Coefficient de Clustering d'intérêt : une nouvelle métrique pour les graphes dirigés comme Twitter

Thibaud Trolliet, Nathann Cohen, Frédéric Giroire, Luc Hogie, Stéphane Pérennes

Coefficient de Clustering d'intérêt : une nouvelle métrique pour les graphes dirigés comme Twitter

Trolliet Thibaud [1], Cohen Nathann[2], Giroire Frédéric[2], Hogie Luc[1], et Pérennes Stéphane[2]

[1] *Université Côte d'Azur/CNRS, France*
[2] *INRIA Sophia-Antipolis, France*

Nous étudions dans ce papier le coefficient de clustering de graphes sociaux *dirigés*. Le coefficient de clustering a été introduit pour capturer le phénomène social selon lequel les amis de mes amis sont mes amis, et a été largement étudié depuis, se montrant d'un grand intérêt pour décrire les caractéristiques sociales d'un graphe. Cependant, ce paramètre est adapté pour un graphe dans lequel les liens ne sont pas orientés, comme les liens d'amitiés (Facebook) ou les liens professionnels (LinkedIn), mais devient inadapté pour un graphe dans lequel les liens sont dirigés d'une source d'informations vers un consommateur d'informations. Nous montrons que les études précédentes ont manqué une grande partie des informations contenues dans la partie dirigée de ces graphes.

Dans ce papier, nous introduisons une nouvelle métrique pour mesurer le clustering d'un graphe social orienté avec des liens d'intérêt, que l'on nomme le *coefficient de clustering d'intérêt*. Nous calculons sa valeur, exactement et à l'aide de méthodes d'échantillonnage, sur un graphe de Twitter comprenant 505 millions de noeuds et 23 milliards d'arêtes.

Nous mesurons en outre les valeurs des coefficients de clustering dirigés et non dirigés précédemment introduits dans la littérature, une première sur un si grand graphe. Nous montrons que le coefficient de clustering d'intérêt est plus grand que les coefficients de clustering dirigés classiques. Cela montre la pertinence de cette nouvelle métrique pour capturer l'aspect informatif des graphes dirigés.

Mots-clefs : Systèmes Complexes, Coefficient de Clustering, Graphes Dirigés, Réseaux Sociaux, Twitter.

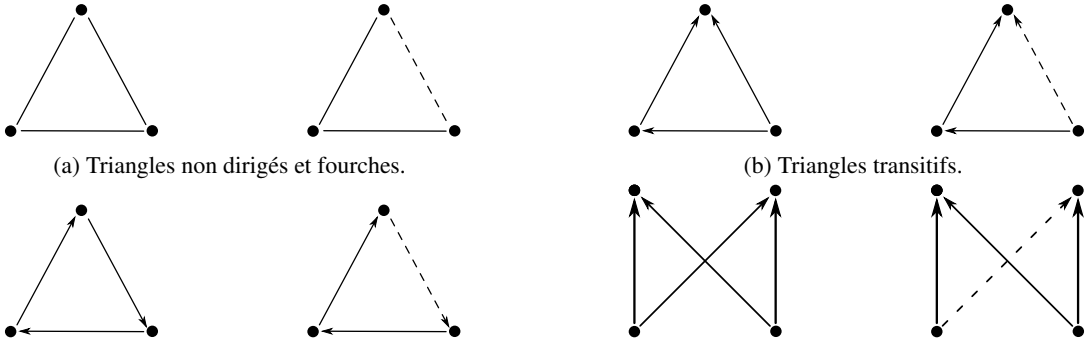
1 Introduction

Les réseaux apparaissent dans un grand nombre de systèmes complexes, que ce soit dans le domaine de la biologie, du social, de l'internet, ... La plupart des réseaux issus du monde réel possèdent des propriétés communes. Dans ce papier, nous porterons notre attention sur le coefficient de clustering dans les réseaux sociaux. Introduit pour la première fois en 2002 par Barrat et Weight [BW00], le coefficient de clustering global est défini comme :

$$cc = 3 \times \frac{\#\text{ triangles dans le graphe}}{\#\text{ fourches dans le graphe}},$$

où une fourche est définie comme un chemin de longueur 2. Dans le cas des réseaux sociaux, cela peut être interprété comme la probabilité pour qu'un ami d'ami soit aussi mon ami. Le coefficient de clustering est devenu une des métriques les plus étudiées dans les réseaux sociaux, et est utilisée dans de nombreuses applications, comme la détection de spams, les recommandations, l'étude de propagation d'informations,...

Cependant si cette métrique est adaptée aux graphes sociaux non dirigés, celle-ci ne l'est plus pour les graphes dirigés comme Twitter ou Instagram, qui ont une fonctionnalité d'information tout autant que social [MSGL14]. Dans ce papier, nous proposons une nouvelle métrique de clustering, nommée *coefficient de clustering d'intérêt*, adaptée aux réseaux d'informations dirigés - voir section 2. Nous proposons en section 3 des algorithmes pour calculer cette nouvelle métrique, à la fois exactement et par échantillonnage, sur une capture du graphe complet de Twitter de 2012, composé de 505 millions de noeuds et 23 milliards d'arêtes. A notre connaissance, il s'agit de la première fois qu'un tel calcul est effectué sur un graphe dirigé de cette envergure. Les valeurs obtenues, présentées en section 4, montrent l'intérêt de la nouvelle métrique.



(a) Triangles non dirigés et fourches.

(b) Triangles transitifs.

(c) Triangles cycliques.

(d) K22s.

FIGURE 1: Triangles et K22 fermés et ouverts (fourches).

2 Coefficient de clustering d'intérêt

Le coefficient de clustering présenté en introduction est défini pour les graphes non dirigés. Il existe plusieurs méthodes pour l'étendre aux graphes dirigés :

— Rendre le graphe non dirigé. Les deux principales façon de faire cela sont, soit en conservant chaque lien du graphe et retirant sa direction - nous appelerons le graphe ainsi formé le *graphe entier non dirigé* -, soit en ne conservant que les arêtes bidirectionnelles, c'est-à-dire garder une arête entre u et v si et seulement si il existe une arête de u à v, et une de v à u - nous l'appelerons *graphe mutuel* -.

— Rendre les triangles dirigés et adapter la définition. Il existe deux possibilités pour rendre un triangle dirigé : *cyclique*, où $u \to v \to w \to u$, et *transitif*, où $u \to v \to w$ et $u \to w$. Il existe ensuite différentes façon de définir les fourches, nous ne nous intéresserons ici qu'aux fourches du type $u \to v \to w$.

Ces quatre possibilités permettent de définir le coefficient de clustering à partir de la définition d'origine de quatre manières différentes - respectivement les *coefficients de clustering non dirigé (ccnd)*, *mutuel (ccm)*, *cyclique (ccc)*, et *transitif (cct)*. Cependant, nous pensons qu'aucune de celles-ci ne capture l'aspect informationnel des réseaux sociaux dirigés comme Twitter. En effet, ceux-ci se basent sur le schéma "les amis de mes amis sont-ils mes amis ?", qui correspond à une question sociale. Nous proposons une nouvelle définition du coefficient de clustering, le *coefficient de clustering d'intérêt* ou *cci*, basé cette fois-ci sur les intérêts communs que deux personnes peuvent avoir, afin de capturer cet aspect d'information. Le coefficient de clustering d'intérêt est défini comme :

$$\text{cci} = 4 \times \frac{\text{\# K22 dans le graphe}}{\text{\# K22 ouverts dans le graphe}} \tag{1}$$

où un K22 est défini comme un ensemble de 4 noeuds tels que deux d'entre eux suivent les deux autres, et un K22 ouvert est un K22 avec un lien manquant. Le facteur 4 est là pour se ramener à une probabilité comprise entre 0 et 1. Cette définition correspond à la question : "sachant que je partage un intérêt en commun avec une personne, quelle est la probabilité qu'il partage aussi cet autre intérêt que j'ai ?" Cette métrique est donc bien centrée sur les centres d'intérêts plus que sur les liens sociaux.

3 Algorithmes pour calculer les différents clustering

3.1 Graphe de Twitter

Pour étudier cette nouvelle métrique, nous nous proposons de calculer sa valeur, ainsi que celle des autres coefficients de clustering discutés, dans un graphe dirigé réel. Le graphe étudié est une capture du graphe de Twitter, récupérée en 2012 par Gabielkov et Legout [GL12]. Chaque noeud correspond à un compte, et il existe un lien dirigé entre les noeuds u et v si le compte associé au noeud u suit le compte associé au noeud v. Le graphe possède 505 millions de noeuds et 23 milliards d'arêtes, faisant de celui-ci l'un des plus gros graphes dirigés disponibles actuellement.

3.2 Algorithmes de calcul

Afin de calculer la valeur des différents coefficients de clustering, nous avons utilisé 3 méthodes : un compte exact, une estimation par échantillonnage sur les arêtes, et une estimation à l'aide d'un algorithme de Monte-Carlo. Nous présentons d'abord ces algorithmes avant de voir les résultats en section 4.

3.2.1 Calcul exact

Nous présentons ici l'algorithme permettant de calculer le nombre exact de K22 dans un graphe dirigé. Le calcul du nombre de K22 ouverts suit une logique similaire, tout comme ceux des triangles et fourches, et ont une complexité moindre, les triangles ne contenant que 3 noeuds contre 4 pour les K22.

Une première idée d'algorithme trivial est de considérer chaque quadruplet de noeuds et, pour chacun d'eux, vérifier la présence de K22 ou K22 ouverts. Celui-ci nous donnerait une complexité de l'ordre de $O(n^4)$. Dans le cas du graphe de Twitter, cela nous amène à 10^{34} opérations environ.

Nous proposons ici un algorithme amélioré. L'algorithme est le suivant : pour chaque noeud x, on compte combien de fois chaque noeud w apparait comme voisin sortant des voisins entrant de x ; notons cette valeur $\#vc_x(w)$. w formera un K22 avec x pour chaque pair de voisins entrants en commun avec x. Ils formeront donc ensemble $\binom{\#vc_x(w)}{2}$ K22. On peut alors exprimer le nombre total de K22 comme :

$$\#K22 = \sum_{x \in V} \sum_{w | \#vc_x(w) \geq 2} \binom{\#vc_x(w)}{2} \tag{2}$$

La complexité de cet algorithme est de $m + \sum_{u \in V} d^+(u)(d^+(u) - 1)$, avec m le nombre de liens du graphe et $d^+(u)$ le degré sortant du noeud u. En effet, chaque lien est seulement considéré une fois en tant qu'arc entrant, et $d^+ - 1$ fois en tant qu'arc sortant. Dans le graphe de Twitter, la somme des carrés des degrés sortants est égale à $8 \cdot 10^{13}$. Le nombre d'itérations nécessaires pour calculer le nombre de K22 dans Twitter avec cet algorithme est donc de l'ordre de 10^{14} opérations, soit prêt de 20 ordres de grandeurs de moins que la méthode triviale.

3.2.2 Méthodes de calcul approché

Le calcul sur des graphes de cette ampleur n'étant pas aisé, nous proposons aussi deux méthodes de calcul approché du nombre de K22 et K22 ouverts :

— Une méthode par échantillonnage sur les arêtes : on crée le graphe G_p de sorte que chaque arête du graphe initial G a une probabilité p d'être encore présente dans le graphe échantillonné G_p. On peut ensuite utiliser l'algorithme présenté précédemment pour compter le nombre de K22 et K22 ouverts de ce nouveau graphe. Notons que chaque K22 a alors une probabilité p^4 de se retrouver dans le nouveau graphe, et chaque K22 ouvert une probabilité p^3. Nous avons vérifié qu'en pratique, la plupart des K22 et K22 ouverts ne partagent aucun lien ; cela nous permet de déduire facilement l'icc du graphe total à partir du graphe échantillonné :

$$\mathbb{E}(cci_G) = 4 \times \frac{\mathbb{E}(\#\text{K22 dans G})}{\mathbb{E}(\#\text{K22 ouverts dans G})} = 4 \times \frac{\mathbb{E}(\#\text{K22 dans } G_p\ )/p^4}{\mathbb{E}(\#\text{K22 ouverts dans } G_p)/p^3} = \frac{\mathbb{E}(cci_{G_p})}{p} \tag{3}$$

— Un algorithme de Monte-Carlo : en répétant n fois la procédure de sélectionner aléatoirement un noeud v proportionnellement au carré de son degré entrant, puis uniformément aléatoirement deux voisins entrants u_1 et u_2, et compter le nombre de K22 et K22 ouverts contenant cette fourche (u_1v, u_2v), on peut montrer que le coefficient de clustering d'intérêt peut être estimé par :

$$4 \times \frac{\sum_{i=1}^{n} \#K22(u_1^i v^i, u_2^i v^i)}{2 \sum_{i=1}^{n} \#K22_{ouvert}(u_1^i v^i, u_2^i v^i)} \underset{n \to \infty}{\to} cci \tag{4}$$

À titre d'exemple, alors que le temps de calcul de l'algorithme exact a pris plusieurs jours, la méthode par échantillonnage avec p=1/1000 et l'algorithme de Monte-Carlo ont permis d'obtenir des erreurs inférieures au pourcent en moins de 1 minute.

	cci	*cct*	*ccc*	*ccnd*	*ccm*
Twitter total	3.3%	1.9%	1.7%	0.11%	10.7%
Twitter sans structures mutuelles	3.1%	0.51%	0.24%	0.057%	x

TABLE 1: Coefficients de clustering dans le graphe de Twitter, avec et sans les structures mutuelles.

4 Valeurs des coefficients de clustering dans Twitter

Le tableau 1 présente les valeurs des différents coefficients de clustering obtenues avec l'algorithme exact sur le graphe de Twitter. La première ligne concerne les valeurs trouvées dans le graphe total. On voit que la valeur du coefficient de clustering d'intérêt (*cci*) est plus élevée que les autres valeurs, hormis celle du coefficient de clustering mutuel (*ccm*). Cette dernière concerne le graphe contenant uniquement les liens bidirectionnels ; or, nous pensons que ce sont ces liens qui contiennent l'aspect social du graphe de Twitter, tandis que les liens unidirectionnels contiennent celui d'information. Ainsi, dans le graphe composé uniquement des liens sociaux (*ccm*), le coefficient de clustering classique est du même ordre de grandeur que celui trouvé dans d'autres réseaux sociaux non dirigés, environ 3%. Par contre, dans le graphe comprenant à la fois les liens sociaux et d'information (*ccnd*), la valeur classique est bien plus faible (0.11%).

La seconde ligne présente les valeurs de clustering auxquelles nous avons retiré les triangles et fourches composés uniquement de liens bidirectionnels. En effet, un triangle bidirectionnel va induire dans le graphe dirigé deux triangles cycliques, et six triangles transitifs. Ces triangles représentant les liens sociaux, les retirer permet de se concentrer sur le caractère d'information du graphe. De la même manière, nous avons retiré les K22 et K22 ouverts composés uniquement de liens bidirectionnels. Une fois l'aspect social retiré, la valeur du *cci* est environ un ordre de grandeur plus élevée que les autres mesures de clustering basés sur les triangles. Cela montre bien la capacité de cette métrique à capturer l'aspect informationnel du graphe.

Nous avons aussi calculé la valeur des différents clustering dans d'autres jeux de données de graphes dirigés, notamment Instagram. Les résultats étaient similaires à ceux trouvés dans Twitter.

5 Conclusion

Dans ce papier, nous avons introduit une nouvelle métrique, le *coefficient de clustering d'intérêt*, permettant de capturer l'aspect informationnel des graphes dirigés. Celui-ci est basé sur l'idée que, si deux personnes ont un intérêt commun, ils ont une probabilité plus élevée d'avoir un autre intérêt commun. Nous avons calculé cette métrique sur un graphe connu pour être à la fois un média social et d'information, une capture du graphe de Twitter avec 505 millions de noeuds et 23 milliards d'arêtes. Nous proposons des algorithmes pour calculer ces clusterings, de façon exact comme approché. La valeur trouvée du coefficient de clustering d'intéret, autour de 3.3%, est plus grande que les valeurs de clustering basés sur les triangles.

Nous avons aussi remarqué parallèlement que le nombre de K22 dans le graphe de Twitter était bien plus élevé que le nombre de triangles. Cela nous pousse à croire que cette nouvelle métrique peut permettre de proposer une nouvelle méthode de recommandation, basée sur la fermeture de K22 ouverts, plutôt que de fermer des triangles ouverts - une des méthodes utilisées aujourd'hui par Twitter. Le grand nombre de K22 permet une bien meilleure précision sur les différentes recommandations possibles, améliorant grandement celles-ci. Les tests préliminaires effectués sont prometteurs, nous incitant à approfondir cette application.

Références

[BW00] Alain Barrat and Martin Weigt. On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13(3) :547–560, 2000.

[GL12] Maksym Gabielkov and Arnaud Legout. The complete picture of the twitter social graph. In *Proceedings of the 2012 ACM conference on CoNEXT student workshop*, pages 19–20. ACM, 2012.

[MSGL14] Seth A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information network or social network ? : the structure of the twitter follow graph. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 493–498. ACM, 2014.