Comparaison de mesures de centralité basées sur les plus courts chemins dans les réseaux dynamiques

Marwan Ghanem*, Clémence Magnien*
Fabien Tarissan**

*Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, France
prenom.nom@lip6.fr,
https://www-complexnetworks.lip6.fr/~nom/
**Université Paris-Saclay, CNRS, ENS Paris-Saclay, ISP UMR 7220, France
fabien.tarissan@ens-paris-saclay.fr
https://www-complexnetworks.lip6.fr/tarissan/

Résumé. Définir l'importance des nœuds dans les réseaux statiques est une question de recherche très étudiée depuis de nombreuses années. Dernièrement, des adaptations des métriques classiques ont été proposées pour les réseaux dynamiques. Ces méthodes reposent sur des approches très différentes dans leur façon d'évaluer l'importance des nœuds à un instant donné. Il est donc nécessaire de pouvoir les évaluer et les comparer. Dans cet article, nous comparons trois approches existes pour mieux comprendre ce qui les différencie. Nous montrons que la nature des jeux de données influe grandement sur le comportement des méthodes, et que pour certains d'entre eux, la notion d'importance n'est pas toujours pertinente.

Depuis de nombreuses années, les chercheurs étudiant les réseaux complexes se sont intéressés à la question de l'importance des nœuds. Cela a conduit à l'introduction de plusieurs notions d'importance : centralité de degré, centralité de proximité ou centralité d'intermédiarité. Les principales approches s'appuient toutes sur la notion de chemin. En d'autres termes un nœud est important s'il est proche des autres nœuds ou si les chemins les plus courts passent par ce nœud. Récemment, des adaptations ont été introduites pour prendre en compte l'aspect temporel des réseaux complexes. Une première approche (Tang et al., 2010; Uddin et al., 2013) consiste à représenter un réseau dynamique comme une séquence de réseaux statiques. Une autre approche proposée par (Nicosia et al., 2013; Magnien et Tarissan, 2015) consiste à définir des chemins temporels comme une séquence de liens qui respecte l'ordre chronologique. Une autre approche encore consiste à construire un réseau statique à partir du réseau dynamique (Takaguchi et al., 2016) en dupliquant chaque nœud à l'instant auquel il interagit. Enfin d'autres propositions introduites notamment dans (Scholtes et al., 2016; Pan et Saramäki, 2011) prennent en compte les aspects temporels dans ces jeux de données mais ne permettent d'obtenir qu'une seule valeur globale d'importance. Dans cet article nous étudions les trois approches qui considèrent l'importance de nœud par plusieurs valeurs, et nous les comparons pour mieux comprendre leurs différences.

1 Définitions

Dans cette section, nous présentons les trois méthodes que nous comparerons dans la suite de cet article.

Une première approche a été présentée dans (Uddin et al., 2013). Les auteurs représentent un réseau dynamique sous la forme d'une séquence de réseaux statiques (*snapshots*). Chaque réseau résulte de l'agrégation de tous les liens pendant une période, qui est de même durée pour tous les *snapshots*. Avec cette représentation, une métrique de centralité classique peut être calculée sur chaque *snapshot*. Les nœuds ayant une centralité élevée dans un *snapshot* sont considérés comme les nœuds les plus importants pendant la période correspondant à ce *snapshot*. Ainsi, les nœuds qui sont le plus souvent important dans chaque *snapshot* sont considérés comme les plus importants sur toute la durée de vie du réseau. Dans cet article, nous considérons cette approche avec la centralité de proximité classique. Cette approche a une complexité de $O(\mid V \mid \cdot \mid E \mid + \mid V \mid)$ par *snapshot*, où V est le nombre de nœuds et E est le nombre de liens . Nous ferons référence à cette méthode par le terme *snapshot*.

La deuxième approche a été présentée dans (Magnien et Tarissan, 2015). Les auteurs étudient un réseau dynamique $G = (V, E)$ où V est l'ensemble des nœuds et E est l'ensemble des liens de la forme (u, v, t) tel que $u, v \in V$ et où t est une étiquette temporelle. Avec cette représentation, un chemin temporel de v_0 à v_{k+1} qui commence à t_s , consiste en une séquence de liens ($(v_0, v_1, t_0), (v_1, v_2, t_1), \ldots, (v_i, v_{i+1}, t_i) \ldots, (v_k, v_{k+1}, t_k)$) tels que $t_i < t_{i+1} \forall i, i = 0..k-1$ et $t_0 > t_s$. La durée de ce chemin est égale à $t_k - t_s$. Ce chemin est considéré étant le plus court chemin s'il a la plus courte durée parmi tous les chemins de v_0 à v_{k+1} qui commencent à t_s. On note $d_{t_s}(v_0, v_{k+1})$ la durée correspondante, appellée la distance temporelle. S'il n'y a pas de chemin de v_0 à v_{k+1} qui commence à t_s, nous considérons alors que $d_{t_s}(v_0, v_{k+1}) = \infty$. Les auteurs définissent la *temporal closeness* d'un nœud u à l'instant t est par $C_t(u) = \sum_{v \neq u} \frac{1}{d_t(u,v)}$.

Notons que cette définition nécessite que la centralité soit calculée à chaque pas de temps t, ce qui est très coûteux en termes de calcul. C'est pourquoi par la suite, nous calculons la *temporal closeness* de chaque nœud toutes les I secondes seulement, c'est qui donne une complexité en $O((D/I)^2)$ ou D est la durée du trace. La valeur de I est basée sur la médiane de la durée du temps inter-contact (le temps écoulé entre deux liens consécutifs) [1].

Dans la troisième approche (Takaguchi et al., 2016), une copie de chaque nœud est créée pour chaque instant où il interagit. Ainsi, un nœud u du réseau dynamique est représenté sous la forme d'un ensemble de nœuds (u, t) où t correspond aux instants où u est actif. Chaque paire consécutive de cet ensemble ((u, t_n) et (u, t_{n+1})) est liée par un lien. Les liens originaux du réseau sont conservés. Cette méthode permet de construire un réseau statique qui respecte la temporalité des interactions. Les auteurs proposent une métrique de centralité, la *temporal coverage centrality*. Elle mesure l'importance d'un nœud (u, t) par la fraction de paires de nœuds ayant un plus court chemin qui passe par (u, t). Le calcul de chacun de ces nœuds ((u, t)) a une complexité en $O(\mid V \mid^2 log(\mid E \mid))$, ce qui rend cette approche coûteuse. Nous faisons référence à cette méthode par le terme *coverage*.

1. Le programme que nous avons utilisé pour calculer les métriques est disponible publiquement sur le lien https ://bitbucket.org/complexnetworks/closeness_centrality_marwan.

2 Jeux de données et analyses

Nous avons utilisés six jeux de données pour cette étude. Cependant, nous avons observés que lorsqu'ils étaient de nature similaire, ils donnaient des résultats similaires. C'est pourquoi nous présentons ici seulement trois jeux de données représentatifs de l'ensemble. [2].

— Enron (Shetty et Adibi, 2005) : contient $47\,088$ courriels échangés entre 151 employés pendant trois ans. Pour chaque courriel nous avons l'expéditeur, le destinataire et la date d'expédition,

— RollerNet (Tournoux et al., 2009) : représente les contacts physiques entre 62 participants lors d'une sortie en rollers à Paris en août 2006. Il contient $403\,834$ contacts entre les participants répartis sur environ trois heures,

— Twitter : enregistre les tweets de comptes associés à des groupes terroristes. Chaque nœud représente un hashtag et chaque lien représente un tweet qui contient deux hashtags (nœuds). Par conséquent, un tweet avec plusieurs hashtags génère plusieurs liens. Le jeu de données contient 3048 hashtags et $100\,429$ liens pendant 22 jours.

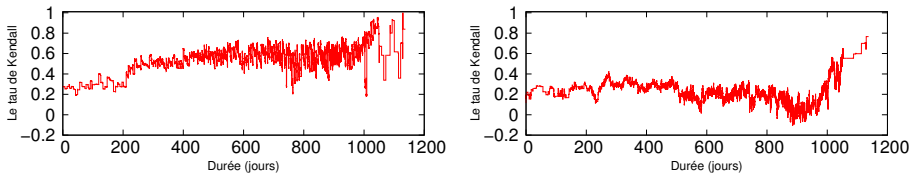
2.1 Comparaison au fil du temps

Afin de pouvoir comparer les méthodes entre elles, ainsi que les nœuds entre eux sur chaque jeu de données, nous ne pouvons pas prendre en compte uniquement la centralité. C'est pourquoi nous commençons par ordonner les nœuds à chaque instant. Nous utilisons pour cela la méthode *inverse compétition ranking* [3]. Grâce à cette méthode de classement, pour chaque instant de chaque jeu de données, nous avons un ordre sur les nœuds pour chaque méthode. Nous pouvons comparer ces ordres entre eux, en utilisant le taux de Kendall. Pour deux ordres, le taux de Kendall renvoie une valeur comprise entre -1 et 1. Cette valeur représente la corrélation entre ces deux ordres. 1 indique que les deux ordres sont parfaitement corrélés alors que -1 indique qu'ils ont une corrélation inverse parfaite. Soit $r_k(i)$ le rang du nœud i dans l'ordre k, pour deux ordres r_1, r_2 le taux de Kendall est défini formellement de la manière suivante : $\mathcal{K}(r_1, r_2) = 1 - \frac{|\{i,j\}:r_1(i)>r_1(j) \text{ and } r_2(i)<r_2(j)|}{|V|(|V|-1)/2}$.

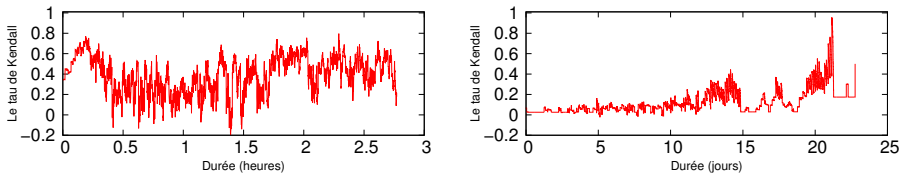
Nous regardons tout d'abord l'évolution du taux de Kendall entre la *temporal closeness* et *snapshot*. Nous présentons dans la figure 1 (a) l'évolution du taux pour Enron. Nous pouvons voir que la corrélation est basse au début et augmente avec le temps. Ceci est dû au fait qu'une grande proportion de nœuds est inactive au début et par conséquent la méthode *snapshot* leur attribue le rang le plus bas. Par contre, certains nœuds ont des chemins temporels vers d'autres nœuds grâce à des interactions qui vont avoir lieu plus tard dans le jeu de données. *Temporal closeness* leur attribue une valeur non nulle, donc un rang non nul. Plus tard, ces nœuds deviennent actifs, puisque le réseau évolue, donc ces nœuds sont pris en compte par la méthode de *snapshot* ce qui augmente la corrélation. Nous pouvons également constater que ce phénomène n'est pas limité au début de l'évolution. Les brusques chutes que nous observons (par exemple au $1000ieme$ jour) sont liées qu' à ces moments la, un grand nombre de nœuds est devenu inactif. Par conséquent la corrélation diminue, pour les raisons présentées ci-dessus.

2. Comme dit précédemment l'approche *coverage* est très coûteuse donc nous avons seulement appliqué *coverage* sur Enron

3. Cette méthode attribue le rang 0 aux nœuds les moins centraux, ensuite chaque nœud a un rang égal au nombre de nœuds moins centraux que lui



(a) *Temporal closeness* et *snapshot* sur Enron (b) *Temporal closeness* et *coverage* sur Enron



(c) *Temporal closeness* et *snapshot* sur Rollernet (d) *Temporal closeness* et *snapshot* sur Twitter

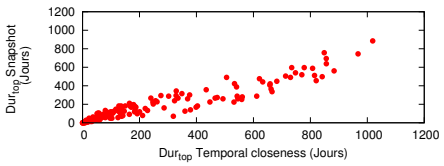
FIG. 1: Évolution du tau de Kendall

Pour Rollernet (Fig. 1 (c)), nous pouvons observer que l'évolution est différente d'Enron. Le taux fluctue fortement et est globalement plus bas et peut également être négatif à certains instants. Ces observations sont liées à l'activité élevée dans Rollernet. Cette activité rend chaque *snapshot* du réseau beaucoup plus dense que ceux analysés dans Enron, ce qui augmente les chances que *snapshot* prenne en compte des chemins temporellement impossibles. Ces chemins ne sont pas pris en compte par *temporal closeness*, donc la corrélation devient naturellement basse.

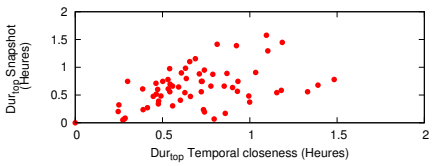
Ensuite, nous nous concentrons sur Twitter (Fig. 1 (d)). Ce jeu de données a une faible activité ce qui induit une corrélation basse. La corrélation augmente seulement au 14 ème, jour quand un grand nombre de nœud devient actif, pour diminuer de nouveau quand l'activité diminue. Nous pouvons identifier certains instants avec des corrélations élevées, qui sont également liées aux pics d'activité.

Nous considérons enfin la corrélation entre *Temporal closeness* et *coverage* sur Enron. La figure 1 (b) présente l'évolution de cette corrélation pour les deux méthodes. Nous pouvons voir que, globalement, la corrélation est assez basse sauf vers la fin. Puisque *coverage* et *temporal closeness* prennent toutes les deux en compte les liens dans le futur, ceci suggère que cette faible corrélation est due à la différence de notion d'importance entre elles. Finalement la hausse à la fin est due au fait que très peu de nœuds sont actifs. Ceci compense la différence de point du vue entre les deux approches, car il suffit d'être actif pour être important.

Nous pouvons constater par cette première analyse que la corrélation entre *temporal closeness* et *snapshot* dépend de la proportion de nœuds actifs. Nous avons remarqué des limitations à la méthode *snapshot* : (i) *snapshot* est incapable de détecter l'importance d'un nœud inactif puisqu'il ne prend pas en compte les liens futurs et (ii) les grands pics d'activité augmentent les chemins temporellement impossibles, en conséquence l'estimation de l'importance peut être biaisée.



(a) Enron



(b) Rollernet

FIG. 2: Les valeurs de Dur_{top} calculées pour chaque nœud pour *temporal closeness* et *snapshot*

2.2 Différence globale

Dans cette partie, nous cherchons à comprendre plus précisément la différence observée entre *temporal closeness* et *snapshot* sur Enron et Rollernet, car la différence de corrélation est relativement élevée. Pour cela, nous définissons une plage de rangs pour laquelle nous estimons que les nœuds sont importants. Cette plage correspond aux 25% des rangs les plus élevés. Autrement dit, pour un réseau avec n nœuds, cela concerne les nœuds ayant un rang supérieur à $\lfloor n * 0.75 \rfloor$. Puis, nous définissons une valeur (Dur_{top}) qui correspond à la durée pendant laquelle un nœud est important. Nous considérons qu'un nœud est présent dans cette plage entre l'instant où nous calculons la centralité jusqu'à l'instant suivant. Plus formellement soit un nœud u et $R(u) = (r_i)_{i=1...k}$ une séquence de rangs de u, nous définissons Dur_{top} la valeur : $Dur_{top}(u) = I \cdot |\{i \leq k-1, r_i \geq \lfloor n * 0.75 \rfloor\}|$.

La figure 2 (a) présente pour chaque nœud de Enron, la valeur Dur_{top} calculée par *temporal closeness* et *snapshot*. Nous pouvons voir que ces valeurs sont corrélées : les nœuds ayant les durées les plus élevées sont communs aux deux méthodes. De plus, nous pouvons voir que certains nœuds ont des valeurs beaucoup plus élevées que les autres. Ils sont donc beaucoup plus importants.

La figure 2 (b) présente les valeurs calculées par *temporal closeness* et *snapshot* pour les nœuds de Rollernet. Nous pouvons observer que la corrélation est moins nette que pour Enron. Comme expliqué précédemment, la forte activité de RollerNet entraîne la prise en compte de chemins temporellement impossible par *snapshot*. De plus, nous pouvons voir que les nœuds ne sont jamais considérés comme importants sur une durée plus longue que la moitié de la trace. Ainsi, aucun nœud ne se démarque des autres par une valeur de Dur_{top} élevée, contrairement à ceux d'Enron. Nous pouvons conclure que la notion d'importance globale n'existe pas forcément dans ce cas, car tous les nœuds sont d'importance similaire.

3 Conclusion

Dans cet article, nous présentons une comparaisons de différentes approches existantes pour mesurer l'importance des nœuds dans les réseaux dynamiques. Nous avons montré que (i) la méthode *snapshot* est incapable d'anticiper l'importance d'un nœud, (ii) la méthode *snapshot* peut être biaisée par des activités élevées, (iii) les différentes centralités ne détectent pas forcément les mêmes nœuds importants, (iv) dans certains jeux de données, la notion d'importance globale peut être sans signification. Il est à noter que nous nous sommes ici intéressés aux

méthodes basées sur les plus courts chemins. Nous espérons pouvoir comparer ces méthodes avec celles basées sur des approches spectrales (vecteurs propres des matrices d'adjacences notamment). Enfin, l'accès aux jeux de données avec une vérité du terrain nous permettrait de mieux comprendre ces différences.

Références

Magnien, C. et F. Tarissan (2015). Time evolution of the importance of nodes in dynamic networks. In *Proceedings of the International Symposium on Foundations and Applications of Big Data Analytics (FAB), in conjunction with ASONAM, 2015.*, FAB '15, New York, NY, USA, pp. 1200–1207. ACM.

Nicosia, V., J. Tang, C. Mascolo, M. Musolesi, G. Russo, et V. Latora (2013). Graph metrics for temporal networks. In *Temporal networks*, pp. 15–40. Springer.

Pan, R. K. et J. Saramäki (2011). Path lengths, correlations, and centrality in temporal networks. *Physical Review E 84*(1), 016105.

Scholtes, I., N. Wider, et A. Garas (2016). Higher-order aggregate networks in the analysis of temporal networks : path structures and centralities. *Eur. Phys. J. B 89*, 61.

Shetty, J. et J. Adibi (2005). Discovering important nodes through graph entropy the case of Enron email database. In *Proceedings of the 3rd international workshop on Link discovery - LinkKDD '05*, New York, New York, USA, pp. 74–81. ACM Press.

Takaguchi, T., Y. Yano, et Y. Yoshida (2016). Coverage centralities for temporal networks. *The European Physical Journal B 89*(2), 35.

Tang, J., M. Musolesi, C. Mascolo, V. Latora, et V. Nicosia (2010). Analysing information flows and key mediators through temporal centrality metrics. In *Proceedings of the 3rd Workshop on Social Network Systems*, pp. 1–6. ACM.

Tournoux, P. U., J. Leguay, M. Dias de Amorim, F. Benbadis, V. Conan, et J. Whitbeck (2009). The Accordion Phenomenon : Analysis, Characterization, and Impact on DTN Routing. In *Proceedings of the 28rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, pp. 1116–1124. IEEE.

Uddin, M. S., P. Mahendra, K. S. K. Chung, et L. Hossain (2013). Topological analysis of longitudinal networks. In *HICSS*.