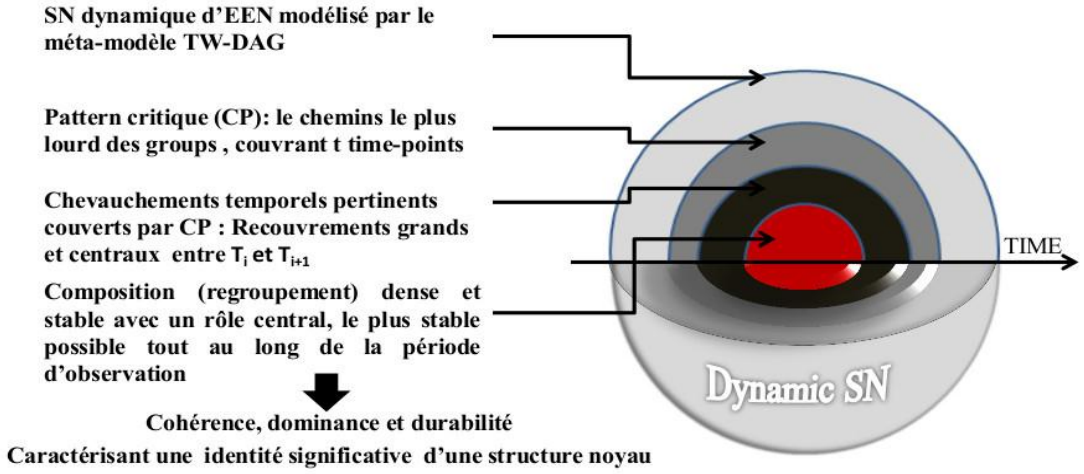


Figure 111. SN dynamique d'EEN présenté par le méta-modèle TW-DAG dans une architecture en couches: CP, chevauchements temporels pertinents, structure de noyau profondément à l'intérieur

8. Conclusion

Devant la dynamique des SNs, l'approche et les expériences proposées comportent une dimension temporelle pour caractériser et identifier une identité plus significative de structure noyau. Sa cohésion, durabilité et dominance sont examinés à travers des paramètres, dérivés, issus de la dynamique des groupes: Persistance, centralité de groupe, stabilité de centralité. Le SN étudié est représenté par TW-DAG, un méta-modèle sous forme de patterns de groupes qui se chevauchent, liés par des arcs pondérés. Bien que les théories des sciences comportementales soient complexes, les chevauchements temporels, la fonction de pondération et la recherche fondée sur les patterns critiques ont montré des atouts majeurs pour détecter les comportements typiques des structures sous-jacentes. Les résultats obtenus ont montré que le CP couvre une succession de chevauchements pertinents (larges et centraux). Ils confirment également que lorsqu'un regroupement persiste profondément à l'intérieur, il a tendance à dominer et régner la centralisation du réseau et préserver de manière la plus stable possible ce statut dans le temps. Il s'agit d'une structure sous-jacente qui reflète une identité plus significative, plus raffinée d'un noyau (une classe élite) qui évolue profondément à l'intérieur d'un processus évolutionnaire.

D'autre part, la complexité du sujet étudié nous laisse introduire plusieurs facteurs et hypothèses qui peuvent influencer ces résultats et leur donnent plus de variabilité : On parle de la consistance/inconsistance de l'ensemble des acteurs/employés du SN, nombre des échanges entre ces employés à considérer pour pondérer les relations, résolution de la taille des fenêtres de temps, comment définir le concept du groupe et la méthode de partitionnement, s'intéresser exclusivement aux chevauchements temporels successifs, omettre les groupes singletons en créant les arcs du méta-modèle, se baser sur d'autres indices pour calculer la similarité entre les groupes dans le temps, définir un seuil de similarité pour créer les premiers arcs. On ajoute aussi les différentes hypothèses qui concernent l'évaluation et le choix de centralité des groupes/ sous-groupes, en la donnant une dimension temporelle, jusqu'à le réglage du poids total de CP, etc. Mais le premier élément à prendre en compte est que les besoins en matière de sécurité et dissimulation d'un côté, et la recherche de l'efficacité la tâche et la coordination d'un autre côté, peuvent centraliser ou de décentraliser les réseaux, et même masquer les classes élites et dominantes.

La pertinence du contexte (quels acteurs ? pour quel usage et objectif ?, les liens sont-ils pertinents, etc.) et la richesse des données sociales qu'on a utilisé par exemple, contribuent à

des modélisations plus réalistes, renforcer significativement les résultats obtenus et multiplier le gain informationnel. Dans notre cas d'étude, la détection d'une telle classe élite dans le réseau de communication de l'entreprise d'Enron qui évolue dans le temps, semble très prometteuse pour des investigations qui s'intéressaient au scandale d'Enron. Chercher une structure noyau plus améliorée est de plus en plus stratégique et son interprétation varie selon différents SNs/ OSNs qui évoluent notamment dans des contextes institutionnels, dans des réseaux illégaux qui cachent des comportements frauduleux ou criminels, pour la lutte anti-terroriste, dans les réseaux de directions imbriquées '*interlocking directorates*', réseaux politiques ou religieuses, mouvements sociaux, en épidémiologie, les réseaux P2P ((Ceyhan et al 2011)), etc.

Notre approche touche plusieurs concepts et aspect dans la littérature. Donc, les conclusions obtenus peuvent enrichir certaines perspectives sur les aspects de durabilité et la sensibilité/ fragilité du SN dans le temps (ex. dans le contexte épidémiologique), et être étendues sur des réseaux à grande échelle et pendant des périodes d'observation plus longues. Cependant, il est à noter que plus l'échantillon est grand, une approche proposée et ses résultats auront une tendance plus statistique, en diminuant la capacité des chercheurs à sonder profondément dans le réseau. Donc, nos études ont exploré le SN en profondeur, en suivant des fondements et une généralisation conceptuels plutôt que statistiques.

Tendances sémantiques vers des nouvelles problématiques et motivations

Au-delà de ces théories mathématiques qui sont définies sur des plans topologiques, la considération des aspects sémantiques peuvent aider à améliorer la détection d'un noyau. Pour enchainer, il faut se rappeler que la dynamique temporelle du SN est animée par des individus qui ont des orientations implicites et des sentiments d'appartenance à des collectivités, etc. Suivant ces aspects et motivations cachés, ils peuvent changer ou maintenir leurs intérêts, leurs relations, menant à créer des régions qu'on détecte comme cohésives, ayant parfois cette tendance à devenir durable et de dominer le réseau au fil du temps (**Figure 112**). Donc, le contexte dynamique semble être un iceberg (**Figure 112**), dont les racines sont sémantiques souvent implicites et supervisent tous les comportements dynamiques des individus, des groupes, des structures sous-jacentes (comme la durabilité) et qui ont été expliqués sur un plan structurel, topologique. De ce fait, les perspectives qui vise l'exploration par exemple d'un caractère sémantique pour une structure noyau nécessiterons un niveau d'abstraction plus élevé, inspiré de la richesse sémantique des entités sociales, leurs collectivités, la sémantique de leurs interactions, leurs intérêts, etc. Cependant, toute analyse ou une fouille multidimensionnelle qui fusionne la sémantique et la dynamique temporelle est susceptible d'être de plus en plus complexe, et implique un coût de calcul massif.

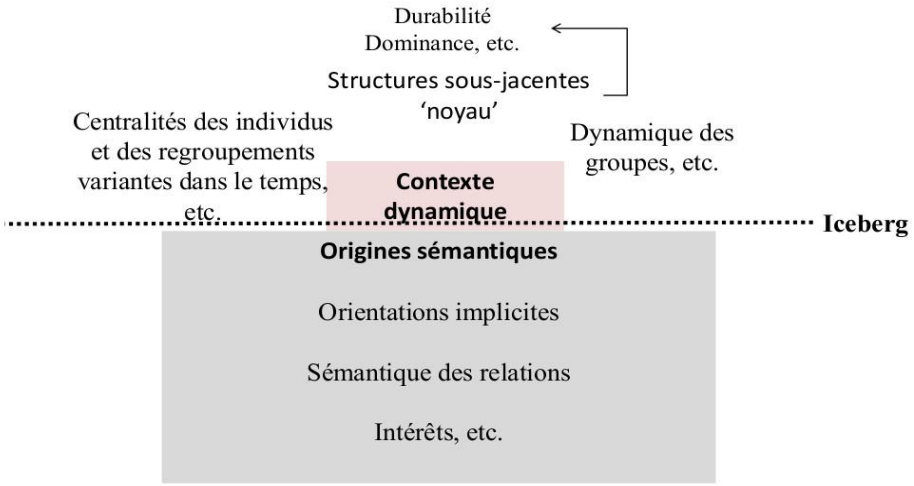


Figure 112. Dynamique temporelle du SN, un iceberg dont les origines sont sémantiques

Une sémantique implicitement inspirée d'une dynamique topologique

En général, même les représentations des données rigoureuses et les modèles mathématiques (graphes), ont un certain niveau de sémantique qui dépend du degré d'expressivité de la syntaxe suivant laquelle ces données sont organisées. La preuve est qu'un même concept défini dans deux représentations différentes, peut se référer à une certaine sémantique dans la première mais pas dans la deuxième. Par exemple ((Traub et al 2010)) ont présenté une collection de textes décrivant une source donnée par un modèle de graphe qui a été qualifié comme graphe sémantique. Deux textes (nœuds) sont liés par un arc pondéré via une mesure de similarité. En effet, cette mesure binaire donne une similarité en termes de chevauchement dit sémantique ((Traub et al 2010)), entre deux textes qui se chevauchent en termes de mots, et de son tour la mesure est également considérée comme mesure sémantique ((Traub et al 2010)). Pour les auteurs, une mesure de centralité associé à chaque texte est aussi sémantique

dans ce cas ((Traub et al 2010)). Cependant, et avec une syntaxe quasiment la même, le méta-modèle TW-DAG qui représente un niveau de dynamique temporelle du SN, a été défini en tant qu'un modèle topologique. En le comparant avec les données textuelles topologiquement présentées par un modèle de graphe en ((Traub et al 2010)), on constate que les deux représentations sont basées sur les mêmes concepts topologiques: nœuds, arcs, poids et chevauchements (Tableau 39). Malgré ça, le graphe des données textuelles a été considéré comme une représentation sémantique ((Traub et al 2010)).

Tableau 39. Comparatif entre le méta-modèle du processus évolutionnaire d'un SN dynamique (TW-DAG) et un graphe 'sémantique' des données textuelles

~~Représentation~~ Concepts	Méta-modèle TW-DAG dynamique topologique	Graphe sémantique de données textuelles
Nœuds	Groupes de partition à chaque 'time-point'	Collection de textes
Arcs	Connectant 2 groupes de deux time-point successifs, qui se chevauchent	Connectant 2 textes qui se chevauchent
Chevauchements	Intersection entre les membres de de 2 groupes dans le temps	Entre les mots de 2 textes (chevauchement 'sémantique')
Pondération	Fonction de pondération incluant la taille de chevauchement étant une mesure de similarité	Pondéré par une similarité 'sémantique' basée sur chevauchement 'sémantique'

Indépendamment de la dynamique temporelle ou la sémantique exprimée, les deux représentations ont la même topologie basées essentiellement sur les chevauchements, même si les composantes sont de nature différente : Groupes formés par des entités sociales devant des textes formés par des mots. Cette illustration montre que le graphe étant un modèle topologique où les données sont assez explicites, peut cacher des informations implicites (sémantiques). Avec les modèles des SNs, la sémantique existe et s'incarne par exemple derrière les causalités des connexions, la formation des équipes implicite (Nettleton 2013), etc., ainsi que les comportements dynamiques des entités sociales et des collectivités comme la durabilité, architecture en couches des structures sous-jacentes, le cas de notre méta-modèle, jusqu'à la sémantique d'une structure noyau.

Si les paramètres de persistance et de stabilité montrent topologiquement la durabilité d'une collectivité, Il peut y avoir d'autres arguments implicites justifiant ce comportement. Cela peut être clarifié en parlant par exemple de la dimension sémantique d'une identité de collectivité (Citoyenneté et de l'identité collective en Europe) dans le contexte politique, déterminée sur deux orientations ((Karolewski 2009)) : La première est une orientation horizontale fondée sur le sentiment d'appartenance qui se montre par les relations entre les membres d'un groupe et leur maintien dans une dynamique interne ((Karolewski 2009)). C'est le côté sémantique implicite manifesté par une dynamicité temporelle topologique. Une entité sociale a des orientations internes pour préserver ou pas son affiliation, d'où ils sont issus les effets topologiques de création et de suppression des liens. Donc le même sentiment d'appartenance est l'une des explications derrière la résistance d'un groupe d'individus dans le temps. Chacun interagit avec l'autre (dynamique interne) sans influencer son appartenance ou la composition.

Une orientation sémantique plus explicite

D'autre part, la deuxième orientation de cette dimension sémantique est verticale et plus expressive. Elle se manifeste par une sorte de loyauté, de fidélité et le degré de solidarité, par exemple, une fierté partagée en gagnant un championnat international de football par une équipe nationale, une perception subjective de similitude culturelle ou un attachement

émotionnel, etc. ((Karolewski 2009)). Cela signifie dans un SN/OSN, plus de richesse informationnelle, différents types de relations, activités, orientations et intérêts développées par les entités sociales. ***C'est à ce stade là qu'il faut améliorer la représentation du SN, en concevant des modèles plus riches de plus en plus explicite, afin de chercher une meilleure exploitation de cette sémantique, enrichir le SNA, révéler un caractère sémantique d'une collectivité vers une identité sémantique d'une structure noyau.***

Il n'est jamais facile d'enchaîner ces phases de traitements sur un plan d'analyse multidimensionnel. D'où il faudra entamer la piste de la sémantique en SNs indépendamment de leur dynamique temporelle comme une première étape. Suivant le cadre général de notre travail de recherche (**Figure 1**), toute proposition qui adopte une nouvelle dimension nécessite une modélisation et une approche d'analyse adéquates, sans omettre la pertinence des données et leur contexte, inspirée des environnements organisationnels, qui est primordiale (**Figure 1**), 'Organization mining from OSNs' ((Fire et al 2013b)). Dans ce sens, quand il s'agit d'étudier des nouvelles traces (des traces sémantiques), nous sommes intéressés par un type de SNs collaboratifs, notamment ceux qui évoluent dans un environnement d'apprentissage collaboratif. Ça sera une bonne motivation et l'illustration de telles contributions.

Chapitre 4 : Analyser un réseau social sémantique d'apprenants dans un environnement d'apprentissage social

1. Introduction et motivations

Les environnements d'apprentissage en ligne sont technologiquement en évolution permanentes, en cherchant souvent à accroître l'efficacité et la qualité de l'apprentissage à distance. En tant qu'une entité sociale, l'apprenant seul devant son PC, peut être découragé. C'est l'un des problèmes dans les systèmes de e-learning. Récemment, les recherches mettent en évidence l'aspect social dans les systèmes de e-learning comme l'une des principales priorités afin d'améliorer le processus d'apprentissage, augmenter le niveau cognitif des apprenants et d'atteindre des objectifs éducatifs. Les environnements d'apprentissage en ligne ont tendance à encapsuler des nouvelles technologies d'information et de communication pour incarner 'l'apprentissage social' ((Frédéric 2009)). L'apprentissage social donne une nouvelle ère à l'e-learning. Il ne s'agit juste d'un simple transfert vertical des connaissances des enseignants aux apprenants ((Halimi et al 2014)). Les outils collaboratifs et les médias sociaux sont intégrés pour favoriser les interactions sociales, les échanges, collaborations, etc., entre les apprenants, les encourager à partager leurs connaissances, expertise et savoir-faire, etc., donc un 'Computer Supported Collaborative Learning' ((Abel & Leblanc 2008)). Les liens sociaux entre les apprenants génèrent des SNs spécifiques, émergents qui ouvrent la voie déjà à des questions particulières et importantes. Quels sont les apprenants/acteurs influant sur le réseau : Quel est le meilleur collaborateur? Les outils collaboratifs ont-ils un impact sur les interactions de l'apprenant et son rôle dans le réseau? Si l'un des succès clés de l'apprentissage est les communautés ((Hathaway et al 2007)), alors ce type de réseau sera un contexte idéal pour discuter ces structures communautaires dans l'environnement d'e-learning. D'où, comment pouvons-nous identifier ce type de communauté, son caractère en tant qu'une communauté d'apprentissage ?, etc.

Évidemment, une méthode fondée sur le SNAM, est susceptible de répondre à telles question en se basant sur des modèles de SNs d'apprenants. Malgré ça, ils n'ont pas la même popularité en SNAM par rapport à d'autres données sociales de collaboration (collaborations scientifiques, relations coauteurs, de coédition etc.) (Girvan & Newman 2002) ((Wang et al 2013)) ((Brandes et al 2009)) ((Leskovec et al 2007)). Mais aujourd'hui, la variété et la richesse des données sociales venant particulièrement des environnements collaboratifs, poussent les analystes à chercher dans ces réseaux suivant des nouvelles dimensions et progrès. À la lumière de notre cadre de travail de recherche et les nouvelles tendances de SNAM, la méthode que nous proposons ajoutera une dimension sémantique en exploitant la richesse sémantique d'un SN d'apprenants. Ainsi, elle exigera un modèle plus riche et donnera plus de fidélité aux études analytiques, menant à des interprétations/ réponses plus profondes et significatives. Concernant la pertinence des données sociales, ce type de SNs émergent clairement dans un environnement d'apprentissage social et sont ainsi qualifiés comme implicites. Ils sont en effet extraits depuis les data logs des applications et systèmes qui ne sont pas strictement des applications OSN (Nettleton 2013). Les interfacions (collaborations) entre apprenants sont plus orientées, plus pertinentes par rapport à d'autres liens explicitement déclarés sur les OSNs. Par conséquent, une analyse sémantique d'un SN semble être prometteuse pour aider le tuteur ou le modérateur dans la prise de décision et la recommandation intelligente des meilleurs parcours d'apprentissage pour un apprenant, dans

le cas par exemple d'un apprentissage personnalisé, 'Personal Learning environment 'PLE' ((Halimi et al 2014)). À cet égard, nous découvrons les propriétés sémantiques des connexions sociales entre apprenants. Ça sera également bénéfique pour les perspectives techniques visant à remplacer la modération coûteuse basée sur les opérateurs humains. Suivant la sémantique (type) d'une relation, nous pouvons par exemple identifier les meilleurs collaborateurs ou la meilleure communauté qui seront recommandés pour un apprenant donné selon ses préférences, ses intérêts et sa positivité à ce type de relation. De plus, nous pouvons fournir des informations utiles sur le dynamisme en termes d'interactions et popularité des apprenants pour anticiper par exemple la réduction du sentiment d'isolement ou l'augmentation du sentiment d'appartenance à une communauté. Les résultats prévus peuvent améliorer les moteurs de recherche qui sont inclus dans ces systèmes en identifiant les rôles de leadership: Les apprenants actifs et stratégiques. Par conséquent, notre étude prend sa place entre les tendances de SNA sémantique (Modéliser/ analyser des représentations de SN plus réalistes) et les études des interactions (traces) sociales sémantiques entre apprenants dans un système d'e-learning.

Techniquement

Tout d'abord, un réseau formé par les interactions d'un ensemble de collaborateur n'a pas été qualifié de manière hasardeuse comme un SN. Plusieurs démonstrations confirment que ce type de réseaux affiche de nombreuses caractéristiques des graphes sociaux réels (distribution de loi de puissance, phénomène de petit monde, tendance au Clustering, etc.). D'autre part, les systèmes de e-learning font souvent appel aux technologies de web sémantique et ontologies pour définir des métadonnées, structurer et annoter des matériaux et ressources pédagogiques d'apprentissage ((Halimi et al 2014)). Le but est de faciliter l'utilisation, chercher plus d'expressivité et de représentation formelle et compréhensible des connaissances. Mais, moins d'attention a été accordée à la sémantique des interactions sociales entre les apprenants, comment la modéliser, représentent et analyser, tant que l'aspect social de l'apprentissage était un nouveau paradigme dans ces environnements. ((Torniai et al 2008)) parmi peu de chercheurs qui ont essayé d'étudier les interactions sociales sémantiques en e-learning.

Au-delà des représentations classiques et topologiques, nous proposons une approche pour construire et analyser un modèle sémantique d'un graphe social d'apprenants. Dans un premier temps, nous proposons un vocabulaire simple (modèle ontologique) qui permet de décrire les entités sociales 'apprenants' (âge, positivité, appréciation, etc.) et leurs interactions (type, orientation) par un graphe social RDF. Cependant, l'analyse d'un graphe (typé) RDF est elle-même un autre défi à cause de sa complexité, le problème d'interopérabilité et la carence des outils nécessaires pour l'analyser. Pour surmonter un tel défi, on propose un processus de 'mapping' vers une représentation équivalente qui conserve la même richesse exprimée selon le degré d'expressivité que nous avons adopté. Nous serons ainsi en mesure d'enrichir et paramétrer nos études analytiques suivant le type de relations. Une analyse locale (individuelle) basée sur des métriques de centralité et de prestige sera implémentée pour quantifier et comparer l'importance des acteurs / apprenants et identifier les rôles de leadership. Les apprenants révélés comme stratégiques, ont un potentiel plus élevé pour échanger et diffuser l'information en encourageant les autres à apprendre. Cette analyse nous amène à interpréter aussi le profil social de l'apprenant. En outre, pour avoir un aperçu sur l'influence individuelle (locale) de l'apprenant par rapport à l'ensemble du réseau, nous étudions des distributions statistiques plus enrichies et des croisements entre certains indicateurs. Les résultats détermineront la centralisation du réseau, en comprenant mieux sa connectivité. Dans ce sens, nous identifions la configuration du réseau en communautés, chacune aura un caractère sémantique interne contribué par des apprenants plus proches et

partageant le même type de relation. Tous ces traitements proposés, y compris le 'mapping' et les études empiriques sont mis en œuvre sous forme d'un prototype expérimental. Il s'agit d'une application logicielle (EA-SemSNL) que nous avons implémenté pour analyser et paramétrer l'analyse d'un SN des apprenants-collaborateurs extrait depuis l'environnement d'apprentissage social 'SoLearn' ((Halimi et al 2011)). Le modèle sémantique de ce SN constituera les données d'entrée pour cette application, en montrant comment exploiter la sémantique exprimées pour analyser ces nouvelles traces sans outils intermédiaires. D'un autre côté, les résultats obtenus révélant, soit les positions stratégiques, les paramètres descriptifs du réseau ou sa structure modulaire varient selon le type des liens sélectionné.

2. Modélisation d'un réseau social sémantique entre apprenants

Partant du principe que les interprétations significatives viennent des représentations réalistes, notre objectif consiste à définir un modèle plus expressif de SN d'apprenants. Le contexte du e-learning social donne tout d'abord une dimension sémantique à exploiter pour modéliser et étudier des nouvelles traces sémantiques d'un SN. Les aspects et les propriétés sémantiques seront extraits depuis les caractéristiques des interactions sociales entre apprenants :

- Leurs interactions sociales ne sont pas explicitement énoncées par rapport aux OSN. Les relations sont ici plus implicites, plus orientées et donc plus pertinentes
- Ces interactions sont issues de « l'acte de collaboration» et ne sont pas des simples connaissances éphémères, innées. Elles sont plutôt acquises et conditionnées par des compétences sociales de l'apprenant.
- La densité de ces liens dépend de la positivité de l'apprenant à interagir qui est à son tour fortement affectée par les outils collaboratifs utilisés.

Partant de ces aspects sémantiques cachés particulièrement derrière les interactions, la représentation du réseau sera améliorée. Au même titre que les modélisations ontologiques (ex. FOAF profile) qui ont été faites pour les OSNs, nous proposons un vocabulaire spécifique: Un ensemble de concepts décrivant les propriétés sémantiques de ce SN (le profil de l'entité sociale –apprenant- et ses interactions). Ces concepts sont formalisés dans un schéma ontologique suivant lequel on modélise le SN sémantique d'apprenants.

2.1.Concepts schématisés pour décrire les interactions et l'entité social-apprenant

On sait que l'aspect social de l'apprentissage contribue à accroître le niveau cognitif pour un apprenant. Mais dans ces environnements, la positivité de l'apprenant à se socialiser et interagir est influencée. Les outils de collaboration utilisés qui déterminent le mode et la synchronisation de l'interaction, présentent l'un des principaux facteurs influant. Par conséquent, nous distinguons deux modes (types) d'interactions (de collaboration): Collaboration synchrone (CS) et asynchrone (CA). La première nécessite que les 2 acteurs–apprenants 'x' et 'y' soient simultanément impliqués dans ce lien suivant par exemple un temps et délais convenus, tandis que la seconde est libre de cette contrainte de temps. En outre, quel que soit le type de relation, l'échange entre des 'x' et 'y' est normalement symétrique, initialisé (lancé) par une demande (de collaboration) émise par 'x' et accepté par 'y' (ou inversement) pour interagir. Donc, nous proposons des propriétés qui distinguent les deux types de relations, ainsi que le récepteur et l'émetteur qui l'ont lancé (**Figure 113**).

Même si ces acteurs présentent sur le plan topologique le niveau le plus bas de granularité, les atomes du SN, ils seront caractérisés sémantiquement aussi par rapport à ce contexte qui les met en interaction. On s'est inspiré de certains critères de recherche de collaborateurs ((Lafifi & Bensebaa 2008)) en CSCL pour décrire le profil de l'entité sociale apprenant (**Figure**

113) : Positivité, appréciation, niveau cognitif (profil cognitif). Des concepts que les expériences précédentes et les spécialistes en sociologie et en dynamique de groupes ont démontré leur efficacité.

La socialisation de l'apprenant peut être également influencée par ses propres compétences: le niveau cognitif (excellent, bon, moyen). Il peut être positif (normal ou négatif) face aux demandes reçues par d'autres apprenants pour collaborer. Il peut avoir une appréciation très agréable, agréable, désagréable, etc., vers ses partenaires et interactions. Cependant, son profil social étant un bon collaborateur, faible ou isolé, sa positivité et son appréciation sont affectés et changent selon le type de relation. On verra que ces propriétés seront déterminées et interprétées de manière plus réaliste en analysant son SN sémantique.

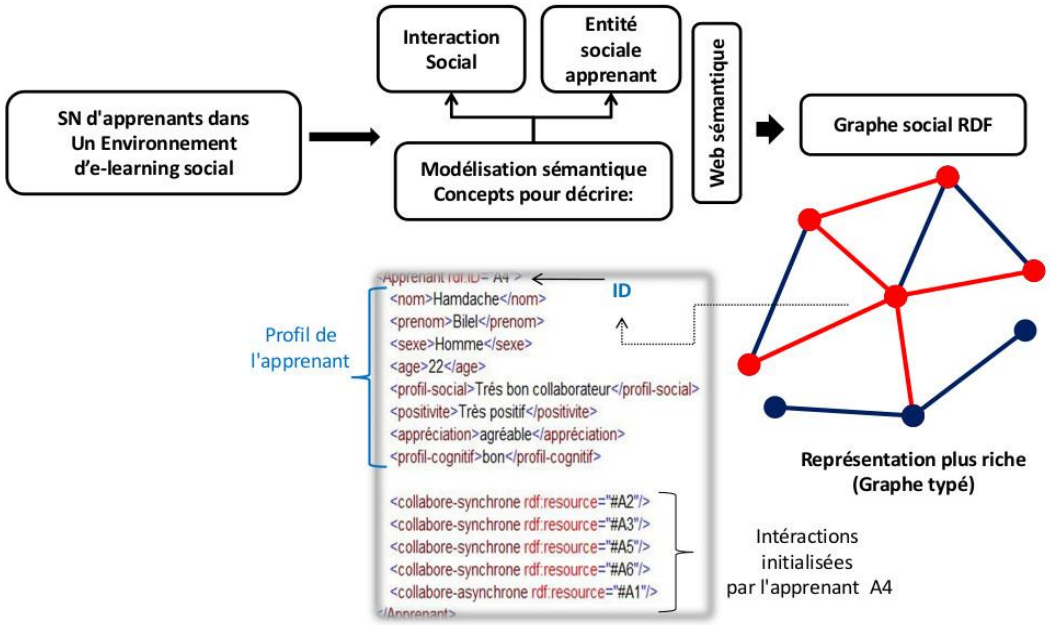


Figure 113. Aperçu sur le modèle sémantique d'un SN de collaborateurs apprenants

Sérialisation des concepts, un schéma ontologique pour générer des annotations RDF

Suivant ce degré d'expressivité qu'on adopte, les concepts proposés seront sérialisés dans un vocabulaire bien adapté : Un schéma ontologique RDFS. Chaque interaction est représentée par une propriété RDF asymétrique (**Figure 113**). La propriété est en mesure de distinguer le type d'interaction, son 'domain & range' sont exploités pour désigner respectivement l'émetteur et le récepteur de la demande d'interaction ayant lancé l'échange. Pour chaque acteur-apprenant (Ai), des annotations (triplets) RDF sont générées pour décrire sémantiquement ses liens typés ainsi que son propre profil social. Par conséquent, l'ensemble de ces annotations constituent le SN sémantique des apprenants sous forme d'un graphe sémantique RDF.

2.2.Formaliser un processus de 'mapping'

Avant d'analyser directement ces nouvelles traces, nous devons mettre en évidence les exigences et la complexité quand il s'agit d'analyser/ fouiller un graphe RDF. Nous proposons un processus de 'mapping' intermédiaire bien adapté pour faire face à telles situations (**Figure 114**). il s'applique sur le graphe social RDF des apprenants : R (N, P). Le but est de passer vers une représentation G équivalente en conservant la même richesse sémantique exprimée dans R. On assure initialement la distinction du type et de l'apprenant-

initiateur pour chaque interaction (**Figure 114**). En d'autres termes, R (N, P) sera traité (analysé) par l'intermédiaire de G (V, E, L)

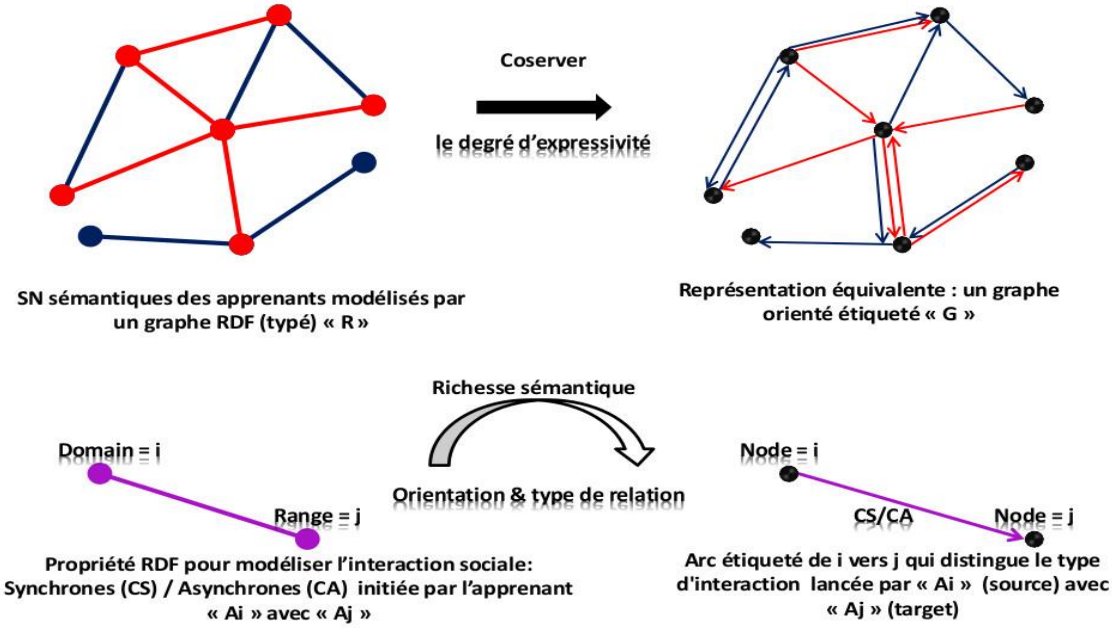


Figure 114. Processus de mapping vers un graphe orienté et étiqueté

Formellement la fonction de 'mapping' est définie par $M: R \rightarrow G$ tel que:
$\forall$ n $\in$ N, $\exists$ v $\in$ V $\setminus$ M(n) $= v$
$\forall\, p_{Type}\,(n_i, n_j) \in$ P, $\exists$ e $(\mathrm{v_i}, \mathrm{v_j}) \in$ E $\setminus$ L(e) $= Type \wedge$ M(n_i) $=$ $\mathrm{v_i}$, M$(n_j) = \mathrm{v_j}$

Ici, N est l'ensemble des ressources (annotations) RDF, où chacune correspond à un acteur-apprenant. P est un ensemble de propriétés RDF décrivant les liens typés entre les ressources RDF. D'autre part, V et E présente respectivement l'ensemble des sommets et des arêtes (arcs) de G, alors que L est une fonction d'étiquetage utilisée pour indiquer le type de lien entre les deux apprenants. Chaque interaction de la forme $p_{Type}\,(n_i, n_j)$ se présente par un arc étiqueté $(\mathrm{v_i}, \mathrm{v_j})$ dans G, soit par l'étiquette 'CS' ou 'CA' et orienté $\mathrm{v_i}$ ('domain' n_i) vers $\mathrm{v_j}$('range' n_j), (**Figure 114**).
Un algorithme itératif (Algorithme 13) est proposé pour exécuter ce processus de 'mapping'. L'algorithme se base sur le mécanisme orienté-objet. Les nœuds et les arcs dans le graphe cible sont présentés par des objets pour pouvoir conserver les mêmes descriptions qu'on trouve dans les ressources du graphe RDF (Algorithme 13). Autrement dit, chaque profil social d'un apprenant ou interaction se présente par un objet dans G. L'aperçu de l'Algorithme 13 en pseudo-code inclut des fonctions prédéfinies de certaines API JAVA qui sont utilisées pour interroger le graphe RDF et créer le graphe G.

Algorithme 13. Pseudo-code du processus de 'mapping d'un graphe RDF vers un graphe orienté étiqueté

```
(1)
 SN <- Create -directed-graph <My-Node, My-Link> ();

// Créer des noeuds
(2)
List-RDF-Resources-Learners <- Recover_ List-RDF-Resources ();
RDF-Resource-Learner <- Null; Nodes-List <- Null

(3)
TANT QUE (! Empty (List-RDF-Resources-Learners) FAIRE
        RDF-Resource-Learner <- List-RDF-Resources-Learners. Retrieve-element ();
        My-Node Node  <- Create-Node (RDF-Resources-Learners);
        Nodes-List.Add (Node);  SN.Add-node (Node);

    end (3)
// Créer des arcs étiquetés  pour chaque acteur
My-Node Node-Learner <- Null;
(4)
TANT QUE (! Empty (List-RDF-Resources-Learners)) FAIRE
        RDF-Resource-Learner<- List-RDF-Resources-Learners. Retrieve-element ();
    (5)
    POUR (i = 0; i < Nodes-List. size (); i++) FAIRE
        SI  (RDF-Resource-Learner.ID = Nodes-List (i).description())  ALORS
            Node-Learner <-  Nodes-List (i);
            break;
        End (5)
    (6)  //Créer des arcs étiquetés vers des collaborateurs en mode synchrone
    My-Node Node-collaborator-Synchronous <-  Null;
    List-RDF-Resources- Collaborators-Synchronous <-
    Recover_ List-RDF-Resources-with-property-Collaborate-Synchronous (RDF-Resource-Learner);

    (7)
    TANT QUE  (! Empty (List-RDF-Resources-Collaborators-Synchronous))   FAIRE
            RDF-Resource- Collaborator <- List-RDF-Resources- Collaborators-Synchronous. Retrieve-element ();
        (8)
        POUR (i = 0; i < Nodes-List. size (); i++) FAIRE
            SI (RDF-Resource- Collaborator.ID = Nodes-List (i).description()) ALORS
                Node-collaborator-Synchronous <-  Nodes-List (i);
                break;
            End (8)
        My-Link Arc-Synchronous <-  Create-Arc ('CS')
        Direction (Arc-Synchronous, Node-Learner, Node-collaborator-Synchronous);
        SN. Add-link (Arc-Synchronous);
        End(7)
    (9)  répéter le  traitement  de (6) pour créer les arcs étiquetés de collaboration asynchrone
    End(4)
```

3. Prototype expérimental basé sur une application logicielle (EA-SemSNL) pour enrichir les expériences d'analyse

Les environnements d'apprentissage social en ligne sociaux (de collaboration) constituent une source fiable des données sociales, Les data logs des outils de collaboration encapsulés à l'intérieur fournissent des SNs pertinents. Dans ce sens, nous étudions un échantillon de SN d'apprenants qui évolue dans l'environnement 'SoLearn' ((Halimi et al 2011)). Le réseau est extrait du fichier log enregistrant les requêtes réussies et les communications lancées entre 20 apprenants. SoLearn ((Halimi et al 2014)) ((Halimi et al 2011)) est un environnement de e-learning social qu'on ne va pas s'en servir seulement pour extraire des données sociales. C'est aussi un environnement d'apprentissage permettant de personnaliser et améliorer le processus

d'apprentissage des utilisateurs, leurs interactions et leurs profils ((Halimi et al 2014)). Par conséquent, l'analyse enrichie que nous proposons ainsi que les interprétations et les conclusions qui seront obtenus seront en bénéfice d'une recommandation plus intelligente, réaliste et plus précise.

3.1.Modèle de SN sémantique, des données d'entrée pour EA-SemSNL

Le réseau étudié est représenté par de nouvelles traces sémantiques (annotations RDF) qui se générèrent en se basant sur le schéma ontologique proposé. Ce schéma ainsi que les annotations RDF sont tout d'abord sérialisés à l'aide d'un éditeur de document à base de web sémantique RDF / OWL: 'Altova Semantic Works'. Cela donne un graphe RDF, le modèle sémantique du SN qui va subir le processus de 'mapping' et des études empiriques. L'ensemble de ces traitements est organisé en un prototype expérimental, implémenté en langage JAVA sous forme d'une application logicielle, qu'on appelle EA-SemSNL: 'Enriched Analysis of Semantic Social Network of Learners'. Donc ce modèle sémantique constitue les données d'entrée pour EA-SemSNL (**Figure 115**).

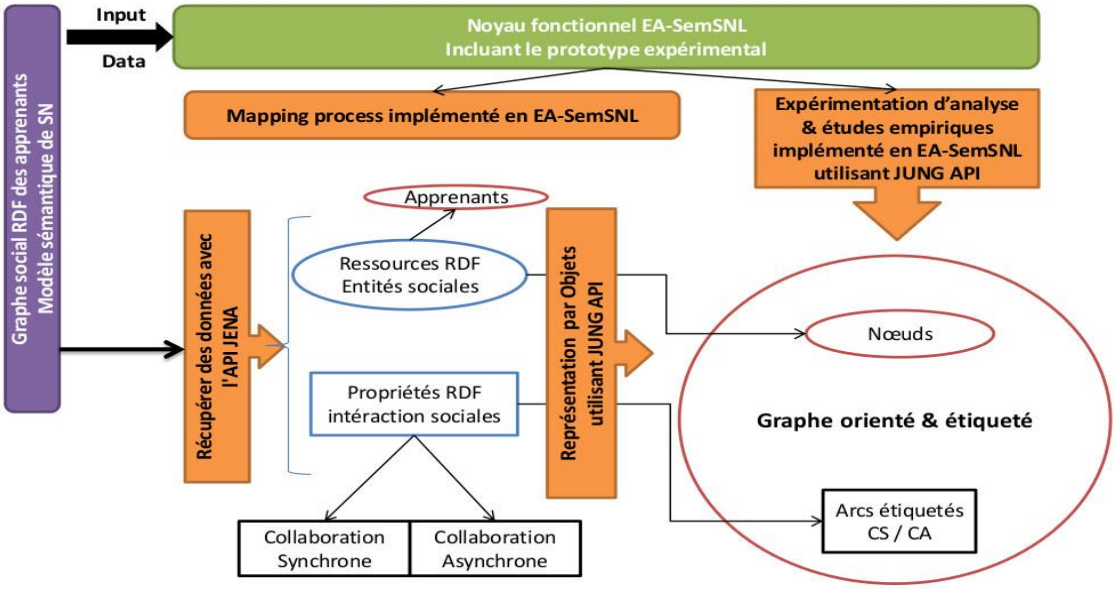


Figure 115. Le graphe social RDF des apprenants étant les données d'entrée d'EA-SemSNL où le processus de mapping et le prototype expérimental sont implémentés dans son noyau fonctionnel

EA-SemSNL est mis en œuvre sur l'environnement de développement de la plate-forme Eclipse et s'exécuté sur 'Java Runtime Environment' (JRE) 8.31, sur Windows 64 bits. De nombreuses API sont intégrés dans EA-SemSNL. Mais les plus importantes, l'API JENA et JUNG (Java Universal Network Graph) sont utilisées dans le noyau fonctionnel pour implémenter le processus de 'mapping' et le prototype expérimental (**Figure 115**). Certaines procédures prédéfinies servent à effectuer certaines opérations, par exemple la récupération des données relationnelles RDF (par JENA) en mapping, ou construire (transformation en objets) et analyser le graphe orienté étiqueté cible (par JUNG), (**Figure 115**). Il est à noter que le mapping s'exécute automatiquement sur le modèle sémantique, une fois que l'application est lancée pour produire la représentation équivalente.

3.2.Aperçu sur EA-SemSNL et méthodologie d'évaluation

La **Figure 116** montre l'interface principale de notre application JAVA (EA-SemSNL). Elle se compose essentiellement d'un menu (barre de menu), des onglets, etc. Le menu offre l'accès aux données d'entrée (le SN : liens et entités sociales) et propose diverses

fonctionnalités interactives, des traitements d'analyse, etc. Avec les fenêtres onglets on peut afficher aussi le graphe social RDF, son vocabulaire RDFS et visualiser la représentation (graphe étiqueté) cible selon le type de lien sélectionné (**Figure 116**). Donc, cette option qui permet à l'utilisateur de choisir le type de relation permet ainsi le paramétrage de toutes les fonctionnalités et expériences d'analyse qui sont classées sur la droite (**Figure 116**). L'utilisateur peut sélectionner un apprenant en entrant son ID et calculer son potentiel dans le réseau à travers d'un ensemble de indices classés dans une même rubrique comme locaux. D'où l'utilisateur découvre différentes positions pertinentes (stratégiques) dans le réseau. Comme il est montré dans la **Figure 116**, EA-SemSNL fournit autres options, par exemple l'opportunité de normaliser les résultats de calcul. D'autre part, l'utilisateur peut étudier des distributions et les corrélations statistiques entre certains paramètres (métriques), apercevoir la connectivité du réseau et la détection de communautés, etc., (**Figure 116**). L'interface contient un onglet démo qui affiche l'état de l'opération en cours et quelques résultats en sortie (**Figure 116**).
Par ailleurs, beaucoup de résultats et interprétations sont affichées par nombreuses autres interfaces et boîtes de dialogue qui se génèrent suite à une manipulation interactive avec le noyau fonctionnel d'EA-SemSNL. La **Figure 117** montre l'architecture du prototype expérimental implémenté dans ce noyau fonctionnel. Il est organisé suivant un schéma d'analyse bien défini et structuré. C'est l'organe exécutif derrière toutes les opérations expériences d'analyse effectuées sur ce modèle de SN. L'architecture proposée dans la **Figure 117** confirme que la majorité des fonctionnalités: visualisation graphique du réseau (**Figure 118**), matrice d'adjacence, etc., études analytiques: le potentiel de l'apprenant (centralité/ prestige), structure communautaire, etc., sont enrichies, paramétrées suivant le type de lien (d'interaction) sélectionné.

Figure 116. Aperçu sur EA-SemSNL, l'interface principale

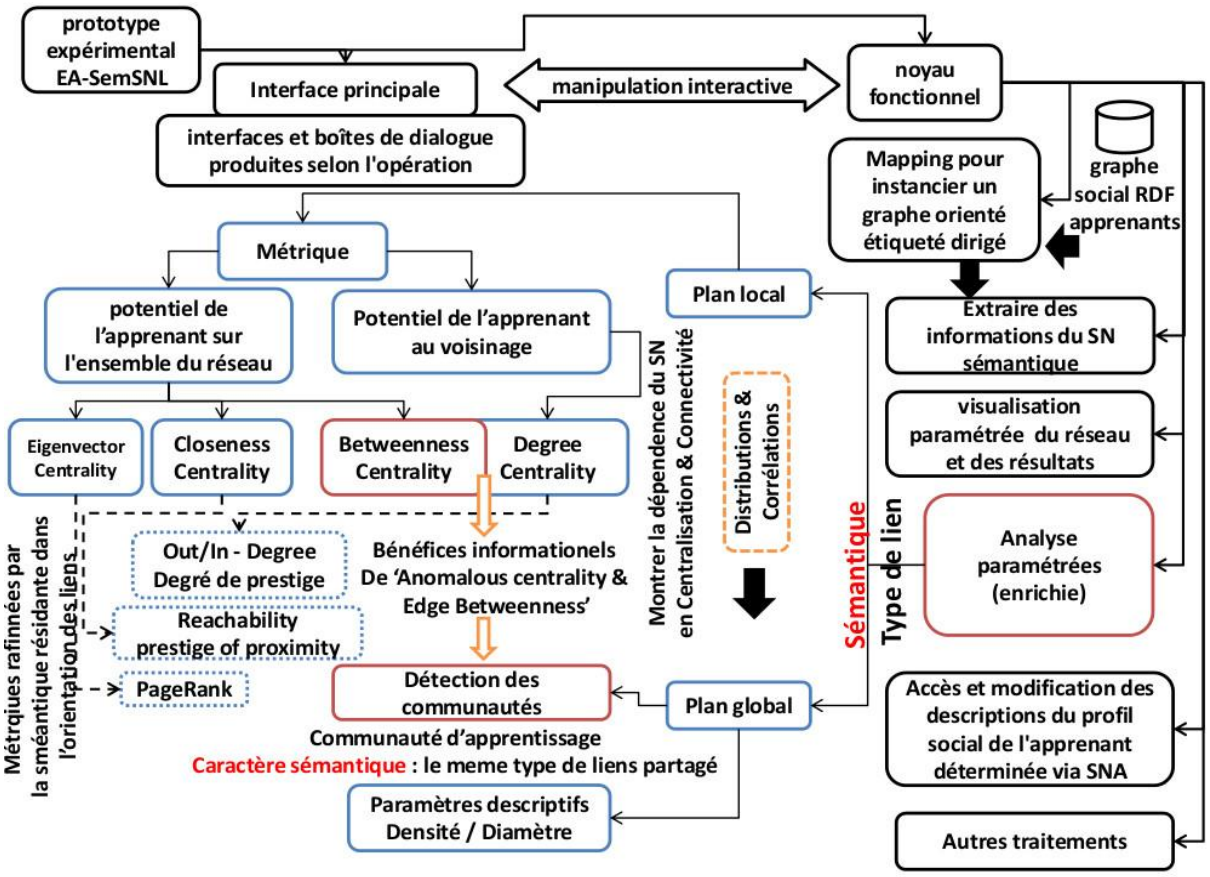


Figure 117. Aperçu sur l'architecture du noyau fonctionnel d'EA-SemSNL comprenant le schéma d'analyse du protocole expérimental

Lors du lancement d'EA-SemSNL, les données d'entrée passent par le processus de mapping et peuvent être ainsi visualisées comme dans la **Figure 118.** Pour les analyser (ce modèle sémantique de SN d'apprenants), on expose un ensemble d'expériences suivant une méthode d'évaluation (schéma d'analyse) divisée en étapes: Partant d'un plan d'analyse locale vers un plan plus global, passant par des études axées sur la dépendance de structure du réseau en centralisation et connectivité (**Figure 117**).

En premier lieu, nous évaluons le potentiel d'un apprenant donné en utilisant différentes mesures proposée dans EA-SemSNL. Nous verrons ainsi comment ce potentiel varie au même temps, quand la richesse sémantique (le type et l'orientation de ses liens) est exploitée. D'où, il sera possible d'interpréter le profil social (positivité, etc.) de chaque apprenant. Avec EA-SemSNL, nous pourrons catégoriser ces acteurs-apprenants selon leur positivité, appréciation, etc., envers un type d'interaction donné (synchrone). Dans ce sens, EA-SemSNL peut nous montrer si un sous-ensemble, d'apprenants, classés par exemple comme bons collaborateurs, changent leur comportement selon un autre type d'interaction (asynchrone). De plus, un autre module de l'application, nous permettra de trouver les rôles leadership: Les apprenants les plus centraux, les plus prestigieux. Différents apprenants dominant seront détectés suivant différents points de vue. Si EA-SemSNL (le prototype expérimental) est bien efficace sur un plan d'analyse locale, il est également muni d'une autre composante destinée à étudier

l'organisation globale du réseau (**Figure 117**). Cette partie est développée pour découvrir la structuration du réseau en communautés d'apprentissage. Certaines configurations et hypothèses qui caractérisent ce type de communautés seront d'abord montrées dans ce réseau. Un algorithme de division est implémenté en le paramétrant avec le type de liens, ce qui nous permettra de détecter des communautés d'apprentissage significatives topologiquement mais aussi sémantiquement.

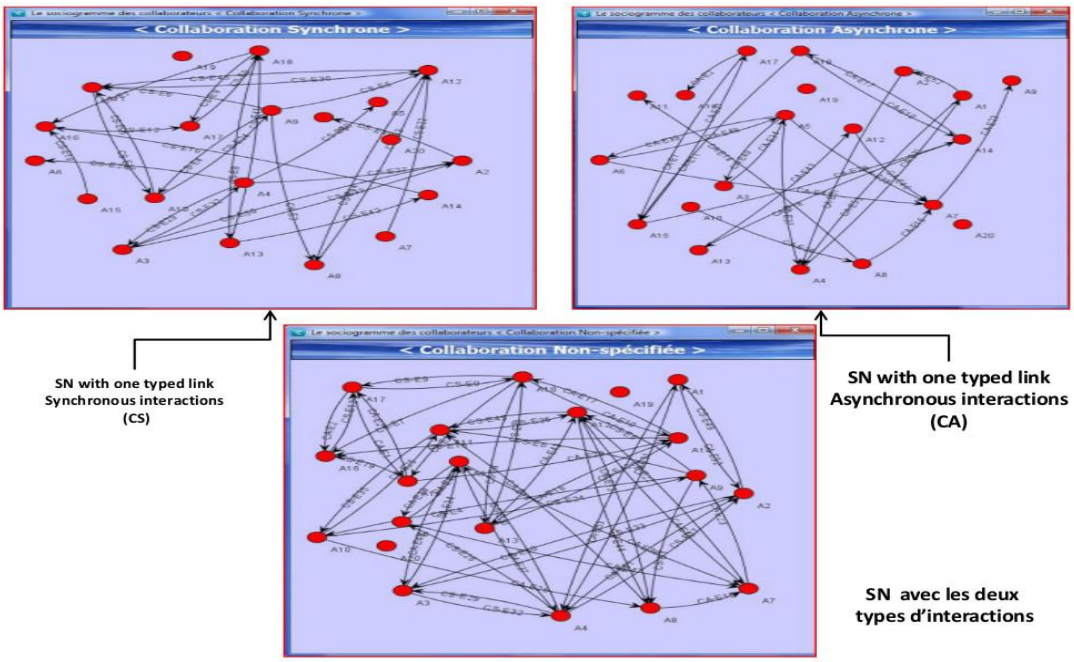


Figure 118. Le graphe social (typé et non typé) des apprenants visualisé par EA-SemSNL

3.3.Études analytiques et empiriques par EA-SemSNL

Sur le plan local, EA-SemSNL permet de sélectionner n'importe quel apprenant et mesurer sa centralité individuelle ou son prestige au voisinage ou sur l'ensemble du réseau (par un ensemble de mesures implémentées), (**Figure 117**). Comme le montre la **Figure 119**, l'application calcule et affiche par exemple la centralité de degré qui désigne le nombre d'interactions réalisées par un apprenant-acteur: A4. A4 a plus d'interactions synchrones qu'asynchrones. Sur le côté droit de la **Figure 119**, on remarque que l'activité individuelle de l'apprenant dans son voisinage peut être raffinée par l'orientation de ses liens en appliquant 'in/out Degree' (**Figure 117**). Les interprétations sont ainsi plus diversifiées. Le nombre d'interactions synchrones issues des demandes reçues et acceptées par A17 (In-Degree) est supérieur au nombre des interactions initiées par A17 (Out-Degree), contrairement au cas asynchrone (**Figure 119**). En d'autres termes, A17 est localement plus prestigieux en interaction synchrones par rapport au mode asynchrone.

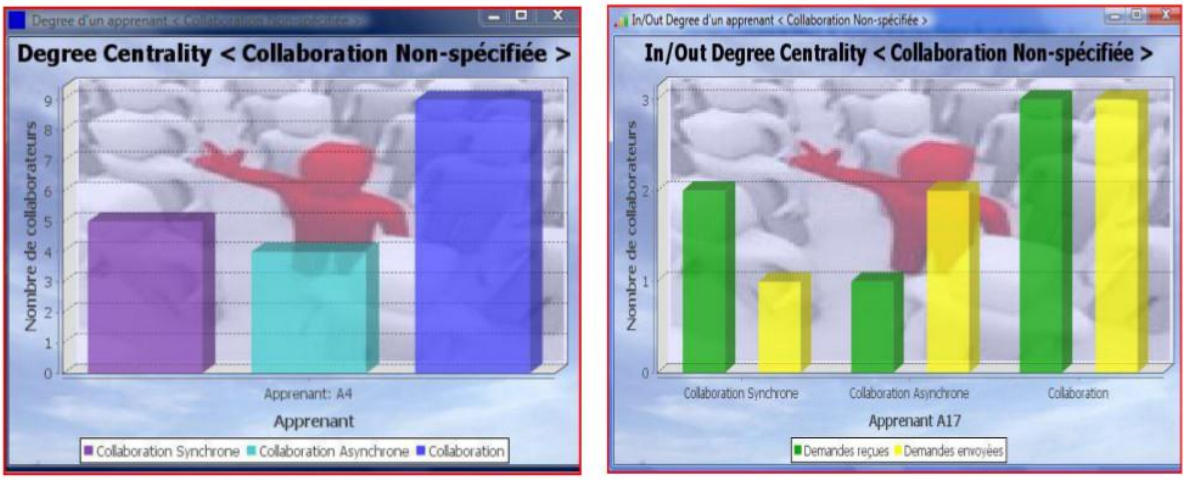


Figure 119. Centralité de degré de l'apprenant A4 & In\Out-Degree de A17 calculés par EA-SemSNL

De plus, EA-SemSNL permet d'évaluer le potentiel de l'apprenant sur l'ensemble du réseau (**Figure 120**). Après avoir fait entrer l'ID d'un apprenant donné, l'application peut évaluer sa proximité (par 'Closeness centrality' : Cc) et son rôle d'intermédiaire (par 'Betweenness centrality') sur les chemins les plus courts (les géodésiques). D'après l'évaluation de la proximité (en vert) et l'indépendance (orangée) de l'apprenant A18, abordées sur le côté gauche de la **Figure 120,** A18 semble être plus proche du reste du réseau à travers ses interactions synchrones. A18 est moins indépendant contrairement au cas asynchrone où il est plus autonome. La signification sémantique portée par l'orientation des liens, faisant référence à l'initiateur de l'interaction est aussi exploitée par notre programme en offrant la possibilité de raffiner cette métrique (**Figure 120**). Nous pouvons faire exécuter le programme exclusivement sur les chemins entrants. Dans ce cas, la centralité de la proximité s'interprète comme l'accessibilité vers un acteur apprenant à partir de tous les autres acteurs du réseau.

D'après le résultat de calcul affiché par EA-SemSNL sur le côté droit de la **Figure 120,** l'application permet de comparer aussi l'intermédiarité d'un apprenant donné. Par exemple, A12 est plus intermédiaire sur les chemins de communications asynchrones (**Figure 120,**). Une fois que le type de lien n'est pas spécifié, l'intermédiarité de A12 augmente (**Figure 120**), car le nombre de géodésiques traversant A12 devient plus important et il joue ainsi un rôle intermédiaire plus central.

Autres options sont également programmés dans EA-SemSNL pour mesurer le potentiel global de l'apprenant en se basant sur la centralité de ses voisins (et les de ses voisins): Ses collaborateurs apprenants, comme par exemple le 'Eigenvector centrality'. Cependant, l'orientation des liens est encore une fois plus significative pour cette évaluation, ce qui permet d'implémenter un algorithme plus adapté, le PageRank (**Figure 117**). D'où, EA-SemSNL est en mesure d'évaluer le prestige de l'apprenant, par exemple A10 (**Figure 120**), mais cette fois sur l'ensemble du réseau. Autrement dit, on évalue la probabilité d'initier une interaction par n'importe quel autre apprenant avec A10. Par conséquent, il est remarquable dans la partie inférieure de la **Figure 120,** que les liens synchrones rendent l'apprenant A10 plus prestigieux par rapport aux interactions asynchrones. Il est à noter qu'EA-SemSNL offre l'opportunité aux utilisateurs de régler certains paramètres liés à l'algorithme de PageRank.

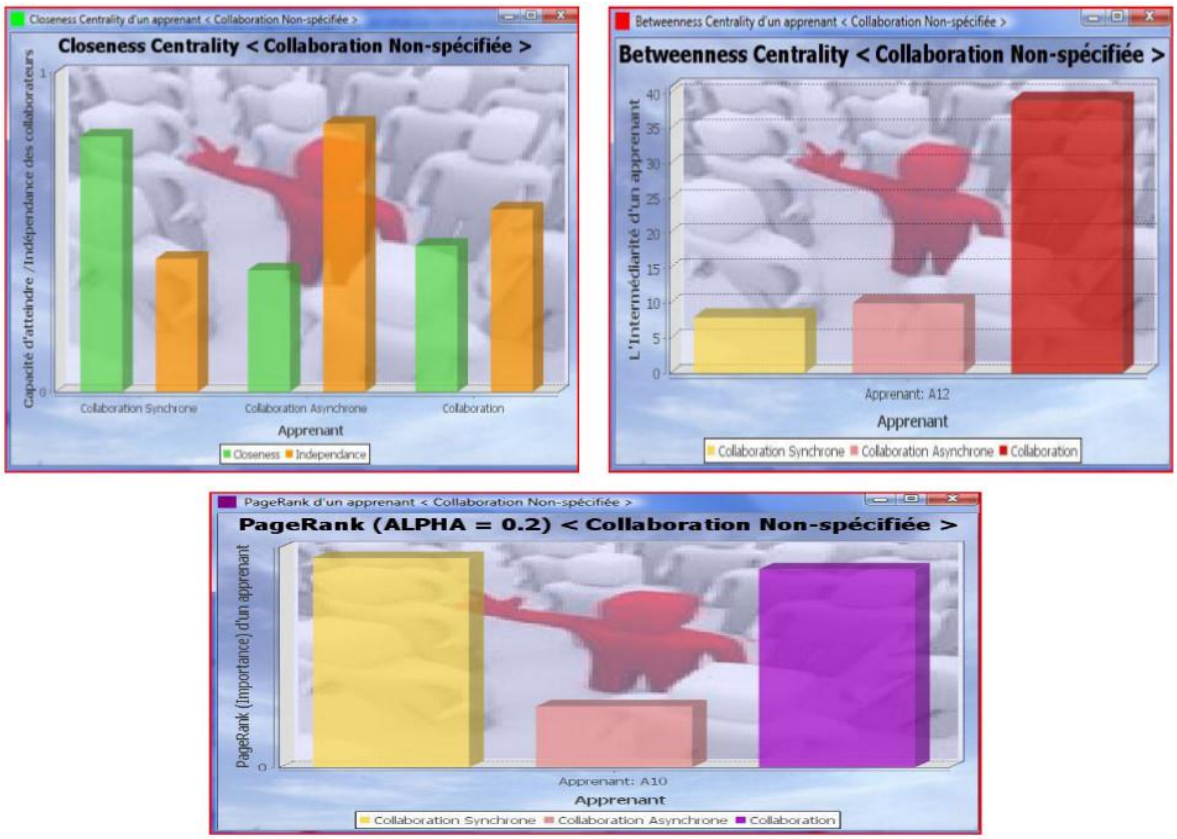


Figure 120. Potentiel en termes de 'closeness & betweenness centrality' et le prestige en termes d'accessibilité et PageRank d'un apprenant donné, calculés sur l'ensemble du réseau par EA-SemSNL

Deuxièmement, nous avons développé dans EA-SemSNL un autre module qui est capable de calculer le potentiel / le prestige de tous les apprenants en même temps (**Figure 121**). L'objectif est de détecter et de montrer quels sont les apprenants stratégiques et ceux qui occupent des positions pertinentes: les plus centraux, les plus prestigieux dans le réseau. Donc, différents rôles de leadership sont révélés suivant différentes métriques (**Figure 117**), ainsi que le type et l'orientation des liens. Dans la **Figure 121**, EA-SemSNL affiche d'abord la centralité de degré de tous les apprenants. Par ce moyen, on peut révéler l'acteur le plus central au voisinage, A12 dans le cas synchrone, A5 dans le cas asynchrone et A4 dans le cas où le type de lien n'est pas spécifié. Dans l'autre côté (**Figure 121**), on détecte l'apprenant le plus central qui joue le rôle le plus intermédiaire sur les flux de communication de réseau, par exemple A18 dans le cas des interactions synchrones (**Figure 121**). Il est aussi remarquable que certains acteurs ayant une centralité d'intermédiarité nulle sont ceux qui ont déjà un degré égale à 0 (**Figure 121**). Par contre, certains n'ont aucune intermédiarité bien que le degré correspondant n'est pas nul. Ces acteurs ont en effet un In-Degree = 0 ou Out-Degree = 0. En général, la **Figure 121** montre que les apprenants les plus intermédiaires sont susceptibles de présenter des degrés élevés et jouent ainsi un rôle de leadership, à l'image de l'apprenant A12.

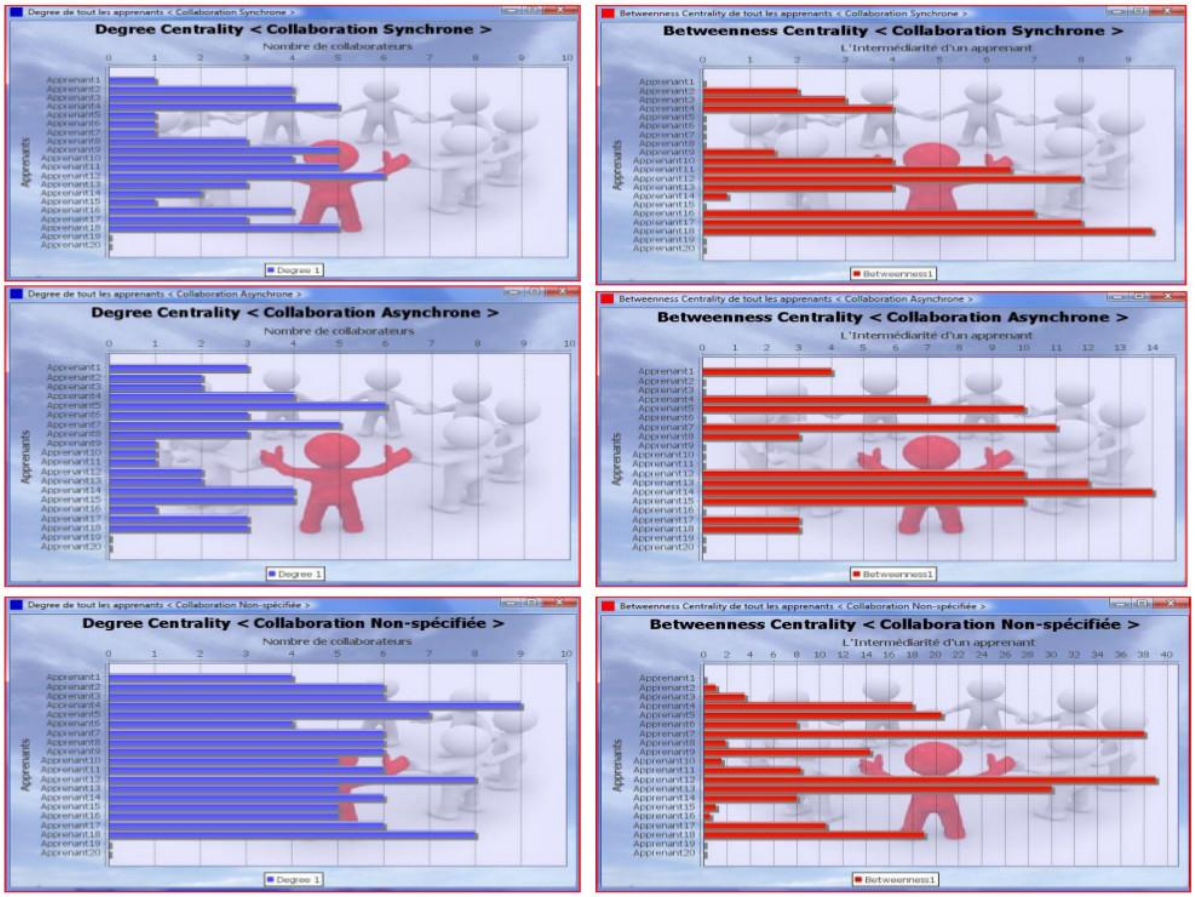


Figure 121. Pertinence de tous les acteurs du réseau suivant la centralité de degré et d'intermédiarité, calculée par EA-SemSNL en spécifiant ou non le type de lien

Ces derniers résultats donnent l'impression que le les centralités de degré et d'intermédiarité dans ce réseau de collaborateurs apprenant sont corrélées. EA-SemSNL est capable de d'étudier, paramétrer et afficher une telle corrélation qui nous permettra de découvrir des cas particuliers assez significatifs (**Figure 122**). Grâce à cette option les acteurs apprenants sont distribués entre les valeurs de degré et d'intermédiarité dans le cas synchrone ou asynchrone (**Figure 122**). On constate que l'apprenant qui a un faible degré, n'a pas toujours une faible intermédiarité. C'est le cas d'un apprenant qui détient une position exceptionnelle qui se réfère à 'Anomalous centrality' en SN. La **Figure 122** montre que tels apprenants jouent un rôle de leadership en termes d'intermédiarité bien qu'ils aient peu d'interactions localement. C'est une configuration qui reflète souvent l'existence d'une structure modulaire de réseau (communautés).

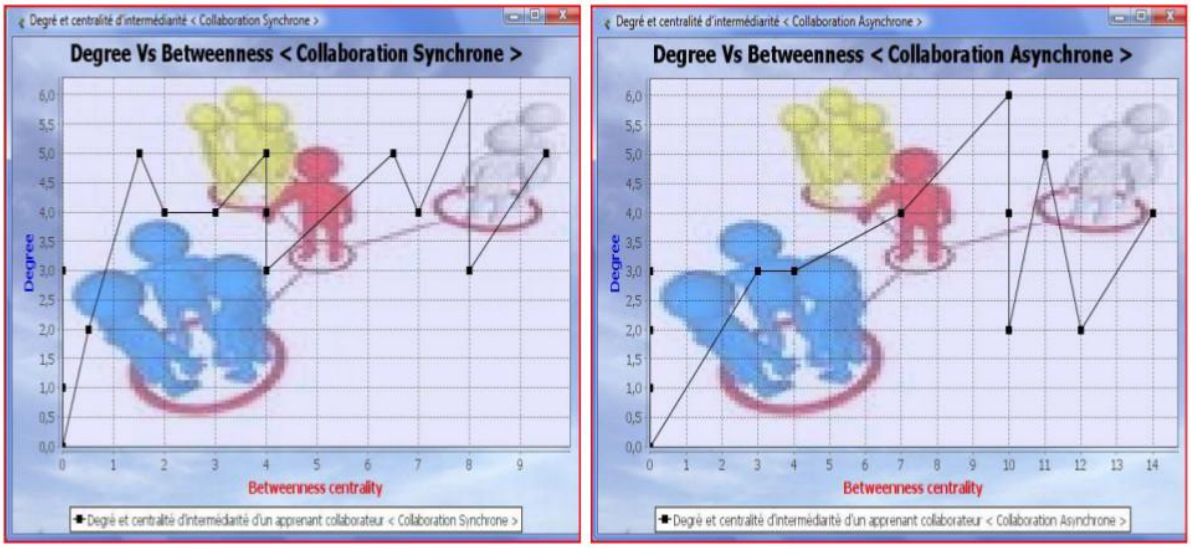


Figure 122. Centralités de degré et d'intermédiarité de l'ensemble des apprenants collaborateurs qui forment le SN

On a trouvé que A12 ou A13 présentent par exemple ces 'Anomalous centralities' sur lesquels EA-SemSNL peut s'appliquer pour afficher la connectivité réseau et son organisation en communautés selon cette configuration. C'est une option programmée dans EA-SemSNL qui prend comme paramètre l'ID d'un acteur, étant donné un 'Anomalous centrality'. L'un des résultats obtenus de cette interprétation communautaire est dans la **Figure 123**. En se basant sur les interactions asynchrones, A12 est identifié comme un pont reliant deux régions cohésives : 2 communautés, (**Figure 123**). Dans le cas où A12 est inactif, la communication est interrompue entre ces deux communautés. C'est exemple typique d'acteur, d'un trou structurel qui présente un avantage informationnel, entre des communautés d'apprentissage. C'est un canal potentiel de communication, privilégié pour fournir des nouvelles informations (non redondantes) pour ces communautés. Tandis que l'information dans chacune de ces communautés est partagée (contacts redondants). Mais, les apprenants les plus proches de A12, ses voisins: A7 ou A13 (**Figure 123**) constituent les frontières de ces communautés. Ils sont mieux informés et susceptibles d'avoir des nouvelles idées avec un profil cognitif plus avancé.

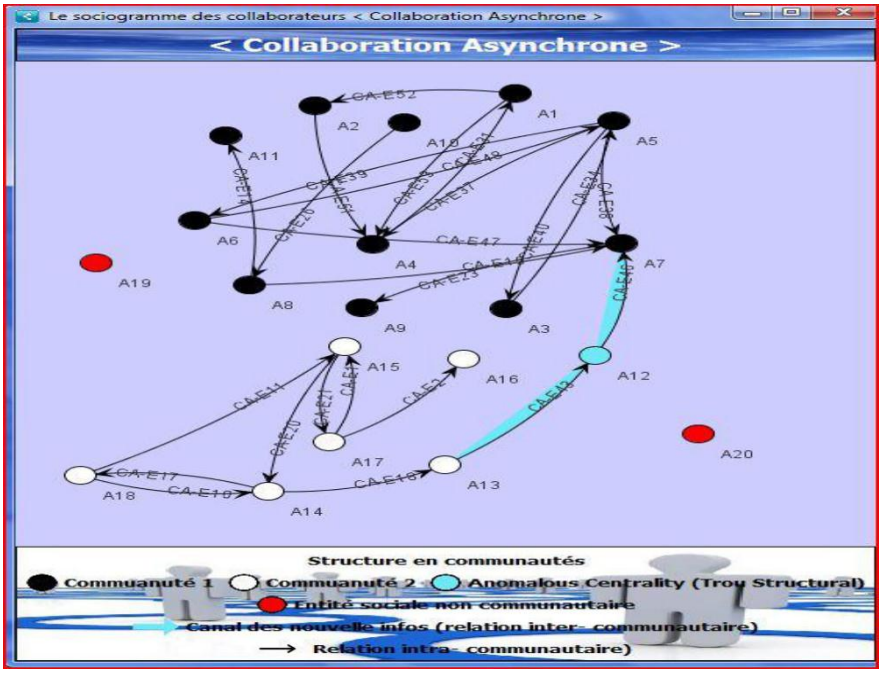
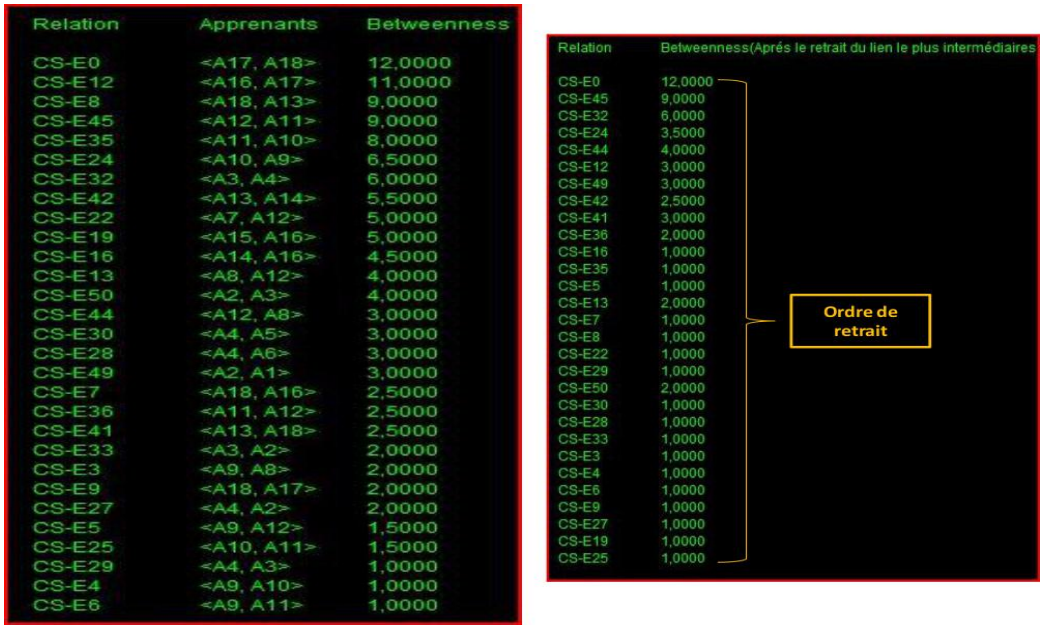


Figure 123. Configuration de communautés d'apprentissage détecté par EA-SemSNL en se basant sur des Ponts bridges (Anomalous centrality) donné comme paramètre (A12)

Troisièmement, nous avons développé une autre composante dans EA-SemSNL pour découvrir des communautés d'apprentissage plus significatives en évaluant les liens et leur sémantique. Comme un nœud, il est possible d'évaluer dans quelle mesure une relation synchrone ou asynchrone est stratégique (intermédiaire) sur les flux de communication. Dans ce cas, les liens stratégiques qui affichent des scores d'intermédiarité supérieurs représentent probablement les points de ***sensibilité du réseau***. EA-SemSNL lance une détection de communauté après avoir calculé itérativement le 'Edge betweenness' de tous les liens et les classer par ordre décroissant (**Figure 124**). L'algorithme itératif de division de (Newman & Girvan 2004) est implémenté dans l'application et programmé pour effectuer une sorte d'attaque ciblée qui supprime les liens les plus intermédiaires. Ces liens sont habituellement des liens intercommunautaires (**Figure 125**, **Figure 126**) dont la suppression affecte sérieusement la connectivité du réseau et le divise efficacement. Après chaque retrait, le programme recalcule les scores et reclasse les liens. Il offre aussi la possibilité de régler des paramètres comme le nombre de liens choisi à supprimer qui détermine ainsi le nombre des itérations et donc plus de flexibilité de l'application.



Relation	Apprenants	Betweenness
CS-E0	<A17, A18>	12,0000
CS-E12	<A16, A17>	11,0000
CS-E8	<A18, A13>	9,0000
CS-E45	<A12, A11>	9,0000
CS-E35	<A11, A10>	8,0000
CS-E24	<A10, A9>	6,5000
CS-E32	<A3, A4>	6,0000
CS-E42	<A13, A14>	5,5000
CS-E22	<A7, A12>	5,0000
CS-E19	<A15, A16>	5,0000
CS-E16	<A14, A16>	4,5000
CS-E13	<A8, A12>	4,0000
CS-E50	<A2, A3>	4,0000
CS-E44	<A12, A8>	3,0000
CS-E30	<A4, A5>	3,0000
CS-E28	<A4, A6>	3,0000
CS-E49	<A2, A1>	3,0000
CS-E7	<A18, A16>	2,5000
CS-E36	<A11, A12>	2,5000
CS-E41	<A13, A18>	2,5000
CS-E33	<A3, A2>	2,0000
CS-E3	<A9, A8>	2,0000
CS-E9	<A18, A17>	2,0000
CS-E27	<A4, A2>	2,0000
CS-E5	<A9, A12>	1,5000
CS-E25	<A10, A11>	1,5000
CS-E29	<A4, A3>	1,0000
CS-E4	<A9, A10>	1,0000
CS-E6	<A9, A11>	1,0000

Relation	Betweenness(Aprés le retrait du lien le plus intermédiaires
CS-E0	12,0000
CS-E45	9,0000
CS-E32	6,0000
CS-E24	3,5000
CS-E44	4,0000
CS-E12	3,0000
CS-E49	3,0000
CS-E42	2,5000
CS-E41	3,0000
CS-E36	2,0000
CS-E16	1,0000
CS-E35	1,0000
CS-E5	1,0000
CS-E13	2,0000
CS-E7	1,0000
CS-E8	1,0000
CS-E22	1,0000
CS-E29	1,0000
CS-E50	2,0000
CS-E30	1,0000
CS-E28	1,0000
CS-E33	1,0000
CS-E3	1,0000
CS-E4	1,0000
CS-E6	1,0000
CS-E9	1,0000
CS-E27	1,0000
CS-E19	1,0000
CS-E25	1,0000

Figure 124. Classement des liens 'CS' selon les scores d'intermédiarité par ordre décroissant et le nouveau classement (à droite) après avoir retiré le lien le plus intermédiaire.

Donc, la configuration des communautés résultante varie en fonction du nombre des liens à éliminer (**Figure 125**) ainsi que le type de lien sélectionné (**Figure 126**). Dans le réseau d'interactions asynchrones, on obtient 4 communautés après avoir éliminé les deux liens les plus intermédiaires (**Figure 125**). Mais, quand on enlève les 3 liens asynchrones les plus intermédiaires, on obtient 5 communautés (**Figure 125**). Même si nous gardons le même nombre de liens à retirer, la composition des communautés varie suivant la nature (la sémantique) des relations (**Figure 126**). Dans ce cas, EA-SemSNL peut détecter des communautés, chacune (sa connectivité) est formée par liens de même type, ce qui la donne une dimension sémantique. Par conséquent, on peut ***découvrir une communauté d'apprentissage dont l'esprit de collectivité, et ce sentiment d'appartenance étant une orientation sémantique horizontale est renforcée***. On constate par ailleurs que ces communautés peuvent se chevaucher au même temps. La **Figure 126** montre que l'acteur A13 à deux communautés en même temps, selon ses interactions synchrones ou asynchrones.

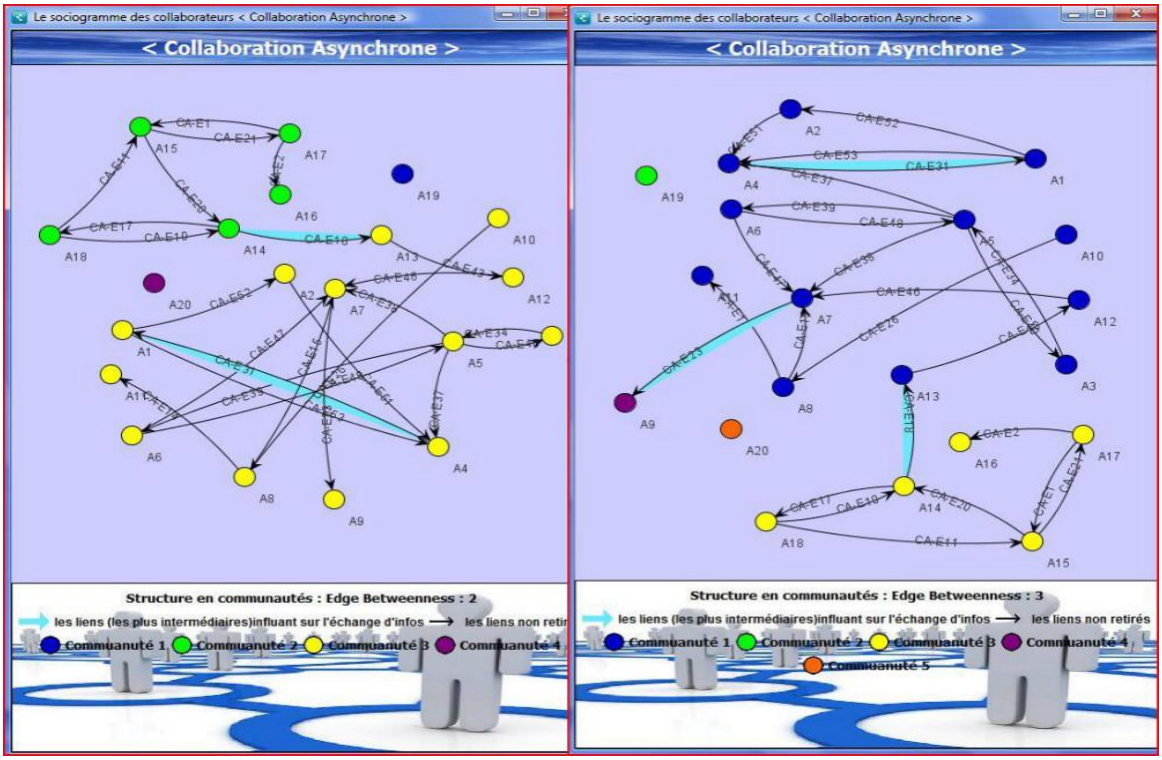


Figure 125. Avec le même type de relations (asynchrones), détection des structures communautaires (par EA-SemSNL) qui varient selon le nombre (2 ou 3) de liens à retirer

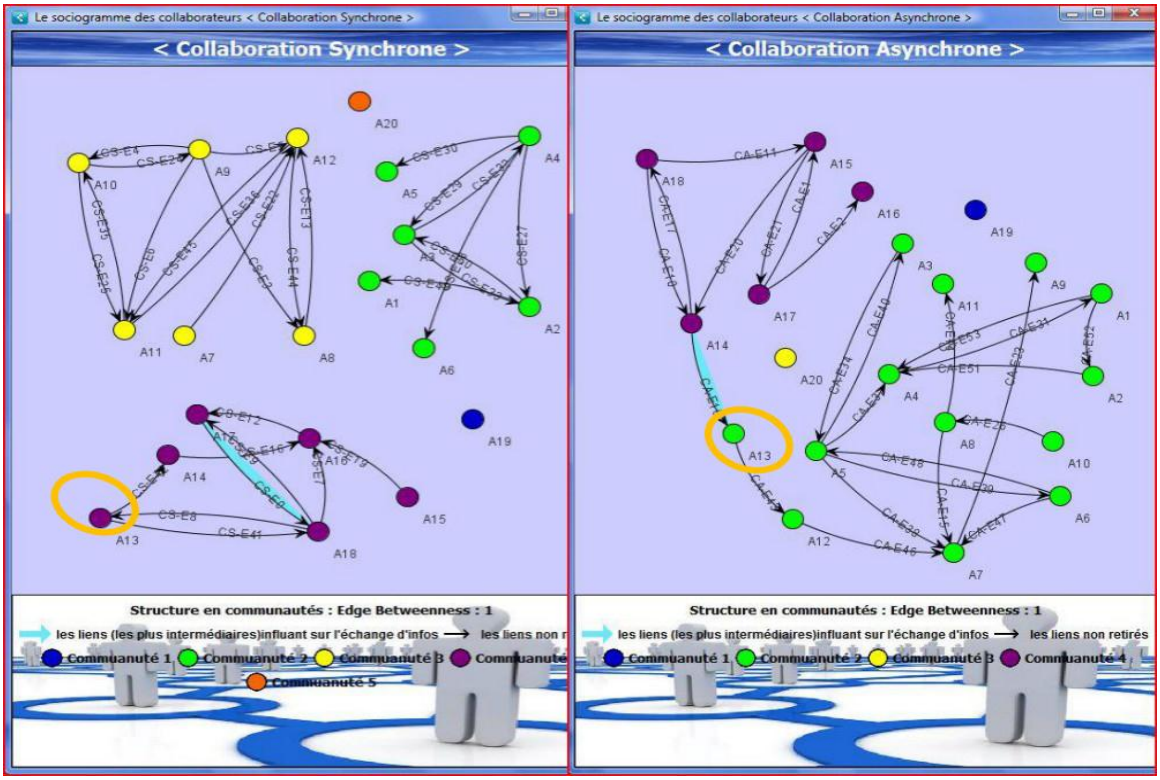


Figure 126. Communautés différentes détectés par EA-SemSNL selon le type des relations

3.4.Discussion

Nous avons décrit un prototype expérimental implémenté dans EA-SemSNL, appliqué sur un modèle sémantique de SN d'apprenants collaborateurs. En tant que premiers utilisateurs, nous avons réalisé une série d'expériences pour analyser ce modèle de SN et montrer les bénéfices informationnels portés par l'exploitation de sa sémantique pour enrichir cette analyse. Au même temps, nous avons décrit les principales fonctionnalités et options proposées par notre application. On s'est basé sur une méthodologie d'évaluation qui part d'un plan d'analyse locale vers l'organisation global du réseau, sa centralisation et sa connectivité. L'analyse individuelle (locale) est plus précise et enrichie. Le potentiel de l'apprenant (centralité ou prestige) est évalué selon différents points de vue qui viennent de cette richesse sémantique axée notamment sur les relations (type et orientation). Pour un type d'interaction sélectionné on peut savoir si apprenant est plus ou moins actif ou prestigieux, s'il est proche/accessible du reste du réseau ou autonome (indépendant), s'il joue un rôle d'intermédiaire, et quelle est la probabilité d'initier une interaction avec lui.

Les résultats de cette analyse peuvent être exploités pour étendre et enrichir aussi les données sociales. On peut évaluer qualitativement le comportement et les compétences sociales (profil social, positivité, etc.) de chaque apprenant suivant les scores de sa centralité et prestige. Par exemple, la centralité de degré permet de déduire son profil social: Un degré élevé signifie un bon collaborateur, un degré nul, signifie un apprenant isolé. Par ailleurs, le degré de prestige ('In-Degree') reflète la positivité de l'apprenant. Un score élevé de 'In-Degree' signifie une attitude positive de l'acteur envers les interactions initiées par les autres dans le réseau. De plus, lorsque le 'In-Degree' est important, il peut être un signe d'un bon profil cognitif, mais pas toujours. Autrement dit, un excellent apprenant peut recevoir de nombreuses demandes d'interactions (de collaborations), mais il ou elle maintient que peu d'interactions. D'autre part, le 'Out-Degree' permet d'interpréter correctement l'appréciation d'un apprenant envers ses interactions. Son appréciation ne sera pas classé agréable à moins qu'il initie de nombreuses interactions (Out-Degree important). Si l'activité ou le potentiel d'un apprenant varie en fonction du type de lien, son profil social, sa positivité, appréciation, etc., changent également.

Dans ce sens, EA-SemSNL offre une autre option capable de classer les acteurs (étant des objets manipulés par l'application) ayant des caractéristiques similaires: la même positivité, etc., ou autre comme le même âge, sexe, etc. Donc différentes classifications peuvent être obtenues. Par exemple, les apprenants sont classés en 5 catégories suivant les niveaux de positivité. Tant que l'indice 'In-Degree' reflète la positivité, EA-SemSNL est capable de calculer la somme de 'In-Degree' (le degré de positivité) des apprenants appartenant à la même classe pour un mode d'interaction donné (**Figure 127**). Cela nous permettra de constater s'il y a un changement de positivité d'une partie d'apprenants si on change le type e relations (**Figure 127**) : Donc apercevoir ***le changement du comportement d'une collectivité qui est basée sur une notion sémantique qui concerne le profil social de ces acteurs.*** On remarque que le degré de positivité des apprenants qualifiés comme positifs ou 'normal' envers les interactions synchrones, diminue considérablement (chute d'In-Degree) avec les interactions asynchrones (**Figure 127**). Cependant, les apprenants qui classés comme négatifs ou très négatifs par rapport aux liens synchrones, deviennent plus positifs envers le mode asynchrone (**Figure 127**).
De la même façon, cette option permet de regrouper sémantiquement ces apprenants en d'autres catégories, par exemple filles et garçons, et comparer leur potentiel réuni (Degré, proximité et intermédiarité). Non seulement le mode d'interaction, mais aussi l'identité de l'apprenant (sexe, âge, etc.) a une influence sur sa centralité ou son prestige dans le réseau.

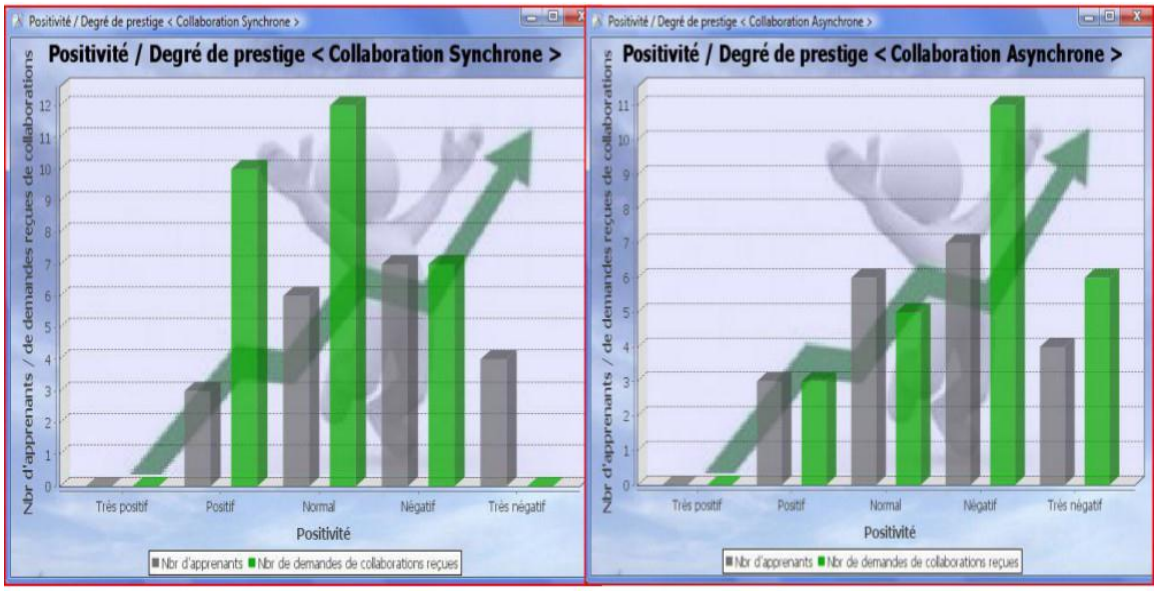


Figure 127. Le changement de positivité des classes d'apprenants devant le changement de type de relations

Deuxièmement, le Tableau 40 montre que les apprenants identifiés comme stratégiques qui jouent des rôles de leadership (et même les moins actifs ou isolés) changent selon différents points de vue : le plus central ou prestigieux au voisinage (Degré, In-Degree, etc.) ou sur l'ensemble du réseau (Betweenness, PageRank, etc.) en spécifiant ou pas le type et l'orientation de liens. EA-SemSNL peut afficher ces positions stratégiques sur le réseau (**Figure 128**). Un apprenant-acteur peut jouer un rôle central au voisinage et en même temps sur l'ensemble du réseau, notamment quand on sait que le degré et l'intermédiarité ainsi que le degré de prestige et PageRank sont corrélés.

Tableau 40. Quelques acteurs stratégiques et dominants le réseau selon différents points de vue détectés par EA-SemSNL

Type de relation	Degree Centrality	Betweenness Centrality	In-Degree	PageRank
Synchrone	A12	A18	A12	A12
Asynchrone	A5	A14	A7	A4
N'est pas précisé	A4	A12	A12	A12

Dans ce sens, l'application peut étudier et afficher des corrélations et distributions entre ces métriques: Closeness et l'indépendance (ou Out-Degree), degré et Eigenvector, etc., comme dans la **Figure 122**. Par ce moyen, il possible de vérifier si le potentiel local d'un apprenant influence ou il est influencé par le reste du réseau (acteurs qui ne sont pas ses voisins directs). Par exemple, l'une de ces corrélations prouve que entre le PageRank est proportionnel au 'In-Degree'. Mais il y a des cas particuliers où l'apprenant devient plus prestigieux quand ses collaborateurs ayant initié ses interactions sont déjà prestigieux. En outre, comme le montre la **Figure 122**, la corrélation entre le degré et l'intermédiarité donne un aperçu sur la connectivité du réseau qui change à son tour selon le type des liens. Une connectivité qu'on a trouvé dominée par des acteurs (avec des centralités anormales) qui contrôlent des trous structuraux ayant des atouts majeurs entre les communautés d'apprentissage.

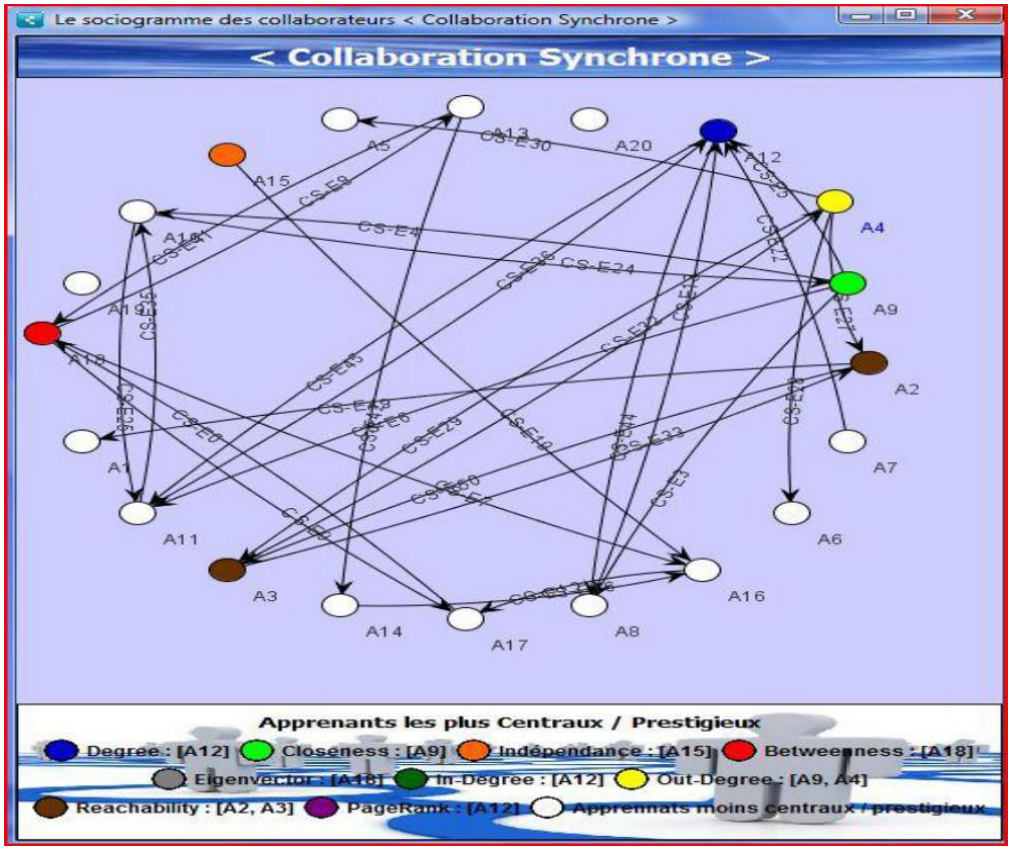


Figure 128. Les apprenants stratégiques, dominants (les plus centraux \ prestigieux) qui jouent des rôles de leadership sur les liens et les flux de communication synchrones au voisinage ou sur tout le réseau

Troisièmement, en ce qui concerne les communautés, EA-SemSNL exploite la sémantique des liens pour les détecter de manière plus significative. On répond d'abord aux exigences topologiques (la cohésion interne) du concept de groupe/ de communautés en se basant sur une méthode de division qui procède par un retrait ciblés des liens (les plus intermédiaires). Ensuite, la division peut être supervisée par le type de lien. C'est-à-dire, l'esprit de collectivité dans chaque communauté se présente non seulement par une composante topologique connectée, mais renforcé en partageant le même type de relations entre ses membres. Pour une communauté plus orientée comme la communauté d'apprentissage, c'est une qualité supplémentaire.

La méthode de division nous a inspiré aussi à implémenter un module optionnel qui permet d'étudier et afficher la résistance/ sensibilité du réseau face à la fragmentation en communautés d'apprenants. Dans ce sens, on donne à l'utilisateur de l'application la faculté de choisir entre une suppression (attaque) ciblée ou aléatoire des liens synchrones ou asynchrones et voir si le réseau résiste ou se fragmente comme c'est montré dans la **Figure 129**. Le taux de fragmentation du réseau est évalué par le nombre de communautés (une granularité en communautés). Comme les résultats précédents, EA-SemSNL affiche cette évaluation sur une fenêtre indépendante, là où on remarque que le réseau n'a pas la même résistance entre les 2 type de relations (**Figure 129**). Plus le nombre de liens stratégiques retirés est important plus le nombre de communautés augmente (**Figure 129**). Cependant, le réseau semble résister plus au retrait ciblé des liens synchrones par rapport aux liens asynchrones, car le nombre de communautés est légèrement plus stable (donc plus cohésives) dans le premier cas. Ici, une justification réaliste peut se tirer de la sémantique de ces relations. En effet, quand on s'implique dans une discussion synchronisée, l'apprenant tend à devenir plus active, plus

proche avec plus de sentiment d'appartenance envers sa communauté qui est susceptible d'être plus cohérente et résistante. D'un autre côté, si les liens sont aléatoirement retirés, le réseau des interactions asynchrones est plus résistant à la fragmentation. De toute façon, si le réseau est plus large, il est moins probable de tomber (de retirer) aléatoirement sur un lien stratégique.

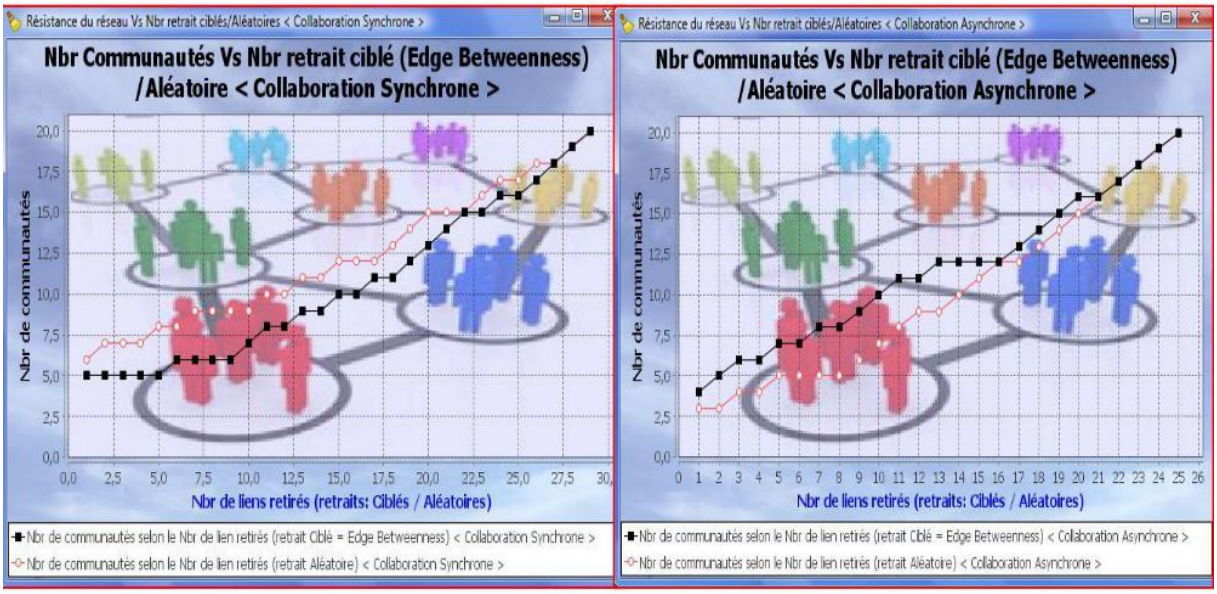


Figure 129. Fragmentation du réseau en termes de nombre de communautés devant le nombre des liens de manière ciblé ou aléatoire

Finalement, le prototype expérimental dans EA-SemSNL inclut des options pour évaluer certaines caractéristiques de ce SN. L'utilisateur peut afficher la densité, le diamètre, etc. du réseau, qui sont paramétrés par le type de relation. On a trouvé par exemple que les interactions synchrones sont plus denses que celles qui sont asynchrones (**Figure 130**), ce qui ouvre la voie à plusieurs d'autres interprétations plus enrichies.

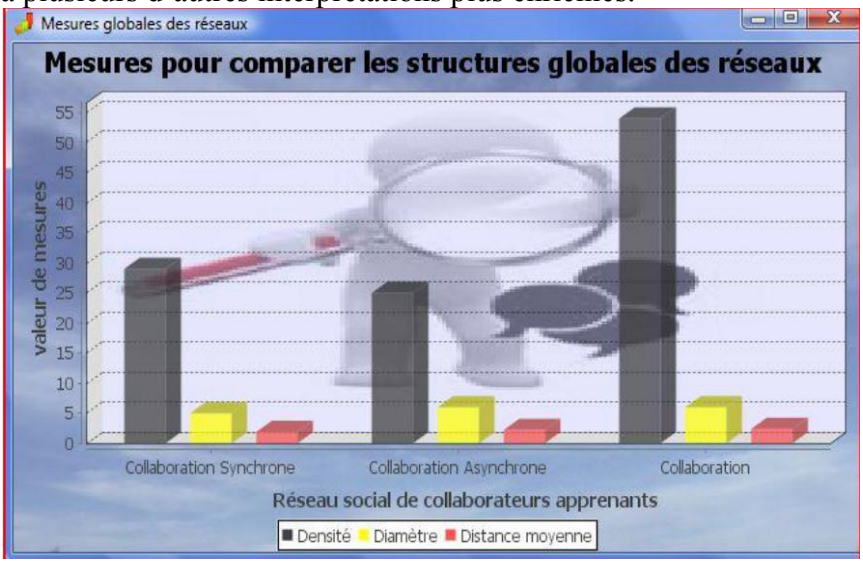


Figure 130. Caractéristiques du SN d'apprenants qui varient selon le type des relations

4. Conclusion

Le SN est une structure précieuse pour comprendre la socialisation des apprenants dans les environnements de e-learning social. Au-delà de SNA classique, nous avons montré

qu'une modélisation et analyse sémantique d'un SN d'apprenants ont des bénéfices informationnels et mènent à des interprétations plus significatives. Un modèle sémantique a été proposé pour améliorer la représentation de ce SN. Il constitue les données d'entrée d'un processus de 'mapping' ainsi qu'un prototype expérimental implémentés dans une application logicielle EA-SemSNL. L'application est en mesure d'exploiter la richesse sémantique de ce modèle (de façon moins coûteuse) et de rendre les études analytiques plus enrichies, paramétrées et plus flexible : Analyse locale, identification des rôles leaderships, corrélations, connectivité et détection des communautés.

Les résultats montrent que le profile social de l'apprenant, son potentiel et sa capacité de dominer le réseau sont influencés par la sémantique de l'interaction : Son type et la positivité de l'initier. Donc, selon différents points de vue, on trouve différents apprenants collaborateurs qui jouent des rôles de leaderships différents, certains d'autres sont moins actifs ou isolés. En identifiant ces positions, un modérateur (ou un moteur de recherche de collaborateurs) dans un environnement de e-learning social peut recommander plus précisément les meilleurs collaborateurs, réduire le sens d'isolation et encourager le sentiment d'appartenance à une communauté***. La sémantisation de l'analyse nous a permis aussi de comprendre mieux la formation des communautés (d'apprentissage) qui n'est pas seulement topologique mais sémantiquement cohérentes et significative aussi***. L'esprit de collectivité est renforcé en partageant le même type de relation et certaines peuvent se chevaucher au même temps. La division de ce réseau (en communautés d'apprentissage) nous a conduits aussi à étudier la résistance de sa connectivité qui varie également selon la sémantique des relations. La causalité de ces liens est un autre aspect sémantique à considérer. Des facteurs hétérogènes sous-jacents peuvent être derrières la création des liens de même type (***hétérogénéité sémantique***).

Conclusion, Perspectives & Challenges

Les travaux en SNAM que nous avons synthétisé ainsi que notre orientation de recherche décryptée sous la lumière de nos contributions se sont basés sur des éléments clés, des motifs d'analyse plus intéressants, plus de pertinence, et de richesse de données sociales qui impliquent plus de dimensionnalité. Nos approches ont été beaucoup plus conceptuelles révélant la volonté de sonder profondément dans le SN en exploitant sa dynamique temporelle et sa sémantique. L'avantage ne se contente pas de montrer par exemple comment caractériser et identifier une identité significative d'un noyau dans le processus évolutionnaire d'un SN ou analyser sa sémantique dans des environnements organisationnels, mais il nous semble avoir amélioré notre capacité de comprendre et distinguer plusieurs sujets, concepts et phénomènes 'hot topics' de ce domaine et dégager même d'autres perspectives et challenges.

À titre d'exemple, il parait que les termes communauté, groupe, collectivité reflètent la même chose mais conceptuellement, nous les avons abordé différemment chacun dans un contexte précis. En avançant dans la thèse, le concept de communauté était beaucoup plus original émergeant dans un contexte statique. En ajoutant la dimension temporelle dynamique, il est devenu plus latent et générique en laissant la place à la notion de groupe qui représente l'une de ses traces à un moment donné. Selon un niveau d'abstraction plus élevé (sémantique), ce groupe n'incarne que topologiquement l'aspect (esprit) de collectivité qui se réfère à une identité implicite derrière sa cohésion topologique, car cette dernière n'est qu'une représentation des flux et des patterns de relations.

Sémantique des groupes et des structures sous-jacentes

Partant des propositions de sémantisation de SNA, on se demande comment peut-on capter le caractère sémantique d'une structure sous-jacente comme le noyau, étant une classe élite. Un noyau sémantique se caractérise souvent dans des systèmes complexes qui sont décrits par des modèles de haut niveau (méta-modèles) basés sur des ontologies, des règles de métier, etc. La sémantique d'un noyau dans un SN dépend du contexte, là où le réseau émerge ou évolue (OSN en ligne/organisationnelle). Pour arriver à ce stade là (sémantique de structure noyau), on a vu qu'il y a des étapes à franchir (**Figure 131**): Modélisation sémantique du SN (étant un méta-modèle aussi), comment exploiter la richesse sémantique exprimée en analyse. On ne se contente pas d'enrichir les métriques d'analyse, mais on veut extraire à partir la caractérisation sémantique des groupes/ des communautés, une identité de noyau (étant un regroupement dominant) sémantiquement significative (**Figure 131**).

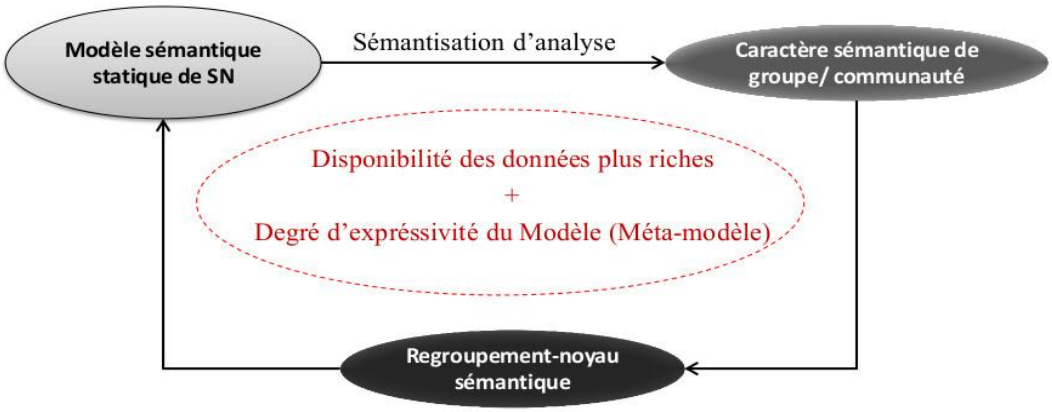


Figure 131. Recherche d'un caractère sémantique d'une structure noyau

Dans les parties précédentes, la sémantisation de SNA ou l'analyse sémantique (modèles, approches d'analyse, application, etc.) exige souvent une compréhension plus approfondie du SN mais aussi de son contexte. Le degré d'expressivité adopté dépend aussi de la disponibilité des données. Par exemple, le modèle sémantique proposé précédemment n'est pas le plus expressif, mais basé sur un vocabulaire améliorable selon des niveaux d'abstraction plus élevés. La sémantique s'incarne dans la causalité des connexions, la positivité à se socialiser, le type de relation, les intérêts, etc. Sur ce plan, nous avons vu que la formation d'une communauté/ groupe peut avoir une dimension sémantique, cette orientation sémantique verticale plus explicite. C'est-à-dire, cet esprit de collectivité est renforcé en partageant le même type de relation, le même intérêt exprimé, etc. Suivant cette orientation, la **Figure 126** a montré par exemple que ces communautés peuvent se chevaucher sémantiquement au même temps. Ces zones de chevauchements sémantiques (statiques) incluent des individus ayant des scores de centralités élevés. Topologiquement, ça peut refléter l'une des conceptions classique d'une structure noyau**.** Mais, une telle zone est sémantiquement plus significative car elle relie des regroupements sémantiquement différents.

Prenant un niveau d'abstraction (degré d'expressivité) plus élevé, là où l'esprit de collectivité d'un regroupement s'exprime dans les intérêts (tags) de ces membres et pas seulement au niveau de sa connectivité (typée). Tout d'abord, les données sociales doivent être suffisamment plus riches (social tagging) pour concevoir des modèles également plus riches avec ce degré d'expressivité. Par exemple, certains environnements de e-learning incluent récemment le 'tagging' collaboratif une sorte de 'social tagging' qui permet aux apprenants d'exprimer leurs intérêts. C'est un moyen pour améliorer la détection sémantique des communautés (détection des communautés sémantiques). Si une communauté est sémantiquement formée par des acteurs qui partagent le même tag (**Figure 132**), une identité sémantique d'un regroupement noyau peut être inspirée de la structuration sémantique des liens entre ces tags (**Figure 132**). C'est-à-dire, un noyau sémantique peut se présenter par un regroupement d'individus qui partagent déjà le même intérêt. Ce dernier occupe un positionnement sémantique particulier (central) par rapport les autres intérêts/ sujets (On peut même parler dans ce cas de centralité sur les liens sémantiques entre tags): Une zone où les autres régions du réseau se croisent sémantiquement (**Figure 132**).

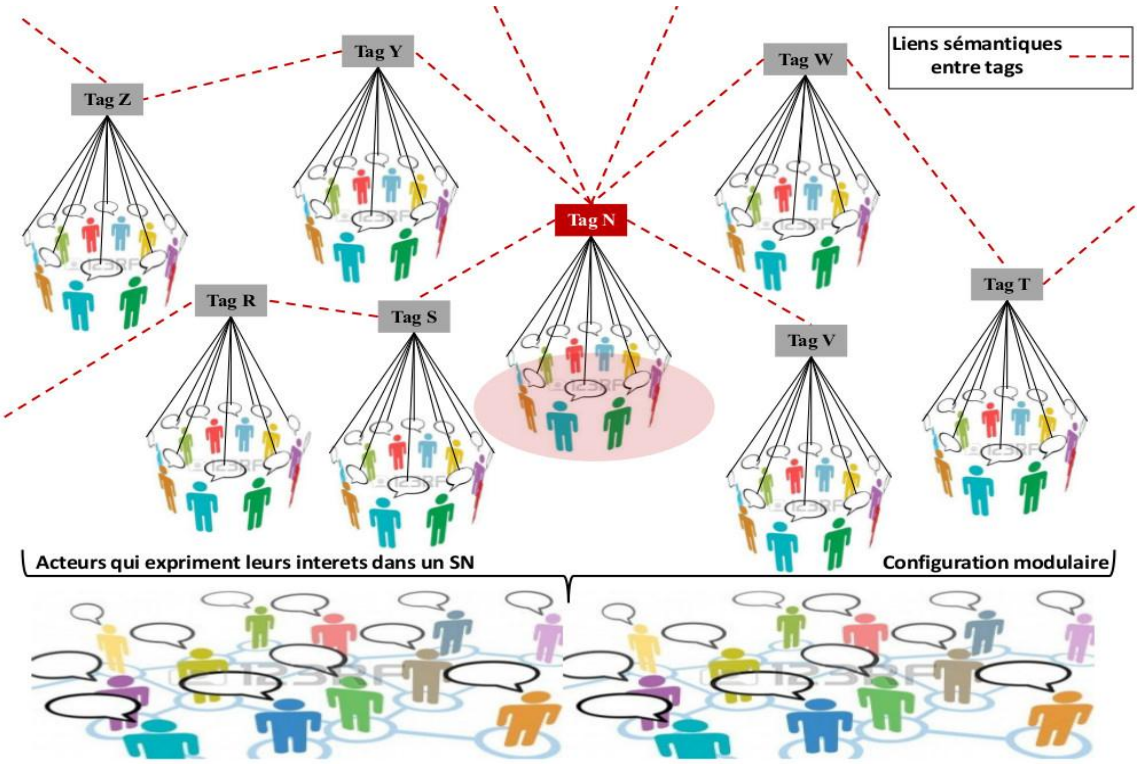


Figure 132. Un regroupement d'individus partageant le même tag qui est sémantiquement le plus lié avec les autres tags

On ne va pas considérer que cette proposition est encore l'image de la fusion entre le modèle sémantique des SNs et topologique/ structurelle de l'analyse/ fouille à moins que l'infrastructure du noyau qui est basée sur le concept de groupe est toujours préservée (**Figure 133**). Mais la configuration statique ne nous laisse focalisés que sur une dimension sémantique plus verticale, explicite et externe. Cette configuration figée ne serait pas adaptée pour aborder par exemple la durabilité d'un noyau dans le temps et donc cette sémantique horizontale qui la contrôle implicitement. Donc la dynamique temporelle (topologique) ne mène pas juste à comprendre et caractériser topologiquement un phénomène complexe sous-jacent comme la structure noyau (**Figure 133**), mais ce sentiment d'appartenance, de dominance dans le temps incarne cette sémantique implicite.

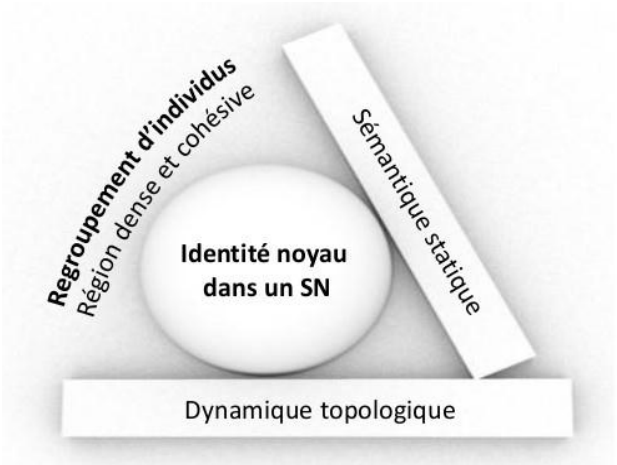


Figure 133. Identité significative d'un noyau entre la sémantique statique et dynamique topologique

Réduire la complexité du traitement sémantique

Un algorithme de détection de communautés sémantiques comme SemTagP est l'illustration des problèmes de complexité en temps et en espace de calcule qui se posent fréquemment dans les processus d'analyse avancée des graphes RDF. Si un mapping vers un graphe étiqueté orienté préserve la sémantique d'un graphe RDF donné, peut-on trouver une approche (un graphe biparti tag-user, etc.) permettant de réduire la complexité et l'utilisation des langages de web sémantique.

Dynamique et durabilité et des regroupements

Revenant au plan topologique, comment saisir et quantifier plus précisément la durabilité/stabilité (durée de vie) d'une structure/ d'un regroupement est l'un des sujets phares des études récentes notamment dans un SN sensible aux contextes sociaux. On peut s'inspirer par exemple de certaines théories entourant la recherche des sous-graphes fréquents pour fournir des réponses précieuses.

L'émergence des comportements collectifs est l'un des aspects liés à la dynamique des structures sociales du monde réel qui se classent encore parmi les 'hot topics' en SNAM. Les chercheurs s'intéressent de plus en plus à l'évolutivité de ces systèmes réseaux complexes, sa modélisation, les infrastructures sophistiquées et l'automatisation de l'analyse, etc., qui ont une influence directe sur le succès de la recherche dans la dynamique des réseaux en général : par exemple la diffusion ou la propagation des idées, des mouvements ou des maladies sur ces réseaux. Autre exemple, après la crise financière de 2007-2008, une analogie plus profonde a été soulignée entre les origines de l'instabilité dans les systèmes financiers et dans ces écosystèmes complexes En particulier, les caractéristiques topologiques des structures de réseau influencent la façon dont elle se propager la détresse dans le système.

D'autre part, les dernières années ont connu un développement systématique des approches algorithmiques dédiées à l'analyse des réseaux dynamiques. Mais selon (Berger-Wolf & Saia 2006), la plupart sont concentrés sur la fréquence, plutôt que de concurrence et l'ordre d'interactions (Berger-Wolf & Saia 2006).

L'exploration des contextes sociaux

L'exploration des contextes sociaux, notamment cultuels dans les modèles des SNs est aussi l'un des progrès, vu que beaucoup de travaux théoriques en SNAM n'incluent pas une théorie sociologique (Culture) ((Scott 2011)).

Fusionner l'aspect sémantique et la dynamique temporelle (multi-dimensionnalité)

L'étude des phénomènes sous-jacents (noyau) montre la complémentarité qui existe entre ces deux aspects dynamique et sémantique, même si les deux sont souvent abordés indépendamment. Par exemple, une identité d'un noyau a été décrite par comportement dynamique typique ou un caractère sémantique statique particulier d'un regroupement d'individus (**Figure 134**). Les chevauchements soit temporels ou sémantiques ont été montrés comme l'un des moyens efficaces pour exprimer une sorte d'identité externe à partir d'une connectivité/ identité interne (**Figure 134**).

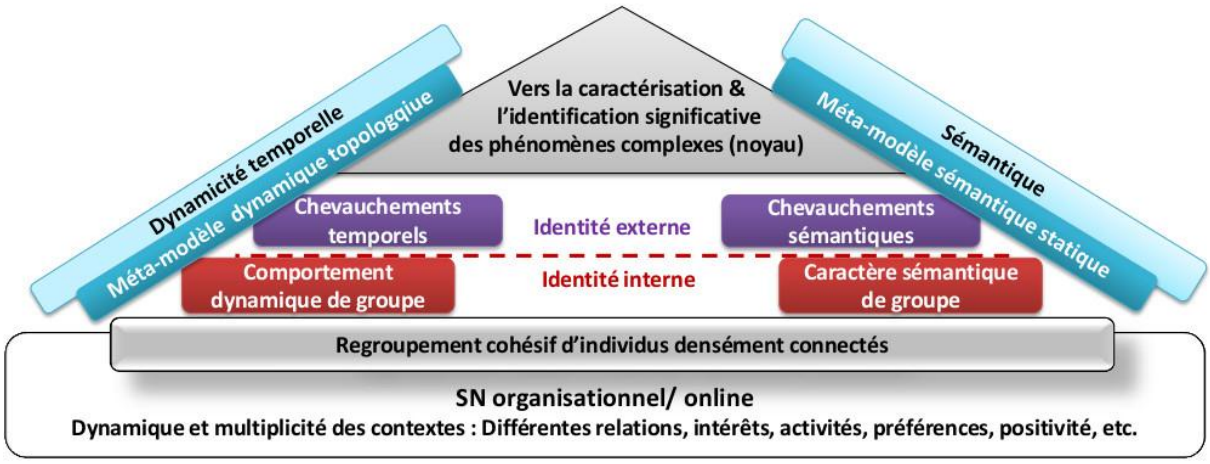


Figure 134. Dynamique temporelle topologique et sémantique, groupes et chevauchements pour comprendre des phénomènes complexes dans les SNs

Au-delà d'une simple analyse, ces hypothèses et propositions de rapprochement ou de fusion entre les deux aspects sont susceptibles de multiplier la complexité notamment quand il s'agit d'étudier des réseaux (OSNs) à grande échelle, avec une diversité d'interactions et activités, évoluant pendant des longues périodes d'observation. Des méta-modèles quasi continues ou à une expressivité massivement riche semblent être infaisables.

D'un point de vue simple concernant cette fusion, des auteurs essayent de découvrir les propriétés d'évolution temporelle des patterns de SN sur le web sémantique ((Zhou et al 2011)). D'autre part, nous pensons à modéliser l'évolution temporelle d'un graphe social sémantique (RDF) par une séquence d'empreintes statiques dans des 'time windows'. (Réseau RDF temporel). Ce qui permet par exemple de suivre l'évolution et le changement temporel des rôles (indices) des acteurs, et pourquoi pas l'évolution des communautés sémantiques dans le temps selon un type de relation donné. Dans ce sens, en identifiant des nœuds avec des intérêts, ((Asur et al 2007)) pensent à intégrer ce contenu sémantique dans le modèle d'évolution de communautés à base évènements qu'ils ont proposé. Par ailleurs, il faut citer que Rowe et Strohmaier (2014) ((Rowe et al 2014)) ont récemment proposé une étude de l'évolution sémantique des communautés en ligne. Selon eux, la communauté évolue également au niveau sémantique et ses concepts (ses intérêts, ses discussions, etc.) émergent, ce qui est largement ignoré. Donc ils avaient pour objectif de capturer la dynamique des communautés sur un niveau conceptuel.

Mais la modélisation est toujours un énorme challenge. Pour certains chercheurs, cette fusion peut être définie par des modèles sémantiques (ontologiques) qui encapsulent l'aspect temporel dynamique. On parle à titre d'exemple d'une syntaxe RDF augmentée comme celle de 'Dublin core'. Ils pensent à ajouter des primitives temporelles qui permettent d'insérer directement des données temporelles dans les annotations RDF, et donc concevoir un graphe social sémantique temporel. Ce dernier ouvre la voie à plusieurs questions qui cherchent l'approche d'analyse capable de prendre d'avantage à la fois de la sémantique et des données temporelle de tel graphe RDF, par exemple pour la prédiction des liens sémantiques ((Takes 2011)).

La multi-dimensionnalité est aussi nouveau paradigme en SNA/ SNAM qui nécessite modèles (méta-modèles) multidimensionnels. Elle semble avoir attiré les chercheurs pour comprendre les causes et les effets des différents phénomènes complexe (propagation des

maladies, d'information) qui ne peuvent pas être inférés sur une seule dimension. Elle combine différents type de données et offre une vision plus large sur les interactions et activités humaines. Une approche d'analyse multidimensionnelle comme celle proposée par (Kazienko et al 2011) est déjà montrée prometteuse pour répondre à des questions posées par exemple par une entreprise sur une nouvelle offre à introduire pour ses clients, la prédiction du type de canal de communication à utiliser et quelles sont les personnes (la communauté) ciblées, etc.

Sensibilité du SN

Nous avons vu aussi que la multiplicité des dimensions et des contextes d'analyse peut enrichir les débats sur des sujets clés comme la sensibilité, fragilité, résistance des réseaux. Une sensibilité testée par rapport à des phénomènes sous-jacents comme la structure noyau dans le temps, des liens stratégiques (intermédiaire) ou à la fragmentation en groupes/ communautés, etc.

Centralités de groupes, Capital social et modération du SN

Les propositions les centralités de groupes peuvent caractériser aussi le 'capital social', une notion liée à la signification et les avantages découlant des liens des individus et leurs groupes dans le SN. En effet des auteurs classifient ces liens en trois catégories, liens forts, faibles et potentiel. Les liens forts sont fréquemment activés (entre personnes qui travaillent dans une même équipe), par rapport les liens faibles. Ils augmentent le partage des ressources et renforcent la confiance (une variable cachée) dans un groupe : Les activités sociales fortement corrélés avec cette confiance ((Srivastava & DeLong 2013)). Cependant, les liens faibles et les SNs 'sparce' facilitent l'accès à des ressources/ sources d'informations plus variées, ils comblent souvent des trous structuraux qui séparent des groupes denses séparés. Les ressources disponibles à travers des liens potentiels (qui représentent une proximité sociale) sont considérées comme le capital social potentiel.

Dans son cercle, l'acteur ne peut pas trouver un nouveau capital social (des ressources supplémentaires). Une action appelée instrumentale lui permet d'y accéder (par exemple via ces trous structuraux). Pour un accès meilleur au capital social il est intéressant de développer des relations avec des gens ayant un faible coefficient de clustering et une centralité d'intermédiarité importante. Mais, les gens ne savent pas forcément les avantages tirés de leur type de relations, leur positionnement, ni comment les développer efficacement. Elles devraient être aidées. La modération ('Community manager', 'SN manager') doit bénéficier des avantages de SNAM e ses progrès pour jouer ce rôle.

Autres perspectives

Comme l'on a déjà cité, il n y a pas que la dynamique et la sémantique pour ajouter plus de dimensionnalité aux études analytiques des SNs, mais les perspectives de SNAM concernent aussi l'utilisation des données de géolocalisation, la prédiction des liens, 'socially-inspired techniques', Social networking in smart environments', etc.

D'autre part, la popularité des OSNs ne fait pas vraiment oublier l'ancienne collection des SNs. Le monde de football (de NBA aussi ((Srivastava & DeLong 2013))) par exemple qui a une immense popularité et un contexte social qui encourage récemment à extraire et analyser des SNs des sélections nationales et des clubs comme dans ((Kooij et al 2009)). La motivation et les bénéfices se multiplient. L'analyse et la visualisation des métriques topologiques et statistiques (des joueurs et du réseau) peuvent être une étape dans la conception des outils qui aident à la décision pour les coaches. Par exemple pour déterminer l'alignement le plus optimal des 11 de départ, ceux qui ont influencés les résultats des matches. Afin de trouver les

automatismes dans la formation de l'équipe, les coaches tendent aussi à choir l'ensemble des joueurs qui ont plus de temps de jeu ensemble, tel que chacun peut anticiper ce que l'autre va faire. C'est l'équivalent par exemple de la recherche des régions cohésives dans le SN de ((Kooij et al 2009)).

Challenge de Big data

Même si on adopte l'une des dimensions d'analyse indépendamment des autres, on tombe aujourd'hui dans un véritable challenge de Big data. L'émergence rapide des OSNs (ex. un billion d'users connectés sur Facebook en 2012 ((Xie & Szymanski 2013))), constitue aujourd'hui la source la plus importante de 'Social Big data'. Par conséquent, des problèmes qu'on croyait résolus en SNAM sont à revoir. Les méthodes utilisées doivent être linéairement évoluées avec la croissance de la taille des OSNs, car la majorité est inapplicable sur ces données massives (des réseaux d'ordre de millions de nœuds). ((Magnusson 2012)) a essayé de déterminer quels sont les algorithmes de SNA qui peuvent avoir cette scalabilité.

D'abord, le Big data nécessite par principe une méthodologie fondamentale pour l'analyse efficace et robuste des données, le traitement et l'extraction de l'information ('Big Data Mining and Analytics'). Les idées s'accentuent sur les parallélisassions, implémenté notamment par le modèle de programmation MapReduce ((Dean & Ghemawat 2008)). Il s'agit de diviser/définir le problème en parties (fonctions mappers/ map) et laisser plusieurs machines traitent chacune des parties séparément en même temps, avant de regrouper les résultats (par des fonctions reduce/reducers) dans un ordinateur puissant, éventuellement à multiples processeurs (ce qui est économiquement moins cher), ((Bello-Orgazet et al 2016)). Google est le pionnier de cette approche avec ses softwares basé sur le paradigme MapReduce et Pregel ((Khan et al 2014)) ((Bello-Orgazet et al 2016)). Alors que Hadoop ((Couldry 2012)) est l'une des alternatives open-source, les plus prometteuses qui trace son chemin en industrie ((Magnusson 2012)).

Donc, la solution consiste à combiner les technologies de Big data et les algorithmes de SNA ((Bello-Orgazet et al 2016)). On sait que les mesures de centralités par exemple sont de haute complexité de calcul dans les réseaux à grande échelle. Pour cela, une deuxième génération d'analyse des graphes à grande échelle se base sur le paradigme de MapReduce est apparue, par exemple Hama, Giraph (basé sur Pregel) qui sont tous les deux des cadres de traitement distribué de graphe ((Bello-Orgazet et al 2016)). D'autre part, la structure modulaire qui est devenue une propriété significative très demandée dans la recherche dans les SNs ou ailleurs. Mais, les algorithmes de détections de communautés ainsi des mesures de qualité de décomposition (modularité) ont montré le challenge de complexité de calcule sur les réseaux à grande échelle. Ce qui oblige aujourd'hui de revoir la détection des groupes cohésifs dans les grands SNs, un moyen pour l'extraction de l'intelligence collective. Par exemple, ((Shi et al 2013)) présente une solution de détection de communautés dans les grands réseaux SNs, en utilisant MapReduce avec un ensemble de méthodes de prétraitement et post-traitement pour optimiser le coût en temps et espace (de stockage) d'exécution et précision de détection.

Bibliographie

(Abel & Leblanc 2008), Abel M-H.,and Leblanc A. "E-MEMORAe2.0 : an e-learning environment as learners communities support." International Journal of Computer Science and Applications (IJCSA), Special Issue on New Trends on AI Techniques for Educational Technologies (2008): Vol. 5, No. 1, pp.108–123.

(Adamic & Adar 2003), Adamic L., Adar E. "Friends and neighbors on the Web." Social Networks (2003): 25(3), pp 211-230.

(Adamic & Glance 2005), Adamic L. A., and Glance N. "The political blogosphere and the 2004 US Election." in Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem. 2005.

(Agarwal et al 2009), Agarwal N., Liu H., Murthy S., Sen A., and Wang X. "A Social Identity Approach to Identify Familiar Strangers in a Social Network." 3rd International AAAI Conference on Weblogs and Social Media (ICWSM09), 2 - 9 May, 2009. San Jose, California, 2009. pp 17-20.

(Ahn et al 2011), Ahn J-w., Taieb-Maimon M., Sopan A., Plaisant C., Shneiderman B. "Temporal Visualization of Social Network Dynamics: Prototypes for Nation of Neighbors." Proceedings of 4th International Conference, SBP (Social Computing, Behavioral-Cultural Modeling and Prediction) SBP'11. College Park, MD, USA: Springer-Verlag Berlin, Heidelberg, March 29-31, 2011. Pages 309-316.

(Aïssani 2009), Aïssani Y. "Changement du noyau central et des éléments périphériques d'une représentation sociale sous l'effet d'un essai contre-attitudinal." Annuaire de Psychologie sociale (2009): vol. 40, nº 2, september 2009, pp. 255-270.

(Alberich et al 2002), Alberich R., Miro-Julia J., and Rossello F. "Marvel Universe looks almostlike a real social network." arXiv :cond-mat/0202174 (février 2002).

(Albert & Barabasi 2002), Albert R., and Barabasi A.-L. "Statistical mechanics of complex networks." Reviews of Modern Physics (2002): 74, 47 .

(An et al 2011), An J., Cha M., Gummadi K. P., Crowcroft J. "Media landscape in Twitter: A World of New Conventions and Political Diversity." International AAAI Conference on Weblogs and Social Media (ICWSM), July 2011. Barcelona, Spain, 2011.

(Anyanwu et al 2007), Anyanwu M., Maduko A., Sheth A. "SPARQL2L:Towards Support for Subgraph Extraction Queries in RDF Databases." Proc. WWW2007. 2007.

(Armstrong-Wright & Mice 1969), Armstrong-Wright, Mice A. T. Critical Path Method: Introduction and Practice. Longman Group LTD, , pp5ff. London, 1969.

(Asimakopoulos 2009), Asimakopoulos J. "Globally Segmented Labor Markets." Critical Sociology (2009): 35(2):175-198.

(Asur et al 2007), Asur S., Parthasarathy S., Ucar D.,. "An event-based framework for characterizing the evolutionary behavior of interaction graphs." In KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM, 2007. pages 913–921.

(Auer et al 2008), Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R., Ives Z. "DBpedia: A nucleus for a web of open data." In Proc. Int. Semantic Web Conf. 2008. pages 722-735.

(Babaei et al 2015), Babaei M., Grobowicz P., Valera I., Gummadi K. P., Rodriguez M.G. "On the Users' Efficiency in the Twitter Information Network." International Conference on Weblogs and Social Media (ICWSM). 2015. Short Paper.

(Backstrom et al 2006), Backstrom L., Huttenlocher D., Kleinberg J., X. Lan X. "Group Formation in Large Social Networks: Membership, Growth, and Evolution." in: Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: KDD, ACM, 2006. pp. 44–54.

(Backstrom et al 2012), Backstrom L., Boldi P., Rosa M., Ugander J., Vigna S.,. "Four degrees of separation." In ACM Web Science 2012: Conference Proceedings. ACM Press, 2012. pages 45−54, Best paper award.

(Baker 2013), Baker S.,. "Critical Path Method (CPM) [Online]." 2013 last changed. Health Services Policy and Management Courses, University of South Carolina, Columbia, SC. 2004. Accessed 30 Mar. 2015 <Hspm.sph.sc.edu. Available at: http://hspm.sph.sc.edu/Courses/J716/CPM/CPM.html>.

(Barnes 1954), Barnes J. A. "Class and Committees in a Norwegian Island Parish." Human Relations (1954): February 1954 7: 39-58, doi:10.1177/001872675400700102.

(Bartz et al 2009), Bartz K., Blitzstein J., and Liu J. "Graphs, Bridges and Snowballs: Monte Carlo maximum likelihood for exponential random graph models: From snowballs to umbrella densities." Unpublished paper, 2009.

(Bastian et al 2009), Bastian M., Heymann S., Jacomy M. "Gephi: An Open Source Software for Exploring and Manipulating Networks." in: Proc. 3rd. Int. AAAI Conference on Weblogs and Social Media. 2009. pp. 361–362.

(Batagelj & Mrvar 1998), Batagelj V., Mrvar A. "Pajek - Program for Large Network Analysis Connections." 1998. 21(1998)2, 47-57.

(Batagelj & Mrvar 2003a), Batagelj V., Mrvar A. "Pajek - Analysis and Visualization of Large Networks." In Juenger M., Mutzel P. (Eds.): Graph Drawing Software, Springer (series Mathematics and Visualization). Berlin: Springer, Amazon, 2003. 77-103. ISBN 3-540-00881-0. PDF.

(Batagelj & Mrvar 2003b), Batagelj V., and Mrvar A. "A density based approaches to network analysis: Analysis of Reuters terror news network." Ninth Annual ACM SIGKDD. Washington, D.C., 2003.

(Batagelj & Mrvar 2006), Batagelj V., and Mrvar A. Pajek datasets. 2006. <<URL: http://vlado.fmf.uni-lj.si/pub/networks/data/>.>.

(Batagelj & Mrvar 2012), Batagelj V., Mrvar A. "Pajek - Program for Large Network Analysis." Home page http://vlado.fmf.uni-lj.si/pub/networks/pajek/, 16 November 2012.

(Beauguitte 2011), Beauguitte Laurent. "Une courte introduction à Pajek." Groupe fmr (flux, matrices, réseaux), 13 p. 2011.

(Bello-Orgaz et al 2012), Bello-Orgaz G., Menéndez H.D., Camacho D. "Adaptive k-means algorit hm for overlapped graph clustering." Int.J.Neural Syst. (2012): 22 (05) (2012) 1250018.

—. "Adaptive k-means algorithm for overlapped graph clustering." Int.J.Neural Syst. (2012): 22 (05) (2012) 1250018.

(Bello-Orgazet et al 2016), Bello-Orgaz G., Jung Jason J., and Camacho D. "Social Big Data: Recent achievements and new challenges, Publication Date." Information Fusion Journal (August 2015-2016): 28 (2016) 45–59, DOI http://doi.org/10.1016/j.inffus.2015.08.005.

(Benz et al 2010), Benz D., Hotho A., Jäschke R., Krause B., Mitzlaff F., Schmitz C., Stumme G. "The social bookmark and publication management system BibSonomy." The VLDB J., dec 2010 (2010): 19(6):849-875.

(Berge 1985), Berge C. "Graphs and Hypergraphs." Elsevier Science Ltd (1985).

(Berger-Wolf & Saia 2006), Berger-Wolf T. Y., Saia J. "'A Framework for Analysis of Dynamic Social Networks." Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, PA, USA, August 2006. 523—528.

(Bhattacharya et al 2014), Bhattacharya P., Zafar M. B., Ganguly N., Ghosh S., Gummadi K. P. "Inferring User Interests in the Twitter Social Network." ACM Recommender System Conference (RecSys). ACM, 2014. Short Paper.

(Blondel et al 2008), Blondel V., Guillaume J-L., Lambiotte R. and Lefebvre E. "Fast unfolding of communities in large networks." Journal of Statistical Mechanics: Theory and Experiment (September 2008): No. 10, P10008, 12pp.

(Boldi et al 2004), Boldi P., Codenotti B., Santini M., Vigna S. "UbiCrawler: A Scalable Fully Distributed Web Crawler." Software: Practice & Experience (2004): volume:34, number: 8, pages: 711-726.

(Boldi et al 2011), Boldi P., Rosa M., Santini M., Vigna S.,. "Layered Label Propagation: A MultiResolution Coordinate-Free Ordering for Compressing Social Networks." Proceedings of the 20th international conference on World Wide Web. publisher: ACM Press, 2011.

(Bonneau et al 2009), Bonneau J., Anderson J., Anderson R., Stajano F. "Eight friends are enough: Social graph approximation via public listings." In SocialNets'09: The Second ACM EuroSys Workshop on Social Network Systems. 2009. pp.13–18.

(Borgatti & Everett 2000), Borgatti S.P., and Everett M.G. "Models of core/ periphery structures." ELSEVIER - Social Networks (2000): Vol. 21, No. 4. (October 2000), pp. 375-395.

(Borgatti 2002), Borgatti S.P. "NetDraw 1.0: Network visualization software." 2002.

(Borgatti et al 2002), Borgatti S.P., Everett M.G., Freeman L.C. Ucinet for Windows: Software for Social Network Analysis. 2002: Harvard, MA: Analytic Technologies, n.d.

(Bothorel & Bouklit 2008), Bothorel C. and Bouklit M. "An algorithm for detecting communities in folksonomy hypergraphs." 8th International Conference on Innovative Internet Community Systems I2CS. Schoelcher, Martinique: Sponsored by IEEE, 2008.

(Boyd & Ellison 2007), Boyd D.M., and Ellison N.B. "Social network sites: Definition, history, and scholarship." Journal of Computer-Mediated Communication (2007): 13 (1) (2007) article 11. Available at: http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html.

(Brandes 2001), Brandes U. "A faster algorithm for betweenness centrality." J. Math. Socio (2001): 25(2): 163-177.

(Brandes et al 2001), Brandes U., Krackhardt D., Tamassia R., Wagner D. "Link Analysis and Visualization." Dagstuhl seminar 01-06 July, 2001. 2001.

(Brandes et al 2007), Brandes U., Delling D., Gaertler M., Gorke R., Hoefer M., Nikoloski Z. and Wagner D. "On fnding graph clusterings with maximum modularity." Graph-Theoretic Concepts in Computer Science (2007): pages 121–132 Springer.

(Brandes et al 2009), Brandes U., Kenis P., Lerner J., van Raaij D. "Network analysis of collaboration structure in Wikipedia." Proceedings of the 18th international conference on World wide web. 2009. 731-740.

(Breiger & Pattison 1986), Breiger R., and Pattison P. "Cumulated social roles: The duality of persons and their algebras." Social Networks (1986): 8, 215-256.

(Breiger 1974), Breiger R. " The duality of persons and groups." Social Forces (1974): 53, 181-190.

(Brožovský & Petříček 2007), Brožovský L., Petříček V.,. "Recommender system for online dating service." In Proc. Znalosti. 2007. pages 29-40.

(Buffa 2008), Buffa M. "Du Web aux wikis: une histoire des outils collaboratifs." http://interstices.info, online journal, issue of 23/05/08 (2008).

(Burt 1992), Burt R.S. "Structural holes. The Social Structure of Competition." Cambridge, ,Harvard University Press (1992).

(Burt 2004), Burt R.S. "Structural Holes and Good Ideas." American Journal of Sociology (2004): 100(2): 339-399.

(Caputo et al 2015), Caputo A., Socievole A., and De Rango F. "CRAWDAD data set unical/socialblueconn (v. 2015-02-08)." Downloaded from http://crawdad.org/unical/socialblueconn, feb 2015, 2015.

(Carlson & Doyle 1999), Carlson J., and Doyle J. "Highly Optimized Tolerance: A Mechanism for Power Laws in Designed Systems." Physical Review E (1999): 60:2.

(Celma 2010), Celma Ò. Music Recommendation and Discovery in the Long Tail. Springer, 2010.

(Ceyhan et al 2011), Ceyhan S., Shi X., Leskovec J. "Dynamics of bidding in a P2P lending service: effects of herding and predicting loan success." In: International World Wide Web Conference Committee (IW3C2), WWW 2011—session: social network analysis, March 28–April 1, 2011. Hyderabad, India, 2011. ACM 978-1-4503-0632-4/11/03.

(Cha et al 2009), Cha M., Mislove A., Gummadi K. P. "A Measurement-driven Analysis of Information Propagation in the Flickr Social Network." World Wide Web Conference (WWW), April 2009. Madrid, Spain, 2009.

(Cha et al 2010), Cha M., Haddadi H., Benevenuto F., Gummadi K.P. "Measuring User Influence in Twitter: The Million Follower Fallacy." International AAAI Conference on Weblogs and Social Media (ICWSM), May 2010. Washington, DC, 2010.

(Chakrabarti et al 2006), Chakrabarti D., Kumar R., Tomkins A. "Evolutionary clustering." In KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM, 2006. pages 554–560.

(Chakrabartiet al 2004), Chakrabarti D., Zhan Y., Faloutsos C. "'R-mat: A Recursive Model for Graph Mining." in: Proc. SIAM Data Mining Conference, SIAM. Philadelphia, PA, 2004.

(Champin & Mrissa 2011), Champin P-A, Mrissa M. Web des données : Linked Open Data. LIRIS, 2010 - 2011.

(Chen et al 2009), Chen J., Osmar R., Zaïane R. and Goebel R. "Local Community Identification in Social Networks." Published in ASONAM '09 Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining. Department of Computer Science, University of Alberta, Canada T6G 2E8: (IEEE Computer Society Washington)., 2009.

(Cheng et al 2008), Cheng X., Dale C., Liu J. "Statistics and Social Network of YouTube Videos." IWQoS. IEEE, 2008. 229-238.

(Chi et al 2007), Chi Y., Song X., Hino K., Tseng B.L. "Evolutionary spectral clustering by incorporating temporal smoothness." US Patent App (2007): October 18 2007, 11/874,395.

(Cho et al 2011), Cho E., Myers S. A., Leskovec J. "Friendship and mobility: User movement in location-based social networks." In Proc. Int. Conf. on Knowledge Discovery and Data Mining. 2011. pages 1082-1090.

(Chuan Shi et al 2013), Chuan Shi, Yanan Cai, Di Fu, Yuxiao Dong, Bin Wu. "A link clustering based overlapping community detection algorithm." Data Knowl. Eng. ELSEVIER (2013): 87: 394-404.

(Clauset et al 2004), Clauset A., Newman M.E.J., and Moore C. "Finding community structure in very large networks." Physical Review E (2004): 70(6):66111.

(Coifman & Lafon 2006), Coifman R.R., and Lafon S. "Diffusion maps." Applied and Computational Harmonic Analysis (2006): 21(1), 5-30, (July 2006).

(Comin 2009), Comin M.N. Réseaux de villes et réseaux d'innovation en Europe : Structuration du système des villes européennes par les réseaux de recherches sur les technologies convergentes. Thèse de doctorat, Université de Paris I Sorbonne. Paris, 2009.

(Corby 2008), Corby O. "Web, Graphs & Semantics." ICCS'2008. 2008.

(Corman et al 2002), Corman S. R., Kuhn T., Mcphee R. D., Dooley K. J. "Studying Complex Discursive Systems." Centering Resonance Analysis of Communication (2002).

(Couldry 2012), Couldry N. Media,Society,World:Social Theory and Digital Media Practice, Polity. 2012.

(Cuvelier & Aufaure 2011), Cuvelier E., Aufaure M-A. "Graph mining & community detection an introduction to social networks data analysis." published in First European Summer School, eBISS 2011. Paris, France, July 3-8, 2011, hal-00704356, version 1 - 5 Jun 2012.

(Davis et al 1941), Davis A. et al. "Deep South Chicago." University of Chicago Press (1941).

(Davis et al 2012), Davis D., Lichtenwalter R., Chawla N.V.,. "Supervised methods for multirelational link prediction." Social Network Analysis and Mining (2012): 1-15.

(De Choudhury et al 2010), De Choudhury M., Lin Y-R., Sundaram H., Candan K. S., Xie L., Kelliher A. "How does the data sampling strategy impact the discovery of information diffusion in social media?" In ICWSM. 2010. pages 34-41.

(De Nooy et al 2004), De Nooy W., Mrvar A., Batagelj V. "Exploratory Social Network Analysis with Pajek." Cambridge: Cambridge University Press (2004): Chapter 5, 11, 12.

(Dean & Ghemawat 2008), Dean J., Ghemawat S. "Mapreduce: simplified data processing on large clusters." Commun. ACM51(1)(2008). doi:10.1145/1327452.1327492., 2008. 107–113.

(Degenne & Forse 1994), Degenne A., Forse M. "Les Réseaux sociaux." (1994): pp.15.

(Dekker 2011), Dekker A.H. "Temporal Social Network Analysis of Discourse." MODSIM 2011, 19th International Congress on Modelling and Simulation, ISBN: 978-0-9872143-1-7 . Perth, Australia, 12–16 December 2011. pp. 447–453.

(Dhillon et al 2007), Dhillon I.S., Guan Y., and B. Kulis. "Weighted Graph Cuts without Eigenvectors: A Multilevel Approach." IEEE Trans. Pattern Anal. Mach. Intell (2007): 29(11):1944–1957.

(Donoser & Bischof 2013), Donoser M., and Bischof H. "Diffusion processes for retrieval revisited." In: Proceedings of IEEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013. 1320-1327.

(Ducruet 2010), Ducruet C. "Les mesures locales d'un réseau." Author manuscript published in Web Science (2010): halshs – 00546814, version 1–14 December, groupe fmr.

(El Akkad & So Long 2009), El Akkad O., and So Long. GeoCities, The Globe and Mail, Available at: http://www.theglobeandmail.com/technology/globe-on-technology/so-longgeocities/article790790/. published Oct. 02 2009.

(Erétéo 2011), Erétéo G. "Semantic Social Network." Ph.D. thesis Défendu le 11 Avril 2011, Orange Labs, Telecom ParisTech, INRIA Sophia Antipolis – Méditerranée . 2011.

(Erétéo et al 2008), Erétéo G., Gandon F., Buffa M., Grohan P. "Analyse des réseaux sociaux et web sémantique." Un état de l'art. Technical report, dans la cadre du projet ISICIL (ANR) : Intégration Sémantique de l'Information par des Communautés de l'Intelligence en Ligne. 2008.

(Erétéo et al 2009), Erétéo G., Gandon F., Buffa M., and Corby O. "Semantic social network analysis." Proceedings of the WebSci'09: Society On-Line. Athens, Greece, 18-20 March 2009.

(Erétéo et al 2011), Erétéo G., Gandon F., and Buffa M. "SemTagP: Semantic Community Detection in Folksonomies." WI-IAT '11 Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. n.d. Volume 01. Pages 324-331, ISBN: 978-0-7695-4513-4.

(Everett & Borgatti 1999), Everett, M. G., & Borgatti, S. P. "The centrality of groups and classes." Journal of Mathematical Sociology (1999): 23(3): 181-201.

(Everett & Borgatti 2004), Everett, M.G., and Borgatti, S.P. " Extending centrality." In Carrington P.J., Scott J., and Wasserman S. (eds.), Models and methods in social network analysis. Cambridge: Cambridge University Press, 2004.

(Everton 2002), Everton S.F. A guide for the visually perplexed: visually representing social networks. Stanford: Stanford University: http://www.stanford.edu/group/esrg/siliconvalley/sivnap.html, 2002.

(Farganis 1993), Farganis J. "Readings in Social Theory: The Classic Tradition to Post-Modernism." McGraw-Hill, New York, 1993.

(Fellman et al 2011), Fellman P.V. & Parnell, Gregory S. & Carley, Kathleen M. "Biowar and Bioterrorism Risk Assessment." In Eighth International Conference on Complex Systems. Boston, 2011.

(Fire & Elovici 2013), Fire M., and Elovici Y. "Data Mining of Online Genealogy Datasets for Revealing Lifespan Patterns in Human Population." (2013): arXiv preprint arXiv:1311.427.

(Fire et al 2011), Fire M., Tenenboim-Chekina L., Puzis R., Lesser O., Rokach L., and Elovici Y. "Link Prediction in Social Networks using Computationally Efficient Topological Features." IEEE Third International Confernece on Social Computing (SocialCom) . 2011. 73-80.

(Fire et al 2012a), Fire M., Puzis R., and Elovici, Y. "Link Prediction in Highly Fractional Data Sets." Handbook of Computational Approaches to Counterterrorism. Springer, 2012.

(Fire et al 2012b), Fire M., Katz G., Elovici Y., Shapira B., and Rokach L. "Predicting student exam's scores by analyzing social network data." Active Media Technology, Springer Berlin Heidelberg (2012): 584-595.

(Fire et al 2013a), Fire M., Tenenboim-Chekina L., Puzis R., Lesser O., Rokach L., and Elovici Y. "Computationally efficient link prediction in a variety of social networks." ACM Transactions on Intelligent Systems and Technology (TIST) (2013): volume 5, number 1, pages 10.

(Fire et al 2013b), Fire M., Puzis R., and Elovici Y. "Organization Mining using Online Social Networks." (2013): arXiv preprint arXiv: 1303.3741.

(Fortunato et al 2004), Fortunato, S., Latora, V., Marchiori, M. "Method to find community structures based on information centrality." Phys. Rev (2004): E 70(5): 056104.

(Foulonneau 2010), Foulonneau M. "Modélisation, environnements, sémantiques et Web de données." Présentation dans le séminaire ISKO (juin 2010). Centre de Recherche Public Henri Tudor, Luxembourg. www.tudor.lu, 2010.

(Francis 2009), Francis D., Horine G., Tittel E.,. The Process of Building a Project Schedule. http://flylib.com/books/en/2.466.1.70/1/, 2009.

(Frankenstein et al 2015), Frankenstein, William & Mezzour, Ghita & Carley, Kathleen M & Carley, Richard L. "Remote assessment of countries nuclear, biological, and cyber capabilities: joint motivation and latent capability approach." Social Network Analysis and Mining, Springer (2015): 5(1).

(Frédéric 2009), Frédéric D. "Social Learning, Entreprise Collaborative – Ecollab – Une Introduction Au Social Learning." October 2009.

(Freeman 1977), Freeman, L.C. "A set of measures of centrality based on betweenness." Sociometry (1977): 40, p:35-41.

(Freeman 1979), Freeman, L.C. "Centrality in social networks: Conceptual Clarification." Social Networks (1979): 1, 215-239.

(Freeman 2004), Freeman L.C. "Graphical techniques for exploring social network data." In Carrington P.J., Scott, J., and Wasserman S. (eds.), Models and methods in social network analysis. Cambridge: Cambridge University Press, 2004.

(Gao et al 2012), Gao H.,Tang J., Liu H. "Exploring Social–Historical Ties on Location-Based Social Networks." in: Proc. 6th Int. AAAI Conf. on Weblogs and Social Media, 4–7 June, 2012. Dublin, Ireland, 2012. pp. 114–121.

(Geisberg et al 2008), Geisberg R., Sanders P., Scultes D. "Better approximation of Betweenness centrality." ALENEX08. 2008.

(Ghosh et al 2012), Ghosh S., Viswanath B., Kooti F., Sharma N. K., Gautam K., Benevenuto F., Ganguly N., Gummadi K. P. "Understanding and Combating Link Farming in the Twitter Social Network." World Wide Web Conference (WWW), April 2012. Lyon, France, 2012.

(Gilbert et al 2010), Gilbert F., Simonetto P., Zaidi F., Jourdan F., Bourqui R. "'Communities and hierarchical structures in dynamic social networks: analysis and visualization." Social Network Analysis and Mining (2010): Published online: 5 October 2010 on Springer-Verlag, (2011) 1:83–95, DOI 10.1007/s13278-010-0002-8.

(Gil-Mendieta & Schmidt 1996), Gil-Mendieta J., and Schmidt S. "The political network in Mexico." Social Networks (1996): 18 (2006) 4: 355-381.

(Gil-Mendieta et al 1997), Gil-Mendieta J., Schmidt S., Castro J., Ruiz A. "A Dynamic Analysis of the Mexican Power Network." Connections (1997): 20(2):34-55.

(Girvan & Newman 2002), Girvan M., Newman M.E.J. "Community structure in social and biological networks." Proceedings National Academy of Sciences of the USA, PNAS 99 (12). 2002. 7821–7826.

(Gjoka et al 2011), Gjoka M., Kurant M., Butts C., Markopoulou A. "Practical Recommendations on Crawling Online Social Networks." IEEE JSAC on Measurement of Internet Topologies (2011): Vol.29, No. 9, Oct. 2011.

(Gleiser & Danon 2003), Gleiser P. M., Danon L. "Community structure in jazz." Advances in Complex Systems (ACS) (2003): Vol: 6 Issue: 4 (December 2003), Page: 565 - 573.

(Gleiser 2007), Gleiser P. M. "How to become a superhero." Journal of Statistical Mechanics: Theory and Experiment (2007): Volume September 2007.

(Gloor & Zhao 2004), Gloor P., Zhao Y. "TeCFlow - A Temporal Communication Flow Visualizer for Social Networks Analysis." ACM CSCW Workshop on Social Networks. ACM CSCW Conference. Chicago, Nov. 6. 2004.

(Gloor 2007), Peter Gloor. Some Applications of Dynamic Social Network Analysis. Feb 28, 2007.

(Golbeck 2011), Jennifer Golbeck. "Temporal Social Networks." Joint institute for knowledge discovery. 2011.

(Goldbeck & Rothstein 2008), Goldbeck J., and Rothstein M. "Linking social Networks on the web with FOAF." Proceedings of the twenty-third conference on artificial intelligence, AAA08. 2008.

(Gregory 2007), Gregory S. "An algorithm to find overlapping community structure in networks." Lect. Notes Comput. Sci. (2007): 4702 91.

(Gregory 2009), Gregory S. Finding overlapping communities in networks by label propagation. http://arxiv.org/abs/0910.5516, 2009.

(Guimerà et al 2003), Guimerà R., Danon L., Díaz-Guilera A., Giralt F., and Arenas A. "Self-similar community structure in a network of human interactions." Physical Review E (2003): vol. 68, 065103(R).

(Gupte et al 2011), Gupte M., Shankar P., Li J., Muthukrishnan S., Iftode L. "Finding hierarchy in directed online social networks." In Proceedings of the 20th International Conference on WWW, Session: Social Network Analysis. 2011. pp. 557-566.

(Hagberg et al 2008), Hagberg A.A., Schult D., Swart P.J.,. "Exploring Network Structure, Dynamics, and Function Using NetworkX." in: Proc. 7th Python in Science Conference, SciPy2008. Pasadena, CA, USA, 2008. pp. 11–15.

(Halimi et al 2011), Halimi K., Seridi-Bouchelaghem H., and Faron-Zucker C. "SoLearn: a social learning network." Proceedings of Computational Aspects of Social Networks (CASoN 2011). University of Salamanca, Spain, 2011.

(Halimi et al 2014), Halimi, K., Seridi-Bouchelaghem H., and Faron-Zucker C. "An enhanced personal learning environment using social semantic web technologies." Interactive Learning Environments (2014): Vol. 22, No. 2, pp.165–187.

(Hamadache & Seridi-Bouchelaghem 2016), Hamadache B., and Seridi-Bouchelaghem H. "How to analyse a semantic social network of learners, in a social learning environment?" Int. J. Web Based Communities: Open Web Communities for Social Evolution (2016): Vol. 12, No. 3.

(Hamadache et al 2012), Hamadache B., Seridi-Bouchelaghem H., Farah N. "Analyse et recherche dans les réseaux sociaux." Les deuxièmes journées JDI 2012 - Session E-Technologies (18-19 Novembre). Guelma, Algérie, 2012. pp 33.

(Hamadache et al 2013a), Hamadache B., Seridi-Bouchelaghem H., and Farah N. "Toward Expressing a Preliminary Core Identity Significantly Characterized from the Social Network Temporal Dynamicity." Third International Conference, MEDI 2013, (September 25-27). Amantea, Calabria, Italy: Springer, 2013. pp: 149-161.

(Hamadache et al 2013b), Hamadache B., Seridi-Bouchelaghem H., Farah N. "Toward characterizing a more significant identity of core structure within dynamic social network." The 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013 , (August 25-28). Niagara Falls, Canada: IEEE/ACM, 2013. Pages 1458-1459.

(Hamadache et al 2015), Hamadache B., Seridi-Bouchelaghem H., and Farah N. "An elite grouping of individuals for expressing a core identity based on the temporal dynamicity or the semantic richness." Chapter in Social Network Analysis – Community Detection and Evolution (Springer) (2014–2015): pp.119–143.

(Hamadache et al 2016), Hamadache B., Seridi-Bouchelaghem H. & Farah N. "A significant core structure inside the social network evolutionary process." Soc. Netw. Anal. Min. (Springer) (2016): (2016) 6: 38. doi:10.1007/s13278-016-0344-y.

(Handcock et al 2008), Handcock M.S., Hunter D.R., Butts C.T., Goodreau S.M., Morris M. "Statistical Modeling of Social Networks with "statnet"." Journal of Statistical Software (2008): 24.(2008).

(Hansen et al 2010), Hansen D.L., Shneiderman B., Smith M.A.'. "Analyzing Social Media Networks with NodeXL." book, Morgan Kaufmann. Insights from a Connected World. 2010.

(Hasan et al 2006), Hasan M.A., Chaoji V., Salem S., Zaki M. "Link prediction using supervised learning." In: Proceedings of the SDM Workshop on Link Analysis, Counterterrorism and Security. 2006.

(Hathaway et al 2007), Hathaway T., Muse E.J., and Althoff T. Report on Pedagogical Practices and Methods in E-Learning. Bangor for the Engaging Diversity Development Partnership. School of Education, University of Wales, January 19, 2007.

(Hayes 2006), Hayes B. "Connecting the dots. can the tools of graph theory and social-network studies unravel the next big plot?" American Scientist (2006): 94(5):400-404.

(Hazar 2014), Hazar Hamad Hussain. "Time management tools and techniques for project management." Socio-economic Research Bulletin (2014): Issue 4 (55).

(Huberman & Adamic 1999), Huberman B.A., and Adamic L. A. "Growth dynamics of the World-Wide Web." Nature (1999): 399 (1999) 130.

(Huisman et al 2011), Huisman Mark, Marijtje A.J., van Duijn. "A reader's guide to SNA software." P.J., In Scott J. and Carrington. (Eds.) The SAGE Handbook of Social Network Analysis. London: SAGE., 2011. pp. 578-600.

(Il-Chul & Kathleen 2007), Il-Chul M. & Kathleen C.M. "Modeling and Simulation of Terrorist Networks in Social and Geospatial Dimensions." IEEE Intelligent Systems, Special issue on Special issue on Social Computing (2007): Sep/Oct '07, 22(5), 40 - 49.

(Il-Chul et al 2008), Il-Chul M & Carley, Kathleen M & Levis, Alexander H. "Vulnerability Assessment on Adversarial Organization: Unifying Command and Control Structure Analysis and Social Network Analysis." SIAM International Conference on Data Mining, Workshop on Link Analysis, Counterterrorism and Security, Apr. 26, 2008. Atlanta, Georgia, 2008.

(Isella et al 2011), Isella L., Stehlé J., Barrat A., Cattuto C., Pinton J-F., Van den Broeck W. "What's in a crowd? analysis of face-to-face behavioral networks." J. of Theoretical Biology (2011): 271(1):166-180.

(Jamali et al 2011), Jamali M., Haffari G., Ester M. "Modeling the temporal dynamics of social rating networks using bidirectional effects of social relations and rating patterns." International World Wide Web Conference Committee (IW3C2), WWW 2011 – Session: Temporal Dynamics. Hyderabad, India: ACM 978-1-4503-0632-4/11/03, March 28–April 1, 2011.

(Jin et al 2007), Jin Y., Matsuo Y., Ishizuka M. "Extracting a Social Network among Entities by Web mining." ESWC 2007. 2007.

(Johnson & Krempel 2004), Johnson J. C., and Krempel L. "Network Visualization: The "Bush Team" in Reuters News Ticker 9/11-11/15/01. ." The Journal of Social Structure's (2004): Vol. 5, No. 1. (2004). http://www.cmu.edu/joss/content/articles/volume5/JohnsonKrempel/.

(Johnson 1984-1985), Johnson D.S. "The Genealogy of Theoretical Computer Science." SIGACT News (1984-1985): Vol. 16, No. 2, pp. 36-44, 1984 , Reprinted in Bulletin of the EATCS, No. 25, pp. 198-211, 1985.

(Kamei et al 2012), Kamei T., Ono K., Kumano M., Kimura M. "Predicting Missing Links in Social Networks with Hierarchical Dirichlet Processes." WCCI 2012 IEEE World Congress on Computational Intelligence, June, 10–15, 2012, in: Proc. Int. Joint Conf. on Neural Networks, (IJCNN 2012)- . Brisbane, Australia, 2012. pp. 1816–1823.

(Kamvar 2003), Kamvar S.D., Haveliwala T. H. , Manning C. D., Golub G. H. Exploiting the Block Structure of the Web for Computing PageRank. Preprint (March, 2003), 2003.

(Kang et al 2007), Kang H., Getoor L., Singh L. "Visual analysis of dynamic group membership in temporal social network." ACM SIGKDD Explorations Newsletter - Special issue on visual analytics, Volume 9, Issue 2, ACM New York December 2007: Pages 13-21.

(Karolewski 2009), Karolewski IP. "Citizenship and collective identity in Europe." Edition, Rindle. Routledge advances in European politics. Routledge, 24 August 2009, p 260, 2009. pp 83–85.

(Kazienko et al 2011), Kazienko P., Kukla E., Musial K., Kajdanowicz K., Bródko P., Gaworecki J. "A generic model for multidimensional temporal social network." ICeND2011, The First International Conference on e-Technologies and Networks for Development, Communications in Computer and Information Science, CCIS 171. Dar-es-Salaam, Tanzania: Springer, 2011, August 3-5, 2011. pp. 1-14.

(Kelley 1961), Kelley J. "Critical Path Planning and Scheduling: Mathematical Basis." Operations Research (1961): Vol. 9, No. 3, p.296-320, May–June.

(Kempe et al 2003), Kempe D., Kleinberg J., Tardos E. "Maximizing the spread of influence through a social network." Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD ’03. 2003. pp. 137–146.

(Kent 1978), Kent D. "The rise of the Medici: Faction in Florence." Oxford: Oxford University Press (1978): 1426-1434.

(Kernighan & Lin 1970), Kernighan B., and Lin S. "An Efficient Heuristic Procedure for partitioning graph." The Bell System Technical J. (1970): 49.

(Kevin 2005), Kevin J. Lang. "Fixing two weaknesses of the Spectral Method." NIPS (2005): 715-722.

(Khan et al 2014), Khan N. ,Yaqoob I., Hashem I.A.T., Inayat Z., Ali W.K.M., Alam M. Shiraz M. Gani A. "Bigdata: survey, technologies, opportunities, and challenges." 2014 (The Sci. World J.): (2014)1–18.

(Kirschner 2008), Kirschner A. "Overview of Common Social Network Analysis Software Platforms." San Fransisco: Monitor Group. 2008.

(Kleinberg 2000), Kleinberg J.M. "The Small-World Phenomenon: An Algorithmic Perspective." Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing, STOC ’00. 2000. pp. 163–170.

(Klimmt & Yang 2004), Klimmt B., Yang Y. "Introducing the Enron corpus." CEAS conference. 2004.

(Knuth 1993), Knuth D. E. The Stanford GraphBase: A Platform for Combinatorial Computing. Addison-Wesley, Reading, MA, 1993.

(Kochut & Janik 2007), Kochut K. J., and Janik M. "SPARQLeR: Extended SPARQL for Semantic Association Discovery." Proc. European Semantic Web Conference, ESWC'2007. Innsbruck, Austria, 2007.

(Kolaczyk 2009), Kolaczyk E. Statistical analysis of network data. Springer, 2009.

(Kooij et al 2009), Kooij R., Jamakovic A., van Kesteren F., de Koning T., Theisler I., Veldhoven R. "The Dutch Soccer Team as a Social Network." International Network of Social Network Analysis - INSNA (2009): Volume 29, Issue 1.

(Kossinets & Watts 2006), Kossinets G., and Watts D. "Empirical analysis of an evolving social network." Science (2006): 311 (5757) (2006) 88–90.

(Krackhardt's 1999), Krackhardt's D. "The ties that torture: Simmelian tie analysis in organizations." Research in the Sociology of Organizations (1999): 16 (1999), 183-210.

(Krebs 2012), Krebs V. unpublished, http://www.orgnet.com/. n.d.

(Kunegis et al 2012), Kunegis J., Gröner G., Gottron T. "Online dating recommender systems: The split-complex number approach." In Proc. RecSys Workshop on Recommender Systems and the Social Web. 2012. pages 37-44.

(Kwak et al 2010), Kwak H., Lee C., Park H., and Moon S. "What is Twitter, a Social Network or a News Media?" Proceedings of the 19th International World Wide Web (WWW) Conference April 26-30, 2010. Raleigh NC (USA), 2010.

(Lafifi & Bensebaa 2008), Lafifi Y., and Bensebaa T. "Criteria for collaborators search." Proceedings of 3rd IEEE International Conference on Information & Communication: From Theory to Application, ICTTA'08, 7–11 April. Damascus, Syria: Digital Object Identifier: 10.1109/ICTTA.2008.4529916., 2008. ISBN: 978-1-4244-1751-3.

(Latapy et al 2008), Latapy M., Magnien C., Del Vecchio N. "Basic notions for the analysis of large two-mode networks." Social Networks (2008): 30(1), 31-48.

(Lathia et al 2008), Lathia N., Hailes S., Capra L. "kNN CF: A temporal social network." Recsys'08: Proceedings of the 2008 ACM Conference (October 23–25, 2008), on Recommender Systems, ASSOC Computing Machinery. Lausanne, Switzerland: ACM, 2008. pp. 227 - 234.

(Leblanc 2008), Leblanc A. "Environnement de collaboration et mémoire organisationnelle de formation dans un contexte d'apprentissage." Thèse présentée et soutenue le 3 Décembre 2008 pour l'obtention du Doctorat dans l'école doctorale (spécialité informatique) de l'Université de Technologie de Compiègne (UTS). 2008 .

(Lee & Adorna 2012), Lee J.B., and Adorna H.,. "Link prediction in a modified heterogeneous bibliographic network." In: Proceedings of International Conference on Advances in Social Networks Analysis and Mining (ASONAM). 2012. 442-449.

(Leskovec & Horvitz 2008), Leskovec J. and Horvitz E. "Planetary-scale views on a large instant messaging network." in Proceeding of the 17th World Wide Web Conference, WWW2008. Beijing, China, 2008.

(Leskovec et al 2007), Leskovec J., Kleinberg, J., and Faloutsos C. "Graph evolution: densification and shrinking diameters." ACM Transactions on Knowledge Discovery from Data (ACM TKDD) (2007): Vol. 1, No 1, Article 2, 41 pages.

(Leskovec et al 2009), Leskovec J., Lang K.J., Dasgupta A., and Mahoney M.W. "Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters." Internet Mathematics (2009): 6(1) 29--123.

(Leskovec et al 2010a), Leskovec J., Lang K. J, Mahoney M. W. "Empirical Comparison of Algorithms for Network Community Detection." In proceeding of the 19th International World Wide Web Conference, WWW2010. Madrid, Spain, 2010.

(Leskovec et al 2010b), Leskovec J., Huttenlocher D., Kleinberg J. "Signed networks in social media." in: Proceedings of CHI 2010, SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM, 2010. pp. 1361–1370.

(Leskovec et al 2010c), Leskovec J., Huttenlocher D., Kleinberg J. "Predicting Positive and Negative Links in Online Social Networks." WWW 2010. 2010.

(Liben-Nowell & Kleinberg 2007), Liben-Nowell D., and Kleinberg J. "The link-prediction problem for social networks." Journal of the American Society for Information Science and Technology (2007): 58(7), 1019-1031,(May 2007).

(Lichtenwalter et al 2010), Lichtenwalter, R.N., Lussier, J.T., Chawla, N.V. "New perspectives and methods in link prediction." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2010. 243-252.

(Limpens 2010), Limpens F. "Multi-points of view semantic enrichment of folksonomies." Ph.D. thesis. 2010.

(Lin et al 2008), Lin Y.R., Chi Y., Zhu S., Sundaram H., Tseng B.L. "Facetnet: a framework for analyzing communities and their evolutions in dynamic networks." In WWW '08: Proceeding of the 17th international conference on World Wide Web. New York, NY, USA: ACM, 2008. pages 685–694.

(Loiacono 2011), Daniele Loiacono. Graph Mining and Social Network Analysis. Data Mining and Text Mining (UIC 583 @ Politecnico di Milano). Milan, 2011.

(Loscalzo & Yu 2008), Loscalzo S., Yu L. "Social network analysis: Tasks and tools." In Liu H., Salerno J.J., and Young M.J. (Eds.): Social Computing, Behavioral Modeling, and Prediction. New York: Springer, 2008. pp. 151-159.

(Lusseau 2006), Lusseau D. "Evidence for social role in a dolphin social network." Springer Science+Business Media B.V. (2006).

(Lusseau et al 2003), Lusseau D., Schneider K, Boisseau O.J., Haase P., Slooten E. and Dawson S. M. "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations." Behavioral Ecology and Sociobiology (2003): 54, 396-405.

(Madan et al 2012), Madan A., Cebrian M., Moturu S., Farrahi K., Pentland A. "Sensing the 'Health State' of a Community." Pervasive Computing (2012): Vol. 11, No. 4, pp. 36-45 Oct 2012.

(Magdon-Ismail & Purnell 2011), Magdon-Ismail M., Purnell J. "Fast overlapping clustering of networks using sampled spectral distance embedding and GMMs." Tech. rep., Rensselaer Polytechnic Institute. 2011.

(Magnusson 2012), Jonathan Magnusson. "Social Network Analysis Utilizing Big Data Technology ." Januray 2012, UPPSALA UNIVERSITET. 2012.

(Maheswaran et al 2010), Maheswaran M., Ali B., Ozguven H., Lord J. "Online Identities and Social Networking." Handbook of Social Network Technologies. 2010. 241-267.

(Maiya & Berger-Wolf 2010), Maiya A. and Berger-Wolf T. "Sampling Community Structure." Proceedings of WWW 2010. Raleigh, NC, April 2010.

(Malek 2009), Malek Maria. "Introduction à l'analyse du réseaux sociaux." Rapport technique – LARIS – EISTI. octobre-novembre 2009.

(Malliaros et al 2016), Malliaros F. D., Apostolos N., Papadopoulos, Michalis Vazirgiannis. "Core Decomposition in Graphs: Concepts, Algorithms and Applications." Published in Proc. 19th International Conference on Extending Database Technology (EDBT), March 15-18, 2016 . Bordeaux, France, 2016.

(Martínez Arqué & Nettleton 2012), Martínez Arqué N., Nettleton D.F. "Analysis of on-line social networks represented as graphs–extraction of an approximation of community structure using sampling." Proc. Congress Modeling Decisions for Artificial Intelligence, MDAI, in: LNAI. Springer-Verlag, 2012. vol. 7647, pp. 149–160.

(Massa et al 2008), Massa P., Souren K., Salvetti M., Tomasoni D. "Trustlet, Open Research on Trust Metrics (link)." Scalable Computing: Practice and Experience, Scientific International Journal for Parallel and Distributed Computing (2008).

(Massa et al 2009), Massa P., Salvetti M., Tomasoni D. "Bowling alone and trust decline in social network sites." In Proc. Int. Conf. Dependable, Autonomic and Secure Computing. 2009. pages 658-663.

(Mastrandrea et al 2015), Mastrandrea R., Fournet J., Barrat A. "Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys." PLoS One. (2015): 10(9): e0136497, Published online 2015 Sep 1,arXiv:1506.03645, http://arxiv.org/abs/1506.03645.

(Max & Kathleen 2005), Max T. & Kathleen C.M. "Structural Knowledge and Success of Anti-Terrorist Activity: The Downside of Structural Equivalence ." Journal of Social Structure (2005): 6(2), elec. pub.

(McAuley & Leskovec 2012), McAuley J., Leskovec J. "Learning to discover social circles in ego networks." In Advances in Neural Information Processing Systems (2012): pages 548-556.

(McGlohon & Faloutsos 2008), McGlohon M., Faloutsos C. "Graph mining techniques for social media analysis." International Conference on Weblogs and Social Media (ICWSM). Seattle, 2008.

(Meeder et al 2011), Meeder B., Karrer B., Sayedi A., Ravi R., Borgs C., Chayes J. "We know who you followed last summer: inferring social link creation times in twitter." International World Wide Web Conference Committee (IW3C2), WWW 2011 – Session: Temporal Dynamics. Hyderabad, India: ACM 978-1-4503-0632-4/11/03, March 28–April 1, 2011.

(Memon & Alhajj 2010), Memon N., Alhajj R. "Introduction to the first issue of social network analysis and mining journal." SOCNET (2011) (November 13, 2010): Springer-Verlag Wien, New York 2010, 1:1–2, DOI 10.1007/s13278-010-0016-2.

(Memon & Alhajj 2011), Memon N., Alhajj R. "Introduction to the second issue of social network analysis and mining journal: scientific computing for social network analysis and dynamicity." Social Network Analysis and Mining (2011): Published online: 29 March 2011 on Springer-Verlag 2011, 1:73–74, DOI 10.1007/s13278-011-0022-z.

(Michalski et al 2011), Michalski R., Palus S., Kazienko P.,. "Matching organizational structure and social network extracted from email communication." In Lecture Notes in Business Information Processing. Springer Berlin Heidelberg, 2011. volume 87, pages 197-206.

(Mika 2005a), Mika P. "Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks." Web Semantics: Science, Services and Agents on the World Wide Web (2005): Vol. 3, No. 2-3., pp. 211-223.

(Mika 2005b), Mika P. "Ontologies are Us: a Unified Model of Social Networks and Semantics." In ISWC. Springer, 2005. volume 3729 of LNCS, p. 522–536.

(Miller 1986), Miller A.G. "The obedience experiments: A case study of controversy in social science." New York, Westport, Preaeger, 1986.

(Min-Joong et al 2016), Min-Joong L., Sunghee C., Chin-Wan C. "Efficient algorithms for updating betweenness centrality in fully dynamic graphs." Information sicences Journal (ELsevier) (2016): Volume 326, 1 January 2016, Pages 278-296.

(Mislove 2009), Mislove A. Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems. PhD thesis. Rice University, 2009.

(Mislove et al 2007), Mislove A., Marcon M., Gummadi K.P., Druschel P., Bhattacharjee B. "Measurement and Analysis of Online Social Networks." Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement (IMC '07). San Diego, CA, October 2007. pp. 29–42.

(Mislove et al 2008), Mislove A., Koppula H. S., Gummadi K. P., Druschel P., Bhattacharjee B. "Growth of the Flickr social network." Workshop on Online Social Networks (WOSN), August 2008. Seattle, 2008.

(Missaoui 2014), Missaoui R., Sarr I. Preface, ebook : Social Network Analysis - Community Detection and Evolution. Lecture Notes in Social Networks, Springer International Publishing Switzerland 2014, 2014.

(Mollenhorst et al 2014), Mollenhorst G., Beate V., Henk F. "Changes in personal relationships: how social contexts affect the emergence and discontinuation of relationships." Soc .Netw (2014): 37, 65–80.

(Nettleton 2013), Nettleton, D.F. "Data mining of social networks represented as graphs." Computer Science Review (February 2013): Vol. 7, pp.1–34.

(Newman & Girvan 2004), Newman, M.E.J., and Girvan M. "Finding and evaluating community structure in networks." Physical Review E (2004): Vol. 69(2), p.026113.

(Newman 2001a), Newman M. E.J. "Clustering and preferential attachment in growing networks." Physical Review E (2001): 64(2) 025102.

(Newman 2001b), Newman M. E. J. "The structure of scientific collaboration networks." PNAS (2001): 98, 404-409.

(Newman 2001c), Newman M. E. J. "Scientific collaboration networks: I. Network construction and fundamental results." Phys. Rev. E (2001): 64, 016131.

(Newman 2001d), Newman M. E. J. "Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality." Phys. Rev. E (2001): 64, 016132.

(Newman 2003), Newman M.E.J. "The structure and function of complex networks." SIAM Review 45 (2003): 167–256.

(Newman 2004), Newman M. E. J. "Fast algorithm for detecting community in networks." Phys. Rev E (2004): 69, 066133.

(Newman 2006), Newman M. E. J. "Finding community structure in networks using the eigenvectors of matrices." Preprint physics/0605087 (2006).

(Newman 2012), Newman M. E. J. "Communities, modules and large-scale structure in networks." Nature Physics (2012): 8, 25-31.

(NGUYEN et al 2013), NGUYEN H. G., LUONG N. L., PHAM T. T. M., TRAN T. D. Label Propagation Algorithm. Final Report First Year Project Master of Software Engineering 2012-2014. University Of Bordeaux, 1- June 2013.

(Noble & Weiss 2004), Noble L., and Weiss S. Following the Enron money trail. Box 4.3, Chapter 4: Corporate money, Global corruption report 2004, https://issuu.com/transparencyinternational/docs/2004_gcr_politicalcorruption_en, Page 74. Center for Responsive Politics, United States, 2004.

(Olguin et al 2009), Olguin D.O., Waber B.N., Taemie Kim, Mohan A., Ara K., Pentland A. "Sensible Organizations: Technology and Methodology for Automatically Measuring Organizational

Behavior." IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics (2009): Volume: 39 , Issue: 1, Page(s): 43- 55 , Feb. 2009.

(Oprah 2009), Oprah Tries Twitter, Crowns Ashton King of It. 2009. <http://blogs.wsj.com/digits/2009/04/17/oprah-tries-twitter-crowns-ashton-king-of-it/>.

(Opsahl & Panzarasa 2009), Opsahl T., and Panzarasa P. "Clustering in weighted networks." Social Networks (2009): 31 (2), 155-163, doi: 10.1016/j.socnet.2009.02.002.

(Opsahl 2013), Opsahl T. "Triadic closure in two-mode networks: Redefining the global and local clustering coefficients." Social Networks (2013): 35 (2),159-167.

(Opsahl et al 2008), Opsahl T., Colizza V., Panzarasa P., Ramasco J. J. "Prominence and control: The weighted rich-club effect." Physical Review Letters 101 (168702) (2008): arXiv:0804.0417.

(Padgett 1994), Padgett J.F. Marriage and Elite Structure in Renaissance Florence. 1994.

(Panzarasa & Opsahl 2009), Panzarasa P., Opsahl T., Carley M. K. "Patterns and dynamics of users' behavior and interaction: Network analysis of an online community." Journal of the American Society for Information Science and Technology (2009): 60 (5), 911-932, doi: 10.1002/asi.21015.

(Paolillo & Wright 2006), Paolillo J. C., and Wright E. "Social Network Analysis on the Semantic Web: Techniques and Challenges for Visualizing FOAF." Information, in Book Visualizing the semantic WebXmlbased Internet And. 2006.

(Parthasarathy et al 2011), Parthasarathy S., Ruan Y., Satuluri V. "Community discovery in social networks: Applications, Methods and emerging trends." book chapter in Social Network Data Analytics. 2011. 79-113.

(Plangprasopchok et al 2010), Plangprasopchok A., Lerman K., and Getoor L. "Growing a Tree in the Forest: Constructing Folksonomies by Integrating Structured Metadata." In Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD), July 2010. 2010.

(Raghavan et al 2007), Raghavan R.N., Albert R., Kumara S. "Near Linear Time Algorithm to Detect Community Structures in Large Scale Network." Phys. Rev. E (2007): 76, 036106.

(Reda et al 2009), Reda K., Tantipathananandh, C., Berger-Wolf, T., Leigh, J., Johnson, A. E. "SocioScape – a tool for interactive exploration of spatio-temporal group dynamics in social networks." Proceedings of the IEEE Information Visualization Conference (INFOVIS '09). Atlantic City, New Jersey, 10/11/2009 -10/16/2009.

(Richardson et al 2004), Richardson M., Agrawal R., Domingos P. "Trust Management for the Semantic Web." in: Proc. 2nd Int. Semantic Web Conference, ISWC. 2003. pp. 351–368.

(Robins et al 1999), Robins, G., Pattison P., Wasserman S. "Logit models and logistic regressions for social networks:III. Valued relations." Psychometrika (1999): 64: 371-394.

(Robins et al 2005), Robins G.,Pattison P., Woolcock J. "Small and other worlds: global network structures from local processes." American Journal of Sociology (AJS) (2005): 110 (4): 894–936.

(Rosen et al 2011), Rosen D., Kim J. H., Barnett G. A. "Social networks and online environments: when science and practice co-evolve." Soc Netw Anal Min, on Springerlink.com (2011): 1:27–42, DOI 10.1007/s13278-010-0011-7.

—. "Social networks and online environments: when science and practice co-evolve." Published online on Springerlink.com, SOCNET (2011): 1:27–42, DOI 10.1007/s13278-010-0011-7.

(Rowe et al 2014), Rowe M., and Strohmaier M. "The semantic evolution of online communities." WWW (Companion Volume) (2014): 433-438.

(Rozenblat 2010), Rozenblat C. "Opening the black box of agglomeration economies for measuring cities' competitiveness through international firm networks." Urban Studies (2010): 47(13):2841–2865.

(Said et al 2010), Said A., De Luca E. W., Albayrak S. "How social relationships affect user similarities." In Proc. IUI Workshop on Social Recommender Systems. 2010.

(Santoro et al 2011), Santoro N., Quattrociocchi W., Flocchini P., Casteigts A., Amblard F. "Time-varying graphs and social network analysis: Temporal indicators and metrics." 3rd AISB Social Networks and Multiagent Systems Symposium (SNAMAS). 2011. 32-38.

(Satuluri et al 2010), Satuluri V., Parthasarathy S., and Ucar D. "Markov Clustering of Protein Interaction Networks with Improved Balance and Scalability." In Proceedings of the ACM Conference on Bioinformatics and Computational Biology. 2010.

(Scellato et al 2010), Scellato S., Mascolo C., Musolesi M., Latora V. "Distance Matters: Geo-social Metrics for Online Social Networks." in: Proc. 3rd Workshop on Online Social Networks, WOSN 2010, . Boston, MA, USA: http://www.usenix.org/events/wosn10/tech/full_papers/Scellato.pdf, n.d.

(Scott & Hughes 1980), Scott J., Hughes M. "The anatomy of Scottish capital: Scottish companies and Scottish capital, 1900-1979." London: Croom Helm (1980).

(Scott 2000), Scott J. Social network analysis, a handbook. Deuxième edition. Edition Sage, 2000.

(Scott 2010), Scott J.,. "Social network analysis: developments, advances, and prospects." SOCNET (2011) (Published online: 6 October 2010 on Springer-Verlag 2010): 1:21–26, DOI 10.1007/s13278-010-0012-6.

(Scott 2011), Scott J. "Social network analysis: developments, advances, and prospects." Social Network Analysis and Mining, Springer-Verlag (2011): 1(1):21–26, DOI 10.1007/s13278-010-0012-6.

—. "Social network analysis: developments, advances, and prospects." Published online on Springer-Verlag SOCNET (2011): 1:21–26, DOI 10.1007/s13278-010-0012-6.

(Seierstad & Opsahl 2010), Seierstad C., and Opsahl, T. "For the few not the many? The effects of affirmative action on presence, prominence, and social capital of women directors in Norway." Scandinavian Journal of Management (2010): 27 (1), 44-54, doi: 10.1016/j.scaman.2010.10.002.

(Seierstad & Opsahl), Seierstad C., and Opsahl, T. "For the few not the many? The effects of affirmative action on presence, prominence, and social capital of women directors in Norway." Scandinavian Journal of Management (2010): 27 (1), 44-54, doi: 10.1016/j.scaman.2010.10.002.

(Seshadri et al 2008), Seshadri M., Machiraju S., Sridharan A., Bolot J., Faloutsos C., Leskovec J. "Mobile call graphs: beyond power-law and lognormal distributions." in: Proc. 14th ACM SIGKDD, KDD'08. New York, NY, USA, 2008. pp. 596–604.

(Shafie 2010), Shafie T. "Design-based Estimators for Snowball Sampling." Workshop on Survey Sampling Theory and Methodology, August 23–27, 2010. Vilnius, Lithuania, 2010.

(Shannon et al 2003), Shannon P., Markiel A., Ozier O., Baliga N., Wang J., Ramage D., Amin N., Schwikowski B., and Ideker T. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome research (2003): 13(11):2498–2504.

(Shi et al 2013), Shi J., Xue W., Wang W., Zhang, Yang B., Li J. "Scalable community detection in massive social networks using MapReduce." IBM Journal of Research and Development (2013): Volume 57 Issue 3-4, paper12, May/July 2013.

(Simmel 1903), Simmel G. " Die Grosstädte und das Geistesleben (The Metropolis and Mental Life)." Petermann, Dresden, 1903.

(Smalla 2013), Smalla M.L., '. "Weak ties and the core discussion network: why people regularly discuss important matters with unimportant alters." Soc. Netw. (2013): 235,470–483.

(Smalla et al 2015), Smalla L. M., Deeds Pamphileb V., McMahan P. "How stable is the core discussion network?" ELSEVIER - Social Networks (2015): Volume 40, January 2015, Pages 90–102.

(Snijders 2005), Snijders Tom A.B. Introduction To Dunamic SocialL Network Analysis. University of Groningen, The Netherlands, November 2005, 2005.

(Snijders 2010), Snijders T.A.B. "Conditional marginalization for exponential random graph models." Journal of Mathematical Sociology (2010): 34 (2010) 239–252.

(Socievole et al 2014), Socievole A., De Rango F., Caputo A. "Wireless contacts, Facebook friendships and interests: Analysis of a multi-layer social network in an academic environment." In Wireless Days (WD), (November 2014), IFIP . 2014. pp. 1-7.

(Srivastava & DeLong 2013), Srivastava J., DeLong C. "Social & Behavioral Analytics Mining Behaviors of a Connected World." The 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013 , (August 25-28). . Niagara Falls, Canada: IEEE/ACM, 2013.

(Steglich et al 2010), Steglich C.E.G., Snijders T.A.B., and Pearson M. "Dynamic Networks and Behavior: Separating Selection from Influence." Sociological Methodology (2010): 40, 329-393.

(Stephen 2012), Stephen M. "For Future Reference, a Pioneer in Online Reading." The Wall Street Journal (January 12, 2012).

—. "For Future Reference, a Pioneer in Online Reading." The Wall Street Journal (January 12, 2012).

(Sudeshna & Birinder 2009), Sudeshna S., Birinder S. T. Analyzing the socio-cognitive of an economic research community. Submitted in partial fulfillment of the requirements of the degree of " Bachelor of technology (Honours)" in computer science and engineering. Department of science and engineering in Indian institute of technology Kharagpur(Master). Kharagpur, India, May 2009.

(Sun et al 2007), Sun J., Faloutsos C., Papadimitriou S., Yu P.S. "Graphscope: parameter-free mining of large time-evolving graphs." In KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM, 2007. pages 687–696.

(Sun et al 2009), Sun Y., Yu Y., and Han J. "Ranking-based clustering of heterogeneous information networks with star network schema." In KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009. pages 797–806.

(Sun et al 2011), Sun Y., Barber R., Gupta M., Aggarwal C.C., Han J. "Co-author relationship prediction in heterogeneous bibliographic networks." In: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM). 2011. 121-128.

(Sundaresan et al 2007), Sundaresan S. R., Fischhoff I. R., Dushoff J., Rubenstein D. I. "Network metrics reveal diVerences in social organization between two Wssion-fusion species, Grevy's zebra and onager." Oecologia (2007): 151:140-149.

(Tab. univ. & biv. 2013), Tableaux (Statistique) univariés et bivariés. n.d. 11 Juin 2016 <http://www.unesco.org/webworld/ 1027 portal/idams/html/french/F2tables.htm>.

(Takac & Zabovsky 2012), Takac L., Zabovsky M. "Data analysis in public social networks." Int. Scientific Conf. and Int. Workshop Present Day Trends of Innovations. 2012.

(Takes 2011), Takes F. "Social Network Analysis." Data Mining. LIACS, Universiteit Leiden, 21 November 2011.

(Tang & Liu 2009a), Tang L., Liu H. "Relational Learning via Latent Social Dimensions." In Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'09). 2009. Pages 817–826.

(Tang & Liu 2009b), Tang L., Liu H. "Scalable Learning of Collective Behavior based on Sparse Social Dimensions." In Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09). ACM, 2009.

(Tang et al 2010a), Tang J., Musolesi M., Mascolo C. Latora V. "Characterising Temporal distance and reachability in mobile and online social networks." ACM SIGCOMM Computer Communication Review (Jan. 2010): vol. 40, no. 1, p. 118.

(Tang et al 2010b), Tang J., Musolesi M., Mascolo C., Latora V., Nicosia V. "Analysing information flows and key mediators through temporal centrality metrics." Proceedings of the 3rd Workshop on Social Network Systems (SNS'10). Paris, France: ACM (2010), April 13, 2010.

(Tantipathananandh et al 2007), Tantipathananandh C., Berger-Wolf T., Kempe D. "A framework for community identification in dynamic social networks." Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD'07. New York: ACM 2007, August 12–15, 2007. pp. 717-726.

(Tauveron 2012), Matthias Tauveron. "De la cooccurrence généralisée à la variation du sens lexical, in La cooccurrence, du fait statistique au fait textuel." (Damon Mayaffre et Jean-Marie Viprey, eds). 2012. CORPUS, 11, 2012.

(Taylor et al 2016), Taylor D., Myers Sean A., Clauset A., Porter Mason A., Mucha Peter J. "Eigenvector-Based Centrality Measures for Temporal Networks." Multiscale Modeling and Simulation: A SIAM Interdisciplinary Journal (2016): physics.soc-ph cs.SI nlin.AO physics.data-an arXiv:1507.01266v2.

(Tommasini & Daolio 2010), Tommasini M., Daolio F. "Structure & Dynamique des Réseaux Socio-économiques." Série 04 - Centralité, Modularité, Communauté. Université de Lausanne, Faculté des hautes études commerciales, 2010.

(Tommasini & Daolio), Tommasini M., Daolio F. "Structure & Dynamique des Réseaux Socio-économiques." Série 04 - Centralité, Modularité, Communauté. n.d.

(Torniai et al 2008), Torniai C., Jovanovic J., Gasevic D., Bateman S., and Hatala M. "E-learning meets the social semantic web." Proceedings of the 2008 Eighth IEEE International Conference on Advanced Learning Technologies. University of Washington, DC: IEEE Computer Society, 2008. pp.389–393.

(Traub et al 2010), Traub MC., Lamers MH., Walter W. "A semantic centrality measure for finding the most trustworthy account." In: Proceedings of the IADIS international conference informatics, July 2010. Freiburg, Germany, 2010. pp 117–125.

(Van Eck & Waltman 2010), Van Eck N.J., Waltman L. "Software survey: VOSviewer, a computer program for bibliometric mapping." Scientometrics (2010): 84(2), 523–538.

(Van Eck & Waltman 2011), Van Eck N.J., and Waltman L. "Text mining and visualization using VOSviewer." ISSI Newsletter (2011): 7(3), 50–54.

(Varghese & Allen 1993), Varghese G., Allen T. "Relational Data in Organizational Settings: An Introductory Note for Using AGNI and Netgraphs to Analyze Nodes, Relationships, Partitions and Boundaries." Connections (1993): Volume XVI, Number 1 & 2, Spring.

(Vickers & Chan 1981), Vickers M., and Chan S. Representing Classroom Social Structure. Melbourne: Victoria Institute of Secondary Education, 1981.

(Vigfusson 2010), Vigfusson Y. "Affinity in distributed systems." Thèse présentée à la faculté de l'école supérieure de l'Université Cornell pour l'obtention le diplôme de docteur en philosophie, février 2010. 2010.

(Viswanath et al 2009), Viswanath B., Mislove A., Cha M., Gummadi K. P.,. "On the Evolution of User Interaction in Facebook." ACM SIGCOMM Workshop on Social Networks (WOSN), August 2009. Barcelona, Spain: ACM, 2009.

(von Frentz 2003), Clemens von Frentz. "ENRON – Chronique d'une faillite record." manager magazin (25/09/2003).

(Von Luxburg 2007), Von Luxburg U. "A tutorial on spectral clustering." Statistics and Computing (2007): 17(4):395–416.

(VOSviewer 2013), Nees Jan van Eck, Ludo Waltman. VOSviewer Manual . (Version 1.5.3 - 5 Dec 2013), VOSviewer page: http://www.vosviewer.com/, 2013.

(Wang et al 2013), Wang L., Hopcroft J., He J., Liang H., Suwajanakorn S. "Extraction the core Structure of Social Network using alpha Beta community." Internet Mathematics, Taylor & Francis Groups (January 1, 2013): volume 9- Issue 1–Pages 58-81.

(Wasserman & Faust 1994), Wasserman S., Faust K. "Social Network Analysis in the Social and Behavioral Sciences ." Social Network Analysis: Methods and Applications, Cambridge University Press (1994): pp. 1–27.

(Wasserman & Pattison 1996), Wasserman S., Pattison P. "Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*." Psychometrika (1996): 61: 401-425.

(Wattenhofer et al 2012), Wattenhofer M., Wattenhofer R., Zhu Z. "The YouTube Social Network." in Proc. 6th Int. AAAI Conf. on Weblogs and Social Media, 4–7 June. Dublin, Ireland, 2012. pp. 354–361.

(Watts & Strogatz 1998), Watts D.J, and Strogatz S. H. "Collective dynamics of `small-world networks'." Nature (1998): 393, 440-442.

(Watts et al 2002), Watts D. J., Dodds P. S., and Newman M. E. J. "Identity and Search in Social Networks." Science (2002): 296, 1302-1305.

(Wellman 2001), Wellman B. "Computer Networks As Social Networks." Science 293 (2001): 2031-34 (2001).

(Xiang 2008), Xiang E.W. "A survey on link prediction models for social network data." Science and Technology (2008).

(Xie & Szymanski 2011), Xie J., and Szymanski B. K. "Community detection using a neighborhood strength driven label propagation algorithm." In IEEE NSW 2011. 2011. pages 188-195.

(Xie & Szymanski 2012), Xie J., and Szymanski B. K. "Towards linear time overlapping community detection in social networks." In PAKDD . 2012. pages 25-36.

(Xie & Szymanski 2013), Xie J., and Szymanski B. K. "LabelRank: A Stabilized Label Propagation Algorithm for Community Detection in Networks." IEEE Network Science Workshop, April 29-May 01, 2013 . West Point, NY, 2013. pp. 138-143.

(Xie et al 2011), Xie J., Szymanski B. K., and Liu X. "SLPA: Uncovering Overlapping Communities in Social Networks via A Speaker-listener Interaction Dynamic Process." In Proc. of ICDM 2011 Workshop. 2011.

(Xie et al 2013), Xie J.,Kelley S., Szymanski B.K. "Overlapping community detection in networks: the state-of-the-art and comparative study." ACM Comput. Surv.(CSUR) (2013): 45(4)(2013)43.

(Xu et al 2010), Xu K., Tang C., Ali G., Li C., Tang R., Zhu J. "A comparative study of six software packages for complex network research." Paper presented at the 2010 International Conference on Communication Software and Networks. Singapore , 2010.

(Yang & Leskovec 2011), Yang J., Leskovec J. "Temporal variation in online media." in: Proc. WSDM '11, 4th ACM Int. Conf. on Web Search and Data Mining. New York, NY, USA: ACM, 2011. pp. 177–186.

(Yessad 2009), Yessad A. Construction d'un Environnement Pédagogique Adaptatif basé sur les Modèles et Techniques du Web Sémantique. Thèse présentée pour l'obtention du diplôme de doctorat pour la saison 2008/2009. Faculté des Sciences de l'ingénieur. Département informatique de l'université de Badji Mokhtar d'Annaba (Algérie), 2009.

(Yongcheng Xu et al 2013), Yongcheng Xu, Chen L., and Zou S. "Ant Colony Optimization for Detecting Communities from Bipartite network ." Journal of Software (2013): Volume 8, Number 11 DOI: 10.4304/jsw.8.11.2930-2935.

(Zachary 1977), Zachary W. W. "An Information Flow Model for Conflict and Fission in Small Groups." Journal of Anthropological Research (1977): Vol. 33, No. 4, pp. 452-473.

(Zafarani & Liu 2009), Zafarani R., and Liu H. "Social Computing Data Repository at ASU." [http://socialcomputing.asu.edu]. Tempe, AZ: Arizona State University, School of Computing, Informatics and Decision Systems Engineering. 2009.

(Zhai et al 2013), Zhai L., Xiangbin Y., Guojing Z.,. "A centrality measure for communication ability in weighted network." Physica A: Statistical Mechanics and its Applications (2013): Volume 392, Issue 23, 1 December 2013, Pages 6107-6117.

(Zheleva et al 2009), Zheleva E., ShararaH., Getoor L. "Co-evolution of social and affiliation netowrks." KDD 2009 (n.d.).

(Zhou & Lipowsky 2004), Zhou H., Lipowsky R. "Network brownian motion: a new method to measure vertex-vertex proximity and to identify communities and sub communities." in: Computational Science-ICCS2004. Springer, 2004. pp.1062–1069.

(Zhou et al 2006), Zhou D., Manavoglu E., Li J.,Giles C.L., and H. Zha. "Probabilistic models for discovering e-communities ." In WWW '06: Proceedings of the 15th international conference on WorldWideWeb. ACM, 2006. page 182.

(Zhou et al 2007), Zhou D., Councill I., Zha H., Lee Giles C. "Discovering temporal communities from social network documents." IEEE International Conference on Data Mining (ICDM 2007). 2007. 745-750.

(Zhou et al 2009), Zhou T., Lu L., Zhang Y.C. "'Predicting missing links via local information." The European Physical Journal B - Condensed Matter and Complex Systems (2009): 71(4), 623-630, (October 2009).

(Zhou et al 2011), Zhou L., Ding L., Finin T. "How is the Semantic Web evolving? A dynamic social network perspective." ELSEVIER - Computers in Human Behavior (2011): Volume 27 Issue 4, July, 2011, Pages 1294-1302.

Sources Web

(1) http://groupefmr.hypotheses.org/3309
(2) http://lab41.github.io/blog/2014/08/22/exploring-the-congo/
(3) http://www.fredcavazza.net/2014/05/22/social-media-landscape-2014/
(4) http://fr.wikipedia.org/wiki/%C3%89cosyst%C3%A8me
(5) http://www.futura-sciences.com/magazines/environnement/infos/dico/d/environnement-ecosysteme-135/
(6) http://www.friendster.com
(7) http://www.orkut.com
(8) http://www.renren.com/
(9) http://d.weibo.com/
(10) https://www.tuenti.com/
(11) http://www.hi5.com/
(12) http://www.studivz.net/
(13) http://www.skyrock.com/
(14) https://myspace.com/
(15) https://www.linkedin.com/
(16) http://www.facebook.com
(17) http://www.twitter.com
(18) http://en.wikipedia.org/wiki/OpenSocial
(19) http://webtrends.about.com/od/glossary/g/tag_def.htm
(20) http://en.wikipedia.org/wiki/Thomas_Vander_Wal
(21) http://www.bibsonomy.org/
(22) Twitter development APIs. Disponible sur: https://dev.twitter.com/.
(23) LinkedIn developer APIs. Disponible sur : http://developer.linkedin.com/apis.
(24) Stanford Network Analysis Platform Datasets. Disponible sur : http://snap.stanford.edu/data/index.html.
(25) Network datasets collectées utilisées par Newman. Disponible sur : http://www-personal.umich.edu/~mejn/netdata/
(26) http://www.orgnet.com/

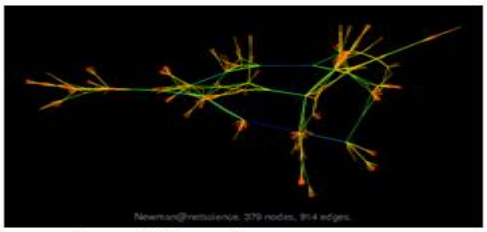
(27) Datasets standard fournis par le software UCINET : http://vlado.fmf.uni-lj.si/pub/networks/data/UciNet/UciData.htm
(28) Liste de datasets de Eric D. Kolaczyk disponibles sur : http://math.bu.edu/people/kolaczyk/datasets.html
(29) Dataset collecté par Tore Opsahl, disponible sur : http://toreopsahl.com/datasets/#online_social_network
(30) Datasets disponibles sur la page personnelle de Cheng-jun Wang http://weblab.com.cityu.edu.hk/blog/chengjun/resources/publicly-accessible-datasets-of-human-online-behavior/
(31) http://www.tagora-project.eu/
(32) http://www.isi.edu/integration/people/lerman/downloads.html
(33) http://digg.com/
(34) http://odysseas.calit2.uci.edu/doku.php/public:online_social_networks
(35) http://www.cise.ufl.edu/research/sparse/matrices/groups.html
(36) http://deim.urv.cat/~alexandre.arenas/data/welcome.htm
(37) http://proj.ise.bgu.ac.il/sns/datasets.html
(38) http://www.anybeat.com/
(39) https://www.academia.edu/
(40) https://plus.google.com/
(41) http://cafe.themarker.com/
(42) http://www.wikitree.com/
(43) http://www.cytoscape.org/
(44) http://yifanhu.net/
(45) http://yifanhu.net/GALLERY/GRAPHS/index.html
(46) http://socialcomputing.asu.edu/pages/datasets
(47) http://www.blogcatalog.com
(48) http://www.buzznet.com/
(49) https://delicious.com/
(50) www.flixster.com/
(51) https://fr.foursquare.com/
(52) www.livemocha.com
(53) http://www.youtube.com
(54) http://socialnetworks.mpi-sws.org/datasets.html
(55) http://konect.uni-koblenz.de/
(56) Orkut network dataset - KONECT, May 2015: http://konect.uni-koblenz.de/networks/orkut-groupmemberships
(57) Teams network dataset - KONECT, May 2015. http://konect.uni-koblenz.de/networks/dbpedia-team
(58) Manufacturing emails network dataset - KONECT, May 2015. http://konect.uni-koblenz.de/networks/radoslaw_email
(59) Uc irvine messages network dataset - KONECT, May 2015. http://konect.uni-koblenz.de/networks/opsahl-ucsocial
(60) Facebook wall posts network dataset - KONECT, May 2015. http://konect.uni-koblenz.de/networks/facebook-wosn-wall
(61) Bibsonomy user–tag network dataset - KONECT, May 2015. http://konect.uni-koblenz.de/networks/bibsonomy-2ut
(62) Twitter user–hashtag network dataset - KONECT, May 2015. http://konect.uni-koblenz.de/networks/munmun_twitterex_ut
(63) Train bombing network dataset - KONECT, May 2015. http://konect.uni-koblenz.de/networks/moreno_train
(64) Infectious network dataset - KONECT, May 2015. http://konect.uni-koblenz.de/networks/sociopatterns-infectious

(65) Seventh graders network dataset - KONECT, May 2015. http://konect.uni-koblenz.de/networks/moreno_seventh
(66) Last.fm song network dataset - KONECT, May 2015. http://konect.uni-koblenz.de/networks/lastfm_song
(67) Filmtipset network dataset - KONECT, May 2015. http://konect.uni-koblenz.de/networks/filmtipset_comment
(68) Facebook (nips) network dataset - KONECT, May 2015. http://konect.uni-koblenz.de/networks/ego-facebook
(69) http://blog.gowalla.com/
(70) Gowalla network dataset - KONECT, May 2015. http://konect.uni-koblenz.de/networks/loc-gowalla_edges
(71) http://www.advogato.org/
(72) Advogato network dataset - KONECT, May 2015. http://konect.uni-koblenz.de/networks/advogato
(73) Brightkite network dataset - KONECT, May 2015. http://konect.uni-koblenz.de/networks/loc-brightkite_edges
(74) Twitter (mpi) network dataset - KONECT, May 2015. http://konect.uni-koblenz.de/networks/twitter_mpi
(75) http://moreno.ss.uci.edu/data.html
(76) http://netsg.cs.sfu.ca/youtubedata/
(77) http://webscope.sandbox.yahoo.com/index.php
(78) http://vlado.fmf.uni-lj.si/pub/networks/data/ucinet/ucidata.htm
(79) The Structure of Information Networks: Instructor: Jon Kleinberg http://www.cs.cornell.edu/courses/cs685/2002fa/
(80) La page de Sepandar D. Kamvar. http://web.stanford.edu/~sdkamvar/research.html
(81) http://www.casos.cs.cmu.edu/computational_tools/data2.php
(82) http://www.insna.org/
(83) http://law.di.unimi.it/datasets.php
(84) http://crawdad.org/
(85) http://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public
(86) http://kevinchai.net/
(87) http://www.trustlet.org/wiki/Advogato_dataset
(88) http://databib.org/databib.php
(89) http://www.sociopatterns.org/datasets/
(90) http://web.stanford.edu/group/sonia/dataSources/index.html
(91) http://realitycommons.media.mit.edu/index.html
(92) NetMiner 4, software tool for exploratory analysis and visualization of network data. Available at http://www.netminer.com.
(93) 'Neo4J' Graph Database System. Available at http://neo4j.org/
(94) https://networkx.github.io/
(95) JUNG, Java Universal Network/Graph Framework. Available at http://jung.sourceforge.net/
(96) 'igraph' library and API. Available at http://igraph.sourceforge.net/
(97) http://www.gmw.rug.nl/~huisman/sna/software.html
(98) http://www.netvis.org/
(99) http://nwb.cns.iu.edu/index.html
(100) http://mrvar.fdv.uni-lj.si/pajek/
(101) http://www.fmsasg.com/Products/SentinelVisualizer/
(102) http://socnetv.sourceforge.net/
(103) http://www.libsna.org/
(104) http://www.yworks.com/
(105) http://www.cfinder.org/
(106) Commetrix (Version 2.4-2012) http://www.commetrix.net/
(107) http://www.analytictech.com/keyplayer/keyplayer.htm

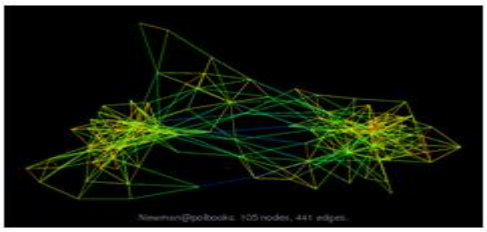
(108) http://www.gmw.rug.nl/~stocnet/StOCNET.htm
(109) http://statnet.csde.washington.edu/index.shtml
(110) https://www.s2.onasurveys.com/
(111) http://www.sfu.ca/personal/archives/richards/Pages/negopy4.html
(112) http://www.absint.com/aisee/index.htm
(113) http://people.apache.org/~stefano/agora/
(114) https://gephi.github.io/
(115) http://www.andrew.cmu.edu/user/krack/krackplot.shtml
(116) https://sites.google.com/site/netdrawsoftware/download
(117) http://www.casos.cs.cmu.edu/computational_tools/datasets/external/gleiser/index11.php

Annexe A: Données sociales

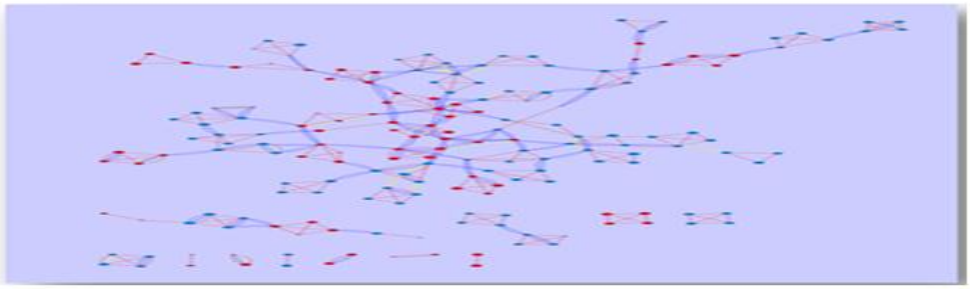
Les 4 figures suivantes montrent des exemples de SNs de la première catégorie cités dans le (Tableau 17), dont certains sont visualisés par Yifan Hu (AT&T Labs Visualization Group) [44, 45]et des outils comme Cytoscape[43] ((Shannon et al 2003)), et autres, etc.



Réseau de collaborations entre co-auteurs scientifiques travaillant sur la théorie et expérimentations dans les réseaux [25, 35] [111]



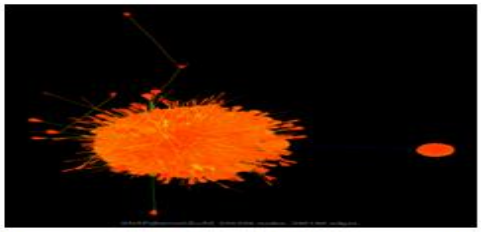
Books about US politics de V. Krebs[26, 25, 35] [135]



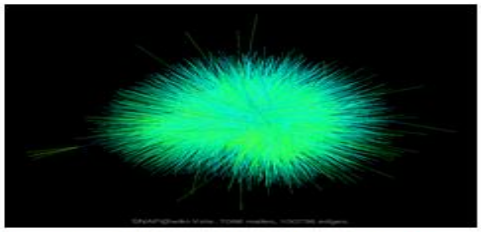
Students' Cooperation Social Network (Multi-Graph) [37] [141]

Figure 135. Exemple de SNs collaboration ou autres qui sont inférés ((Newman 2006)) ((Krebs 2012)) ((Fire et al 2012b))

La figure suivante montre des SNs de la deuxième catégorie (Tableau 17).

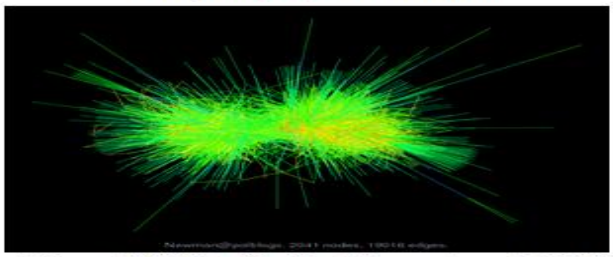


Réseau [35, 45] de 'Email – EuAll' d'une grande institution de recherche européenne depuis Octobre 2003 au mai 2005 [122]



Réseau [35, 45] de 'wiki-Vote ' [128, 129] Données de 2 794 élections en Wikipédia depuis la création de Wikipédia jusqu'à Janvier 2008,

Réseau de 'Academia Dataset' [37, 39] [139, 140]



Réseau [25, 35, 45] de 'Political blogosphere Feb 2005' des hyperliens entre les weblogs des politiciens d'US en 2005 [112]

Figure 136. Exemple de SNs extraits depuis les data logs des applications qui ne sont pas strictement d'OSN ((Adamic & Glance 2005)) ((Leskovec et al 2007)) ((Leskovec et al 2010b)) ((Leskovec et al 2010c)) ((Fire et al 2011)) ((Fire et al 2013a))

La figure suivante montre des exemples de SNs de la troisième catégorie (Tableau 17).

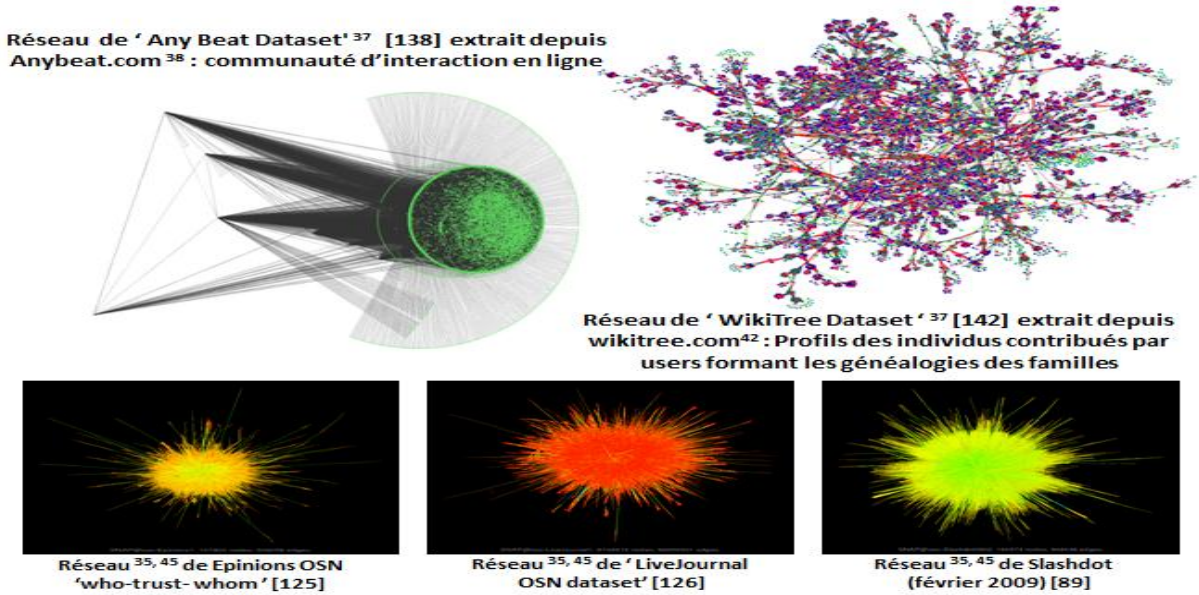


Figure 137. Exemples de SNs 'crawled' depuis les applications de OSNs ((Leskovec et al 2009)) ((Richardson et al 2004)) ((Backstrom et al 2006)) ((Fire et al 2012a)) ((Fire & Elovici 2013))

Annexe B: Dynamique des réseaux sociaux

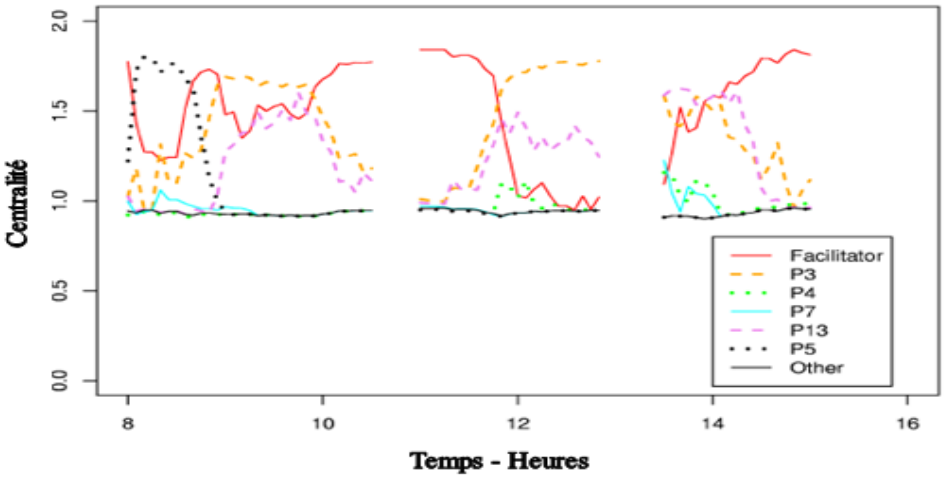


Figure 138. Evolution des scores de centralités individuelles dans un groupe de discussion ((Dekker 2011))

Cette figure décrit l'étude de (Dekker 2011) sur l'évolution des scores de centralités (proximité) qui ont été normalisés par rapport la centralité moyenne. Les phases d'interactions se détachent, ce qui rend la résolution des fenêtres de temps plus facile. Cela évite aussi le risque de perdre le fil de certains changements intermédiaires qui peuvent passés inaperçus. La (**Figure 138**) montre que l'animateur (Facilitateur) de l'atelier domine brièvement la communication au départ: Une sorte de mise en scène (Dekker 2011). La période qui vient par la suite est dominée par le participant P5, avant que l'animateur reprenne de nouveau le contrôle. Après, la discussion est dominée à la fois par P3 et P13 et ainsi de suite dans les phases suivantes.

Annexe C : Sémantique des réseaux sociaux

1. Représentations sémantiquement plus riches

On part du dernier exemple cité dans l'analyse des SNs dynamiques, là où (Kazienko et al 2011) ont représenté les relations multiples d'un SN qui évolue avec différents types de relations par des couches sémantiques. Chacune représente un type de relation : amitié, affaire, famille, etc. (Kazienko et al 2011). C'est une approche de représentation de la sémantique des relations qui s'intéressait précisément au système de partage de photos Flickr en distinguant 11 types de relations entre utilisateurs (Kazienko et al 2011) et qui a mené à un SN multicouches, ou encore un multi –graphe (Kazienko et al 2011). Mais le typage de relation ne résume pas seule toute sa sémantique, mais c'est un fragment.

1.1. Sémantique générale des interactions sociales médiatisées par ordinateur

Avant de chercher dans le typage des connexions sur les OSNs/SNs et dans leurs causalités, les auteurs proposent 3 catégories de relations extraites du contexte général de ces réseaux (systèmes informatiques, médias sociaux, etc.): Relation directe, pseudo directe et indirecte (Kazienko et al 2011).

- ***Relation directe*** entre deux utilisateurs différents qui sont conscients d'être en relation (Kazienko et al 2011). à chaque fois l'un des deux est émetteur et l'autre un récepteur qui savent qu'un processus d'échange d'informations a lieu, par exemple dans la communication par email, messagerie instantané, etc., (**Figure 139**).

- ***Relation pseudo-directe*** qui relie deux différentes utilisateurs x et y à travers un objet intermédiaire qui est soumis aux activités de 'x' et 'y' (**Figure 139**), par exemple le partage, la coédition, etc. Les deux utilisateurs ne sont pas forcément impliqués directement dans la création et le maintien de cette relation. D'ailleurs l'un des deux peut être inconscient de l'existence de l'autre (Kazienko et al 2011). 'x' et 'y' pratiquent des activités parfois identiques ou différentes, ce qui donne 2 formes des relations pseudo-directes (Kazienko et al 2011). La première est une relation pseudo-directe avec les mêmes rôles. C'est à dire x et y appliquent la même action sur l'objet, par exemple commenter une photo (Kazienko et al 2011). La deuxième est une relation pseudo-directe avec des rôles différents, tel que l'un des utilisateurs agit sur l'objet différemment de l'autre (Kazienko et al 2011).

- ***Relation indirecte*** se déduit suite entre 2 utilisateurs ayant des profils similaires ou identiques (Kazienko et al 2011) (par exemple à partir des réseaux d'affiliation ou une sorte de filtrage démographique). Aucun des utilisateurs n'est conscient de l'existence de cette relation (**Figure 139**).

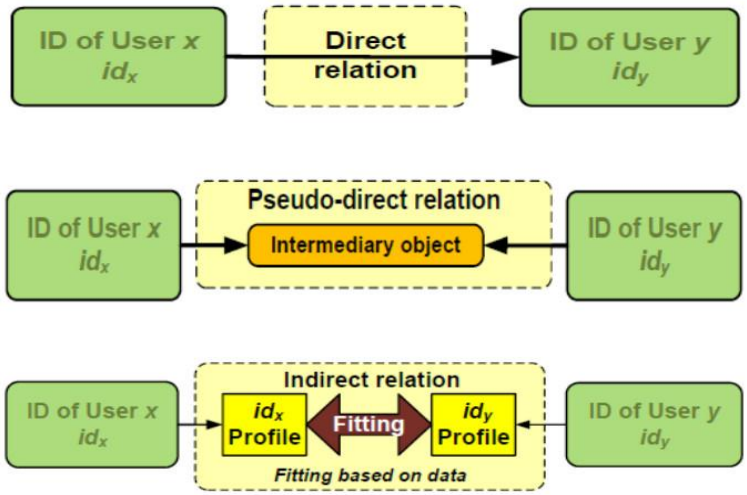


Figure 139. Catégories des liens sociaux, une sémantique générale de relation selon un niveau abstraction basique ((Kazienko et al 2011))

Ces aspects sémantiques sont basiques et reflètent un niveau d'abstraction directement lié au contexte général des OSNs

1.2.Web sémantique, une tendance pour enrichir les modèles de SNs

Les techniques, métriques et approches en SNA évaluent et étudient systématiquement les caractéristiques des SNs, le potentiel de ses acteurs, sa structuration en communautés/ groupes et leur dynamique temporels, etc., sur des graphes sociaux structurels, topologiques. Souvent, il n y a pas de sémantique traitée. Donc, des représentations sémantiquement plus riches, plus expressives seront exigées en premier plan pour enrichir le SNA et donner plus de signification à ses interprétations. Dans ce sens, le web sémantique s'est imposé comme une solution efficace selon une communauté de chercheurs (Erétéo et al 2009) (Erétéo et al 2008) ((Erétéo et al 2011)).

1.2.1. Ontologie et Web sémantique

L'expression de web sémantique est attribuée à Tim Berners-Lee au sein de W3C ('World Wide Web Consortium') ((Yessad 2009)), une version étendue du web actuel, là où le contenu est quasiment inaccessible/ inexploitables par les traitements machines. Aujourd'hui, le volume de données structurées par les technologies du web sémantique est en croissance rapide ((Takes 2011)) en représentant intelligemment les connaissances. C'est un ensemble de normes, de langages et protocoles utilisés pour décrire, structurer et interroger des ressources, des activités, et même les profils sociaux des utilisateurs et leur interactions sur le web, sous formes de données structurées et normalisées (des métadonnées) lisibles par la machine. Sachant que les recherches récentes dans ce sens s'appuient énormément sur les résultats de l'ingénierie de connaissances particulièrement, ces métadonnées sont des annotations basées sur des vocabulaires qui sont en effet des ontologies.

Les ontologies sont avantageuses pour le web sémantique pour définir des vocabulaires. L'ontologie semble être une composante logicielle qui apporte une dimension sémantique dans les systèmes informatique. C'est une spécification formelle explicite d'une conceptualisation selon Gruber, un ensemble de concepts, de propriétés, d'axiomes, de fonctions et de contraintes explicitement définis dans un langage formel de représentation de connaissance. Le W3C a standardisé plusieurs langages dont certains les plus connus sont cités brièvement ci-dessous.

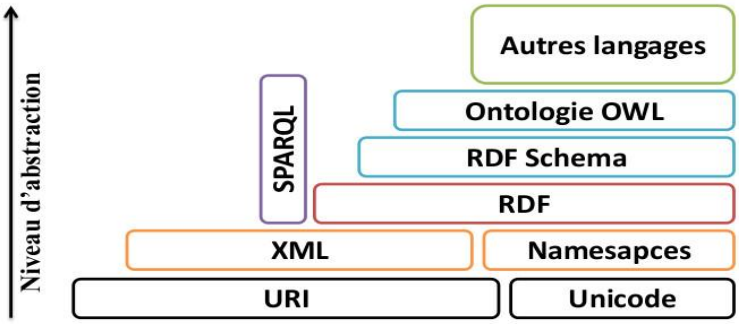


Figure 140. Langages de web sémantique en couches

Ces langages (**Figure 140**) sont avec des formalismes dotés d'une syntaxe (couche de transport syntaxique) XML (langage interopérable) et constituent une infrastructure du web sémantique. Mise à part la couche sommet (Autres langages), les autres couches reflètent aujourd'hui les langages relativement stabilisés, RDF, RDFS, OWL pour les ontologies, et présentent des caractéristiques principales comme:

- La capacité d'identification de diverses ressources en se basant sur notion d'URI ('Uniform Resource Identifier') : Un identifiant unique est attribué à une ressource (profil social, compte utilisateur, etc.) sur le Web ((Champin & Mrissa 2011)).
- L'utilisation des triplets pour représenter ces métadonnées sous la forme (sujet, prédicat, objet). Le sujet représente une ressource. Le prédicat est un type de propriété applicable sur la ressource. L'objet c'est la valeur de la propriété, soit un littéral (une donnée) ou une autre ressource.

RDF ('Resource Description Framework') : C'est le langage relationnel de base de web sémantique, utilisé pour générer des annotations (triplets) RDF. Ces triplets forment un modèle de graphe RDF. Un vocabulaire sera nécessaire pour générer un graphe RDF car ce langage en termes de conceptualisation ne permet pas d'exprimer les propriétés et valeurs autorisées pour un type de ressources, etc.
RDFS ('RDF Schema') : C'est le langage utilisé pour définir un vocabulaire, des primitives ontologiques (ontologie légère) selon un degré d'expressivité donné. Les ressources sont définies par des classes 'rdfs : Class' (une hiérarchie de classes 'rdfs: subClassOf'). Avec RDFS on définit aussi une propriété RDF d'une classe, son domaine de valeurs 'rdfs: domain', et son applicabilité 'rdfs : range'.
OWL ('Ontology Web Language') : Dans certains contextes, on demande plus d'expressivité pour décrire plus de richesse sémantique (par exemple spécifier plus de contraintes sur des propriétés, prendre qu'une seule valeur, etc.). Le langage OWL (une révision du langage de création d'ontologies DAML + OIL) est une extension de RDFS permettant d'enrichir les schémas RDFS, définir des vocabulaires plus expressives, des ontologies plus complexes. Il y a plus de contraintes sur les propriétés (symétrie, transitivité, cardinalité, etc.). Selon le degré d'expressivité souhaitée. Il existe trois versions d'OWL : OWL Lite, OWL DL, et OWL Full.

SPARQL ('Protocol And RDF Query Language'): C'est un protocole et un langage de requête dédié à interroger les annotations RDF. les requêtes s'expriment par une syntaxe adaptée, basée également sur des triplets. Les requêtes interrogatives (clause 'SELECT') permettent d'extraire des sous-graphes RDF correspondant à un ensemble de ressources, en vérifiant éventuellement des conditions dans la clause 'Where'. Il y a aussi des Requêtes constructives (clause 'CONSTRUCT') permettant d'instancier un nouveau graphe RDF.

1.2.2. Graphes sociaux sémantiques RDF

En se basant sur ces langages de web sémantique, certains chercheurs sont convaincus du fait que le graphe RDF peut présenter un modèle plus riche d'un graphe social. Des modèles ontologiques sont proposés pour représenter, annoter et exploiter la sémantique des profils utilisateurs, leurs interactions sociales, leurs usages (Erétéo et al 2008). Il s'agit d'un pont entre le web 2 et le web sémantique, en améliorant la qualité de représentation des SNs.

1.2.2.1.Modèle FOAF ('Friend Of A Friend')

FOAF est la plus célèbre des ontologies, car elle est utilisée pour modéliser les profiles et leurs relations sur les plateformes des OSNs ayant une forte audience. Elle propose des propriétés, pour la description d'un profil ('foaf : name', etc.), et des classes pour les usages d'autres ressources ('foaf : OnlineAccount', 'Document', etc.) avec des propriétés correspondantes ('foaf : holdsOnlineAccount', etc.), (**Figure 141**). La propriété 'foaf: knows' est la plus importante puisqu'elle permet de connecter les profils et former ainsi un SN sémantique de profils FOAF annotés en RDF (SN d'accointances). C'est le cas des profils FOAF sur livejournal et flickr.

1.2.2.2.Modèles plus étendues

Il y a des extensions (ontologies plus expressives) de la modélisation FOAF.

RELATIONSHIP est une ontologie qui propose des propriétés (des relations) plus spécialisées que foaf: knows, en distinguant les relations familiales, amicales ou professionnelles.

SIOC ('Semantically-Interlinked Online Communities Project') est une ontologie qui propose une description plus détaillée et spécialisée de 'foaf : holdsOnlineAccount', etc., concernant les activités en lignes sur des applications sociales (Weblog (**Figure 141**)). En outre elle permet la gestion des propriétés de documents (basée sur l'ontologie du Dublin Core, un standard de métadonnées décrivant les ressources numérique).

SKOS ('Simple Knowledge Organization System') : Elle est spécialisée dans la gestion et l'articulation des tags décrivant les ressources manipulées (la propriété 'IsSubjectOf' (**Figure 141**)). Cette ontologie (qui peut être en RDFS) donne la possibilité de définir des labels associés à un tag (prefLabel, altLabel, etc.) et les relations sémantiques ((Foulonneau 2010)) entre tags (Skos : narrower, Skos : broader, Skos : related), (**Figure 141**).

Au sein de ce trio ontologique FOAF (RELATIONSHIP), SIOC, SKOS (**Figure 141**), qui permettent de générer des annotations de profiles (entité sociales), de ressources et de tags, une ontologie appelée ***SCOT*** 'Social Semantic Cloud Of Tags' s'impose comme un moyen pour représenter la sémantique de relations entre les annotations de ces données de 'social tagging' (folksonomies). 'hasTag', hasLink, 'exists' sont les propriétés les plus connus de SCOT pour décrire ces relations. Un profile FOAF utilise un tag appartenant à un nuage de tags pour décrire une ressource qu'il manipule, annotée par les concepts de SIOC. Ce nuage de tags se compose de ('hasTag') de plusieurs tags. Un tag annoté par les concepts de SKOS, décrit 'exists'. Cette ressource peut être décrite ' hasLink' par plusieurs tags dans le nuage de tags. Donc, les données (de folksonomies) structurées par SCOT peuvent être aussi une source pour des SNs sémantiques.

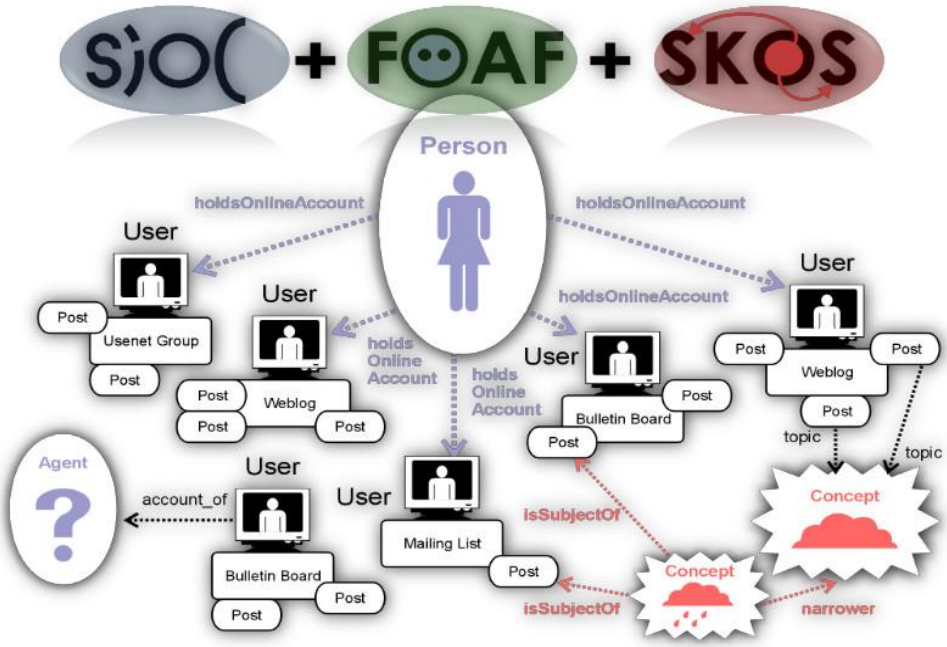


Figure 141. Trio ontologique FOAF, SIOC, SKOS permettant d'annoter les données de social tagging : utilisateur, ressource, tag ((Erétéo et al 2008))

2. Analyse des SNs sémantiques

Les SNs sémantiques affichent-ils des propriétés inhérentes d'un SN ? C'est la première question à poser. Des chercheurs ont facilement démontré que ces réseaux possèdent normalement des caractéristiques des SNs. L'analyse des SNs sémantiques porte un bénéfice informationnel menant à enrichir le SNA. Mais l'analyse de ces nouvelles traces, comment peut-on exploiter la sémantique exprimée constituent un vrai défi. La complexité des techniques de SNA (calcule d'intermédiarité, détection de communautés, etc.) pose déjà une problématique même dans un cadre statique. Elle est susceptible de de se multiplier en analysant des graphes sociaux RDF. D'autre part, les langages et les outils de web sémantiques ne disposent pas d'opérateurs qui répondent aux exigences de SNA.

2.1.Premiers exemples

Paolillo & Wright 2006 ((Paolillo & Wright 2006)) ont essayé d'analyser un graphe social RDF, formé par des profils FOAF, extrait de LiveJournal. Il se compose de deux types de relations (propriétés) 'knows' et 'interest'. Mais ils l'ont réduit à 2 graphes non-typés, un réseau d'accointance formé que par les propriétés « knows » et le réseau d'intérêts, formé par les propriétés de type 'interest'. Ils ont étudié par exemple l'organisation en communautés de ces deux réseaux indépendamment. Pour cela, ils ont appliqué un clustering hiérarchique. Ensuite, les deux réseaux ont été fusionnés en un graphe bipartite afin de déterminer selon eux les centres d'intérêts de chaque groupe d'utilisateurs ((Paolillo & Wright 2006)).
((Goldbeck & Rothstein 2008)) sont aussi parmi les premiers qui ont étudié les profils FOAF en introduisant des propriétés supplémentaire relatives à la 'confiance' accordée entre personnes.
Au lieu d'appliquer une analyse routinière, ces auteurs ont tenté de résoudre à partir de ces traces sémantiques l'un des problèmes des OSNs, la multiplicité des profils d'une même personne physique ((Paolillo & Wright 2006)). Devant la prolifération des profils et donc des profils FOAF qui est décentralisée dans différents OSNs ce problème de multiplicité se pose en utilisant le web sémantique. Mais ((Goldbeck & Rothstein 2008)) ont transformé le problème en atout pour le web sémantique en proposant une méthode de fusion de profils

FOAF. Elle se base sur certaines propriétés de profils FOAF qui sont de nature associées à une personne unique (Courriel, identifiant de messagerie, page web personnelle, etc.). Si deux profils partagent une valeur identique dans l'une de ces propriétés, la méthode détecte qu'il s'agit de la même personne, et donc elle fusionne les deux profils. Par conséquent, les personnes qui ont des profils FOAF fusionnés sur plusieurs sites de réseautage social sont susceptibles d'occuper des positions de hubs (vue précédemment) entre ces plateformes sociales. Donc ils n'ont pas seulement une forte centralité de degré, mais la fusion augmente leur centralité d'intermédiarité.

2.2.Analyser par des extensions de web sémantique

Accessibles via des APIs appropriés, les représentations classiques des OSNs sont améliorées par des modélisations sémantiques (comprenant les structurations sémantiques de 'social tagging') qui ne sont pas évidentes (**Figure 142**). Par ailleurs, la sémantisation de son analyse en exploitant sa richesse sémantique constitue un autre défi. Certains d'autres chercheurs ((Erétéo 2011)) n'ont pas hésité à fusionner les deux modèles : les technologies de web sémantique et la théorie des graphes classique pour enrichir SNA. Ils ont ainsi proposé un cadre qui vise une analyse sémantique de SN (**Figure 142**).

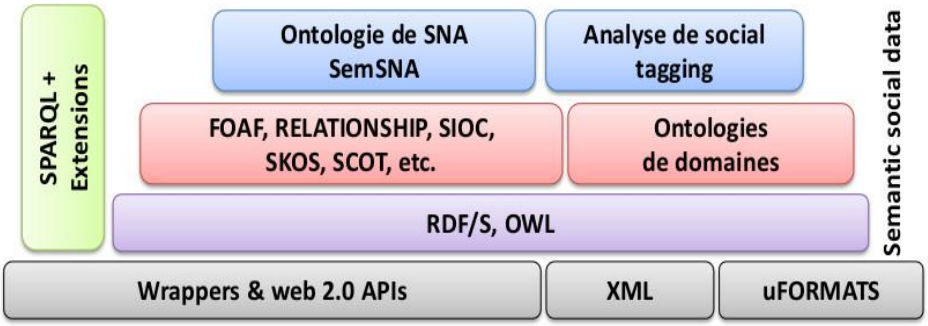


Figure 142. Un cadre de modélisation et d'analyse des SNs sémantiques, extrait de ((Erétéo 2011))

Il fallait d'abord envisager des extensions ou adapter des langages de web sémantique comme SPARQL ((Erétéo 2011)) pour répondre aux exigences de SNA appliquée sur un modèle de graphe social RDF (**Figure 142**). Il ne faut pas oublier que la complexité des algorithmes de calcule de beaucoup de métriques (centralité de proximité, d'intermédiarité, etc.) s'appuient sur la notion de chemin (les géodésiques) qui n'est pas supportée par SPARQL. Sa version standard ne traite le nœud (classe RDF) que dans son voisinage. Elle n'est pas assez expressive pour écrire des requêtes plus 'globales'. Donc, ((Anyanwu et al 2007)) ((Kochut & Janik 2007)), ((Corby 2008)) proposent des extensions (**Figure 142**) de SPARQL afin de pouvoir extraire des chemins entre des ressources sémantiquement liées. SPARQ2L, SPARQLeR ((Kochut & Janik 2007)) sont des exemples de ces extensions qui offrent la possibilité d'appliquer de plus en plus des contraintes sur :

- La longueur des chemins, la présence d'une ressource sur le chemin.
- L'orientation des chemins étant donné que le graphe est orienté.
- L'utilisation des expressions régulières pour filtrer la séquence et type des propriétés (relations) qui forment les chemins.
- Prise en compte de l'Homomorphisme des ressources

Cette extension de SPARQL a été intégrée dans un moteur sémantique appelé CORESE ((Erétéo 2011)). Mais, ce dernier devrait recevoir des modifications pour l'adapter à la résolution des nouvelles requêtes d'analyse là où on manipule des chemins dans des graphes RDF. Ici on fait appel à plusieurs notions pour exécuter l'opération de résolution. C'est une association/ projection (un mapping, E-mapping, ER-mapping, Homomorphisme, etc.) des

entités, des relations et les type de relations d'une requête à ceux du graphe RDF étant un ER-Graph : 'Entity Relation graph'. Ces notions ont été étendues vers PER-graph, PER-mapping qui constituent le principe d'extraction de chemin composé de propriétés (type de relation) qui a une longueur définie entre deux ressources donnée. D'où, les auteurs ((Erétéo 2011)) (Erétéo et al 2008), ont proposé des requêtes pour calculer et paramétrer (enrichir) des indices de SNA (Algorithme 14) en s'appuyant sur des agrégations et des conventions syntaxiques, par exemple une variable chemin est préfixé par '$', utiliser les expréssions régulières, etc. Dans Algorithme 14, on trouve par exemple le code d'une requête qui calcule une centralité de degré (n-degree) d'un nœud dans un graphe social RDF, paramétrée par le type de relation et la longueur de chemins : $Cd_{<type,\ length>}(y)$.

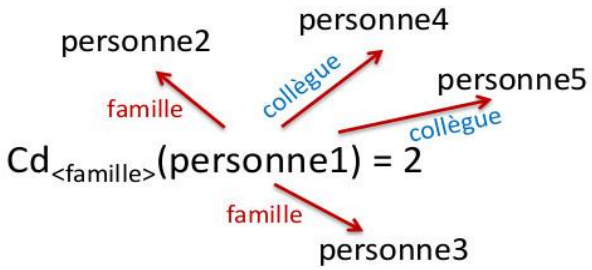


Figure 143. Degré paramétré.

Algorithme 14. Exemple de requête SPARQL pour calculer une centralité de degré paramétrée

```
select ?y count (?x) as ?degree where {
{?x  $path ?y
– filtrer par type de relation (propriétés ) en utilisant les expréssions régulières
filter (match($path, star(param[type])))
– filtrer par longueur de chemin
filter (match(pathLength <= param[length]))
}
UNION
– Refaire le calcule pour les chemins sortants
{?y  $path ?x
filter (match($path, star(param[type])))
filter (match(pathLength <= param[length]))
}
group by ?y
}
```

Dans ce sens, le calcule d'une centralité d'intermédiarité paramétrée $Cb_{<type>}(y)$ est beaucoup plus complexe et ne peut pas se faire directement avec une requête SPARQL. Évidemment, on doit passer par une requête qui calcule l'intermédiarité paramétrée de 'y' entre chaque pair de nœud (j, k), notée $bjk_{<Type>}(y)$ et qui implique à son tour l'extraction des géodésiques paramétrées notée $gjk_{<Type>}(y)$. Ensuite, un post-traitement (une itération) sur les résultats est nécessaire pour calculer les centralités finales.

En s'appuyant toujours sur le web sémantique, les auteurs ((Erétéo 2011)) ont pensé à exploiter les résultats d'une telle analyse pour enrichir encore les données sociales sémantiques, en proposant un modèle ontologique 'SemSNA' (**Figure 142**). Si les graphes RDF basés sur différentes modélisation ontologiques (ontologies de domaines) présentent les entités sociales, leurs usages (social tagging) sémantiquement liées, 'SemSNA' a pour avantage de gérer le cycle de vie de SN ((Erétéo 2011)). Les concepts/primitives de SemSNA (**Figure 144**) proposent d'annoter le SN en injectant les résultats d'analyse de sa représentation RDF. Ils se décomposent en catégories :

- Concepts noyau: Ils décrivent le contexte d'analyse: Le type de relation, le graphe (sous graphe) nommé par 'analyzedGraph' ((Erétéo 2011)). 'SNAConcept' est la superclasse de tous les concepts qui décrivent les indices de SNA (SNAindice, Path).
- Concepts pour décrire les chemins : On trouve la sous classe Path avec ses propriétés ('hasPathLength', 'pathExtremity') et ses sous classes ('DirectedPath', 'Geodesicpath', etc.), **(Figure 144).**
- Concepts positions stratégiques (métriques) qui se mettent sous la classe 'SNAindice', par exemple la sous classe 'Centrality' et ses sous classes : 'Closeness', etc., **(Figure 144).**
- Concepts décrivant les structures communautaires (Annoter des groupes de ressources selon des propriétés particulières). À partir d'une sous classe 'Group' on dérive une sous classe 'Component' pour décrire les restrictions théoriques correspondantes et une autre sous classe 'Community' pour définir sémantiquement les types de communautés **(Figure 144).**

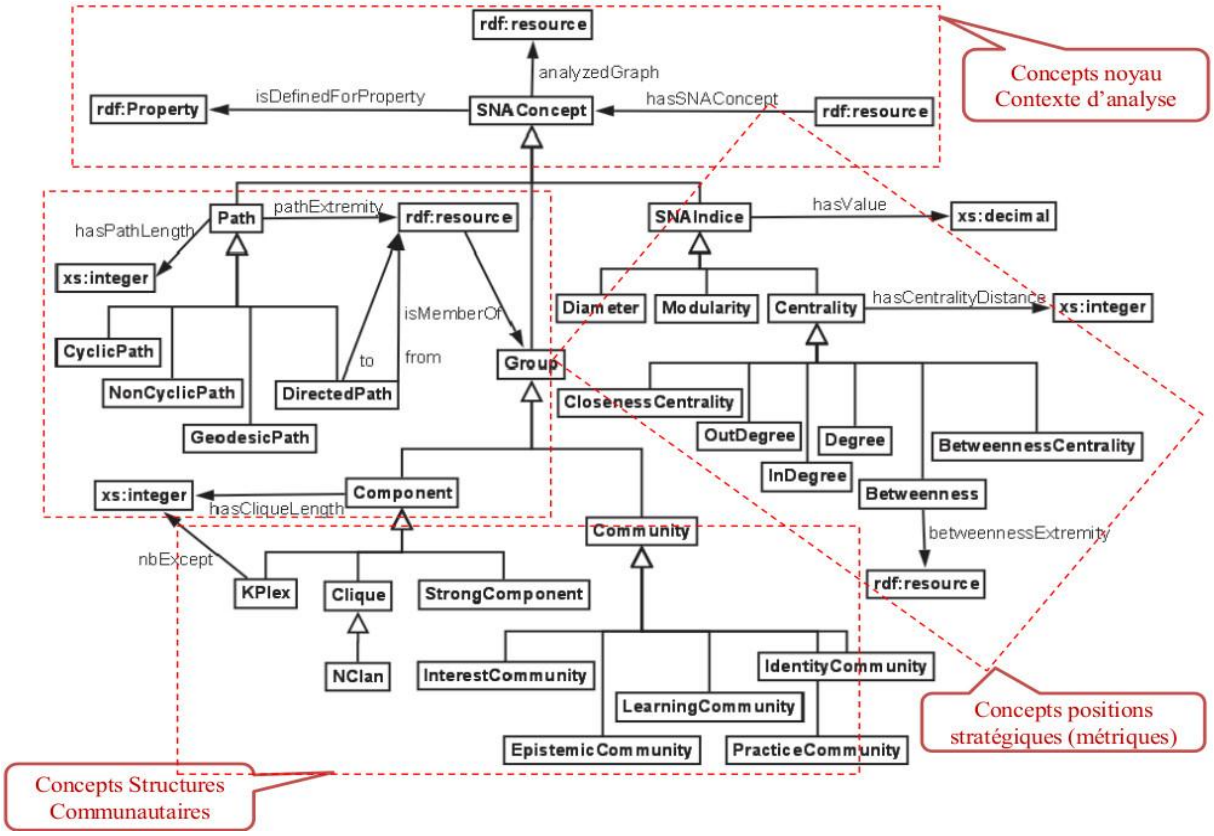


Figure 144. L'ontologie SemSNA pour enrichir les données sémantiques par les résultats de SNA. Un extrait de ((Erétéo 2011))

Ce cadre sémantique a été évalué sur un OSN de 'Ipernity.com', un graphe orienté étiqueté par 3 types de relations (préfère, ami et famille). Après avoir évalué les performances de certaines requêtes, il était claire que l'espace et le temps de calcule de constituaient le problème majeur ((Erétéo 2011)).

2.3.Regroupement et communautés sémantiques

Partant du même principe qui consiste à fusionner le caractère structurel des techniques de SNA et la richesse sémantique des SNs exprimée en web sémantique, des auteurs ont cherché à découvrir des structures communautaires ayant une dimension sémantique. L'une des approches proposées s'est basée sur l'exploitation des données de 'social tagging' notamment les tags et leur structuration sémantique pour étiqueter et détecter des communautés plus

significatives. C'est l'algorithme sémantique 'SemTagP' proposé par ((Erétéo et al 2011)) ((Erétéo 2011)). L'idée derrière 'SemTagP' est de fusionner trois aspects :

- Il se base sur l'algorithme de propagation (P) des étiquettes LPA ou RAK ((NGUYEN et al 2013)) ((Xie & Szymanski 2011)) ((Raghavan et al 2007)).
- Les étiquettes aléatoires laissent la place aux tags de 'Social Tagging' (Tag) ce qui va modifier significativement la stratégie de propagation.
- La structuration sémantique (Sem) de 'Social Tagging'/ folksonomies (Relations entre les tags et entre les tags et utilisateurs). Intuitivement, un tag représente un intérêt autour de lequel, toute une communauté d'intérêt peut se former (acteurs utilisant le même tag) ((Erétéo 2011)). En outre, les tags sont sémantiquement liés.

Autrement dit, 'SemTagP' transforme la propagation des étiquettes aléatoires en une propagation sémantique de Tags ((Erétéo 2011)) ((Erétéo et al 2011)). L'algorithme s'applique sur un graphe RDF typé formé SN sémantique proprement dit (profils et relations FOAF), plus une folksonomie structurée : relation entre tags (SKOS), relation user-tag (SCOT), (**Figure 145**). L'algorithme attribue aux nœuds, acteurs, les étiquettes (les tags) qu'ils utilisent. Les tags se propagent suivant la connectivité des users au niveau d'un type de relation donné et aussi suivant les relations de généralisation (skos:broader) et de spécialisation (skos:narrower) entre les tags (Algorithme 15). À chaque itération, on associe à un utilisateur 'u' le tag qui a le plus grand nombre d'occurrences parmi les tags utilisés par les voisins de 'u' et les tags sémantiquement liés par skos :broader avec les tags des voisins de 'u' (Algorithme 15). la méthode est implémentée dans le moteur de graphe sémantique KGRAM ((Erétéo 2011)) sous forme de requêtes.

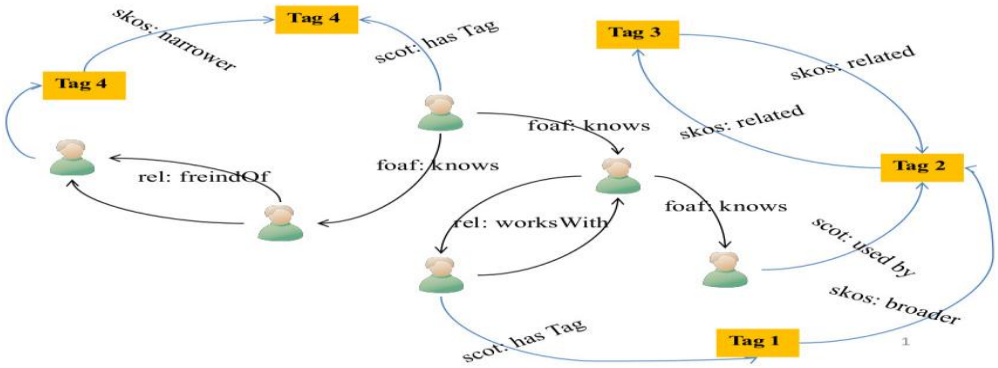


Figure 145. Structuration sémantique des relations user-tag et tag-tag

Algorithme 15. Pseudo-code de l'algorithme de propagation sémantique de Tags 'SemTagP' ((Erétéo et al 2011))

```
Algorithme SemTagP(RDFGraph network  Type type_relation)
{
RDFGraph old_network = network;
TANT QUE modularité (network) >= modularité (old_network)  FAIRE
–     Découvrir une nouvelle partition (une propagation)
      POUR chaque user dans  network.users  FAIRE
            user.tag = mostUsedNeighborTag(user,  type_relation);
RETURN old_network;
}
–  Méthode pour attribuer à un user pendnant une propagation le tag  choisi
–  qui a le plus grand nombre d'occurence parmi les tags  utilisé par ses vosins
–  et les tags sémantiquement liés avec les tags de ses voiisns
Tag  Méthode  mostUsedNeighborTags(User user, Type type_relation)
{
TagChoisi = null; max = 0; tagTable = new hashTable();
POUR chaque user_voisin dans user.neigbors[type_relation] FAIRE
–     Compter le nombre d'occurences du tag d'un voisin topologiquement lié
      SI tagTable.exists (user_voisin.tag)   ALORS
            tagTable[user_voisin.tag] ++;
      SINON
            tagTable[user_voisin.tag]  = 1
–     Chercher le tag le plus utilisté parmi les voisins
      SI max < tagTable[user_voisin.tag] ALORS
            TagChoisi = user_voisin.tag; max = tagTable[user_voisin.tag];
–     Compter le nombre d'occurences du tag lié sémantiquement par skos:braoder
–      avec les tags  du voisin
      POUR Chaque broaderTag dans  user_voisin.tag.broaders FAIRE
            SI tagTable.exists (broaderTag )   ALORS
                  tagTable[broaderTag] ++;
            SINON
                  tagTable[broaderTag]  = 1
            SI max < tagTable[broaderTag] ALORS
                  TagChoisi = broaderTag; max = tagTable[broaderTag];
Return TagChoisi
}
```

La qualité de partition obtenue après chaque itération de propagation est évaluée par une modularité particulière (modularité d'ER-Graph) ((Erétéo 2011)). On s'arrête quand modularité cesse de progresser. Les auteurs ont comparé l'évolution de modularité après chaque itération à partir de 4 variantes de propagation d'étiquettes, appliquées sur des données de collaborations de chercheurs universitaires (académiques) sur des thèses financées ((Erétéo 2011)). La centralisation de tel réseau ne permet pas d'avoir une structure modulaire de qualité avec une propagation des étiquettes aléatoire (RAK). Selon ((Erétéo et al 2011)), SemTagP amplifie l'avantage de propagation des tags en découvrant des groupes/sous-groupes bien connectés et avec plus de collectivité (intérêts sémantiquement liés). Sur le plan de visualisation, les communautés étiquetées par des tags désignant des sujets connexes forment des zones thématiques. Mais l'un des inconvénients se pose quand les tags ont énormément de relations sémantiques entre eux. Les tags décrivent des contextes et des sujets et ces sujets se chevauchent entre eux et les nuages de tags se chevauchent au point où la méthode peut agréger beaucoup d'acteurs dans une même communauté. Les auteurs ont pensé à une intervention manuelle (SemTagP contrôlé) en omettant certains tags pour éviter trop de généralisation ((Erétéo 2011)).

Loin du cadre de web sémantique :

Devant la complexité et le cout de calcule de SNA et notamment la détection des communautés qui se multiplient dans le cadre de web sémantique, une autre parti de chercheurs procèdent différemment pour expliquer la sémantique des communautés/ groupes, même sur un plan topologique. D'ailleurs, ((Zhou et al 2006)) sont les premiers qui ont introduit la notion des communautés sémantiques à partir des documents de communications.

Certains auteurs considèrent le groupe comme par exemple une collection dans un contexte spécifique : Des caractéristiques de nœuds, d'arcs, ou des propriétés structurelles (par exemple dynamique) du réseau (Kang et al 2007). Dans ce sens, (Kang et al 2007) pensent que la sémantique derrière la construction des groupes est basée sur deux éléments essentiels : les acteurs (un acteur (ID_A, $Attribut_1$,…, $Attribut_m$)), événements (un évènement E (ID_E, E_1,…, En, E_{time})), en ajoutant une relation de participation entre acteur et évènement P (ID_A, ID_E, P_1, P_2,…,P_p) (Kang et al 2007). Les évènements sont associés à la notion de temps (intervalle de temps, un point de temps de départ + duration, etc.). Dans le contexte de collaborations sur des publications scientifiques: l'acteur est un auteur, les évènements sont des publications, la participation l'auteur (co-auteur) d'un papier. La sémantique de regroupement (de collection) a été définie par des constructeurs qui s'appuient sur les valeurs d'attributs partagées par acteur, événement, ou une relation de participation (Kang et al 2007). On parle par exemple d'un regroupement par attribut d'acteur. La sémantique d'un groupe ($G_{acteur.Attribut_i = x}$) peut être définie à partir d'un sous-ensemble d'acteurs (connectés) qui ont la valeur x pour l'attribut Ai (Kang et al 2007) $G_{acteur.Attribut_i = x} = \{acteur \mid acteur.Attribut_i = x\}$. Les attributs qualitatifs (catégoriques) sont dans ce cas avantageux. C'est un mécanisme souple sur lequel il se base l'outil C-Group (Kang et al 2007). Comme les montre la Figure 42, les acteurs du réseau ont été regroupés par exemple selon le sujet de publication. ***Ce mécanisme sémantique de regroupement donne plus de flexibilité dans la définition des groupes selon plusieurs points de vue.***

D'un autre point de vue plus original des chercheurs pensent que les communautés se distinguent sémantiquement en 4 catégories :

- Une communauté d'intérêt est un type très fréquent, essentiellement thématique (de la musique, de la mode, etc.). Mais elle est peu compacte car les membres impliqués ne sont pas étroitement liés et la majorité est et ne participent qu'épisodiquement (Erétéo et al 2008).
- Une communauté d'apprentissage est une communauté prolifique où les membres sont plus actifs avec des échanges de savoirs (des maitres apprentis, groupe d'experts qui partagent ses connaissances avec les autres) ce qui engendre un sentiment d'appartenance plus fort (Erétéo et al 2008). Par exemple les contributeurs et administrateurs de Wikipédia forment une communauté d'apprentissage.
- Une communauté de pratique ou professionnelle est typiquement orientée vers la l'amélioration ou la réalisation d'un projet (Erétéo et al 2008), en empruntant le formalisme de l'entreprise.
- Une communauté d'identité s'appuie sur un sentiment d'appartenance plus fort, provoqué par une entière adhésion à une identité. Ce type de communautés est engendré souvent par les associations, les mouvements idéologiques, politiques, religieux et les minorités.