Trouver des séquences de contacts pertinentes dans un flot de liens

Noe Gaumont

Trouver des séquences de contacts pertinentes dans un flot de liens

Noé Gaumont[1] †

[1] *Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu 75005 Paris.*

Un flot de liens est une suite de quadruplets (b,e,u,v) indiquant qu'une interaction a eu lieu entre les entités u et v sur l'intervalle $[b,e]$. Les échanges d'emails, le trafic IP, ou les appels téléphoniques se modélisent de cette façon. La recherche de structures dans ce type de données est un problème très étudié. Nous explorons ici la recherche de sous flots pertinents, c'est-à-dire de sous-ensembles d'interactions denses à la fois structurellement et temporellement. Nous proposons une méthode utilisant un algorithme de détection de communautés et un moyen d'évaluer les groupes trouvés utilisant une définition de densité adaptée aux flots de liens. Nous montrons la pertinence des structures trouvées sur plusieurs réseaux de contacts réels.

Mots-clefs : flot de liens, réseaux de contacts, densité, groupes denses

1 Introduction

Les flots de liens permettent de modéliser et d'étudier les réseaux dynamiques. Un flot de liens est une suite de quadruplets (b,e,u,v) indiquant qu'une interaction a eu lieu entre les entités u et v sur l'intervalle $[b,e]$. Nous cherchons à trouver dans un flot de liens des sous-ensembles d'interactions qui soient pertinents. Dans le cadre statique, cela se traduit par la recherche de groupes denses [BBC+15]. La densité a été étendue aux flots de liens [VL14]. Elle permet de mesurer à quel point un groupe est dense structurellement et temporellement. Cependant un groupe dense n'est pas forcément pertinent. En effet, un sous-ensemble d'interactions n'est pertinent que s'il est plus dense que son voisinage. La notion de voisinage, que nous définissons formellement par la suite, est cruciale pour prendre en compte le contexte dans lequel les interactions ont lieu. Par exemple, une réunion de 5 personnes est plus surprenante si ces 5 personnes ne se rencontrent qu'une seule fois que si elles se rencontrent tous les midis.

Pour rechercher une structure dans un réseau dynamique, beaucoup de méthodes s'appuient sur des séquences de graphes, où chaque graphe est une agrégation temporelle, voir l'état de l'art de Hartman *et al.* [HKW14]. D'autres méthodes [RTG14, ELS15] cherchent la zone la plus dense directement dans le réseau dynamique. Elles n'obtiennent donc qu'un seul groupe et il n'y pas de prise en compte du voisinage De plus, les notions de densité utilisées sont le degré moyen du groupe dans un graphe agrégé.

Notre démarche se distingue sur deux aspects. Nous utilisons une notion de densité qui mélange les aspects topologiques et temporels, et nous considérons des groupes de liens au lieu de groupes de nœuds. Les groupes de liens ont l'avantage d'être intrinsèquement temporels comparé aux groupes de nœuds. Pour les trouver, nous construisons tout d'abord un ensemble de groupes candidats calculés par la méthode de Louvain appliquée sur une projection du flot de liens en un graphe statique. Puis nous gardons uniquement les groupes qui passent notre critère de pertinence. Nous appliquons notre méthodes sur plusieurs jeux de données d'interactions réelles afin de valider nos résultats.

2 Méthode de détection de groupes pertinents

Un flot de liens est défini comme un triplet : $\mathcal{L} = (T,V,E)$, où $T = [\alpha,\omega]$ est un intervalle de temps, V un ensemble de nœuds et $E \subseteq T \times T \times V \times V$ un ensemble de liens. Les liens de E sont des quadru-

†This research was supported by a DGA-MRIS scholarship, by a grant from the French program "PIA Usages, services et contenus innovants" under grant number 018062-44430 and by the CODDDE project ANR-13-CORD-0017-01.

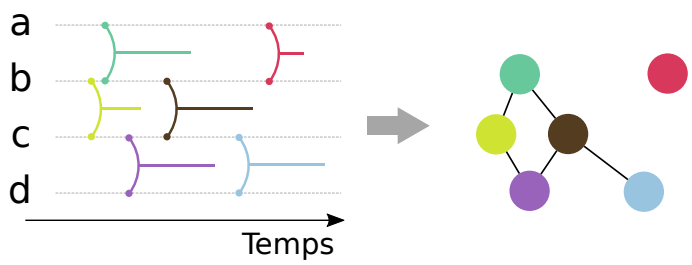


FIGURE 1: Transformation d'un flot de liens avec 4 nœuds (a-d) et 6 liens à gauche en un graphe à droite à 6 nœuds. La couleur d' un nœuds dans le graphe indique le lien du flot qu'il représente.

plets (b,e,u,v), signifiant que la paire de nœuds (u,v) ‡ est connectée sur l'intervalle $[b,e] \subseteq [\alpha,\omega]$. Nous considérons les flots de liens où pour tout $(b,e,u,v) \in E$ et $(b',e',u,v) \in E$, $[b,e] \cap [b',e'] = \emptyset$.

Nous adaptons la définition de la densité dans les flots de liens afin de traiter des liens ayant une durée au lieu de liens instantanés. La densité est alors la probabilité que deux nœuds pris aléatoirement soient reliés par un lien à un instant t tiré dans un intervalle défini par un temps et une durée. Formellement, la densité d'un ensemble de nœuds $V' \subseteq V$ à l'instant t pour une durée δ, est définie de la manière suivante :

$$d(V',t,\delta) = \frac{1}{|V'| \cdot (|V'|-1)} \cdot \sum_{u,v \in V', u \neq v} \frac{\tau_{t,\delta}(u,v)}{\delta}, \tag{1}$$

où $|V'| \geq 2$, $\delta > 0$ et $\tau_{t,\delta}(u,v) = \sum_{(b,e,u,v) \in E} |[b,e] \cap [t,t+\delta]|$ est la somme des durées des liens entre u et v dans l'intervalle $[t,t+\delta]$. Cette formule permet de calculer la densité d'un groupe de nœuds. La densité d'un groupe de liens $E' \subseteq E$ est $d(V_{E'}, t_{min}(E'), d_{E'})$, la densité des nœuds induits où $V_{E'} = \{u, \exists (b,e,u,v) \in E'\}$ à l'instant $t_{min}(E') = min_{(b,e,u,v) \in E'}(b)$, sur la durée du groupe $d_{E'} = max_{(b,e,u,v) \in E'}(e) - t_{min}(E')$.

2.1 Calcul des groupes candidats

Nous proposons une transformation du flot de liens en un graphe non dirigé et non pondéré. Chaque lien du flot est représenté par un nœud dans le graphe. Deux liens (b,e,u,v) et (b',e',u',v') sont connectés dans le graphe s'ils partagent un nœud, *i.e.* $\{u,v\} \cap \{u',v'\} \neq \emptyset$, et si les intervalles s'intersectent, *i.e.* $[b,e] \cap [b',e'] \neq \emptyset$, voir figure 1. Ainsi, un lien dans le graphe représente une connexion structurelle et temporelle entre deux liens du flot de liens. Les groupes denses dans le graphe représentent donc des groupes de liens connectés temporellement et topologiquement dans le flot.

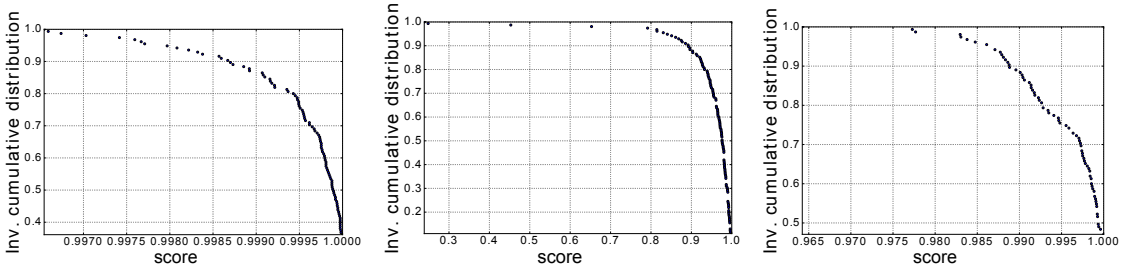
Afin de les détecter, nous utilisons sur le graphe la méthode de Louvain [BGLL08], qui est une méthode de détection de communautés. Nous obtenons ainsi une partition des nœuds du graphe et donc une partition des liens du flot. Les groupes sont trouvés en optimisant la modularité et sont par conséquent denses mais ils ne sont pas évalués sur leurs pertinences par rapport à leurs voisinages. Une méthode d'évaluation des groupes est donc nécessaire afin de garder uniquement les groupes de liens pertinents.

2.2 Sélection des groupes pertinents

La densité est un bon indicateur pour évaluer un groupe mais n'est pas suffisante, car il faut une référence à laquelle se comparer. Nous proposons de comparer la densité d'un groupe de liens avec celle de son voisinage. La densité définie par l'équation 1 dépends de trois facteurs distincts : l'ensemble de nœuds, le temps de début et la durée. Nous définissons trois voisinages d'un groupe en faisant varier chacun de ces facteurs indépendamment. Soit un groupe de liens $E' \subseteq E$ que l'on souhaite évaluer. Pour le voisinage au temps de début, les temps de début possibles sont ceux dans l'intervalle $[\alpha, \omega - \delta]$. Les densités associées sont définies par : $d(V_{E'}, x, d_{E'})$ où $x \in [\alpha, \omega - \delta]$, c'est-à-dire la densité des même nœuds induits sur la même durée que E' mais à un instant de début quelconque.

Pour le voisinage à la durée, les durées possibles sont celles dans l'intervalle $[0.8\, g_{min}, 1.2\, g_{max}]$ où g_{min} (resp. g_{max}), est la plus petite (resp. grande) durée d'un groupe dans la partition. Les densités associées

‡. Nous considérons un flot non orienté, *i.e.* $(b,e,u,v) = (b,e,v,u)$, et sans boucle, *i.e.* $u \neq v$



(a) Score du voisinage au temps de début. (b) Score du voisinage à la durée. (c) Score du voisinage au groupe de nœuds.

FIGURE 2: Distribution cumulative inverse des scores pour chaque voisinage pour Socio Pattern.

sont alors définies par : $d(V_{E'}, t_{min}(E'), x)$ où $x \in [0.8\, g_{min}, 1.2\, g_{max}]$. Utiliser l'intervalle $[0.8\, g_{min}, 1.2\, g_{max}]$ au lieu de $[0, \omega - t_{min}(E')]$ permet d'éviter les cas limites peu informatifs notamment lorsque la durée est proches de zéro ou bien proche de $\omega - \alpha$.

Pour le voisinage de l'ensemble de nœuds, il n'est pas possible de tester l'ensemble des sous-ensemble de nœuds. C'est pourquoi, nous nous limitons aux groupes de nœuds de taille $|V_{E'}|$ et qui partagent $k-1$ nœuds avec $V_{E'}$. Ainsi nous évaluons $k(|V| - k)$ groupes de nœuds.

Nous obtenons pour chaque voisinage un ensemble de valeurs de densité auquel comparer un groupe. Si le groupe est réellement pertinent, alors il devrait être plus dense que les groupes de ses voisinages. Pour quantifier cela, nous utilisons un score qui est la proportion de groupe dans le voisinage ayant une densité plus faible que le groupe en question. Plus cette proportion est importante, plus le groupe est pertinent. Une évaluation est donc composée de trois scores compris entre 0 et 1, un pour chaque voisinage, et un groupe est jugé pertinent si chaque score est plus élevé qu'un seuil déterminé. Les seuils peuvent être différents pour chaque voisinage et dépendent des caractéristiques souhaitées et du jeu de données.

3 Résultats

Nous avons étudié plusieurs réseaux d'interactions et nous présentons les résultats uniquement pour un réseau provenant de Socio Pattern [FB14] [§]. Ce réseau regroupe les interactions de 180 étudiant dans une classe préparatoire sur 9 jours. Les interactions sont capturées par des appareils bluetooth lorsque deux individus sont à moins de 1.5 mètre. Au cours de ces 9 jours, 19774 interactions ont eu lieu. De plus, nous connaissons la classe de chaque élève ainsi que les moments où ont lieu les pauses et le déjeuner.

Pour ce jeu de donnée en appliquant la méthode de Louvain, nous obtenons 155 groupes de liens de plus de 10 liens [¶]. Ces groupes durent en moyenne 7 minutes et concernent en moyenne 22.6 personnes. Les distributions du nombre de liens, de nœuds et de durées des groupes sont hétérogènes et semblent suivre des lois de puissances.

Nous appliquons sur ces groupes le processus d'évaluation et de filtrage. Les distributions des scores pour chaque voisinage sont sur les figures 2(a-c). On remarque qu'en majorité, les groupes obtenus ont des scores très élevés pour chaque voisinage. Pour le temps de début, ces scores élevés sont dûs en partie aux temps de début possibles dans le voisinage qui considère notamment les nuits où aucun lien n'apparait. C'est pourquoi lors du choix des seuils, il est nécessaire d'avoir une connaissance du jeu de données. Les seuils sont choisis en observant les fortes décroissances sur les distributions des scores et sont respectivement de 0.98, 0.998 et 0.85 pour le voisinage de l'ensemble de nœuds, le voisinage à la durée et le voisinage au temps de début. Avec ces seuils proche de 1, nous assurons que les groupes de liens capturés sont plus denses que la grande majorité des groupes dans leurs voisinages. Nous avons par ailleurs vérifié que les scores obtenus n'étaient pas redondants entre eux en observant que chaque voisinage pouvait invalider différent groupes.

§. Les résultats sont similaires pour les autres jeux données.

¶. Les groupes de moins de 10, bien qu'ils soient extrêmement nombreux, ont été exclus car nous ne les considérons pas comme pertinent.

Nous avons étudié manuellement quelques groupes. Nous avons pu observer des groupes capturés étant des regroupements d'élèves de la même classe ayant lieu avant le premier cours de la journée ou lors d'une pause, soit des élèves pendant le déjeuner.

3.1 Caractéristiques des groupes capturés

Nous étudions également le recouvrement topologique et temporel des groupes capturés. Nous observons que chaque nœud est au moins dans un groupe et en moyenne dans 7 groupes pour Socio Pattern. La structure est donc très recouvrante sur les nœuds. Au niveau temporel, les groupes capturés sont à contrario très dispersés avec la majorité du temps aucun groupe présent, notamment à cause de la nuit. Cependant lors de certains instants, *e.g.* les repas, plusieurs groupes peuvent être présents en même temps.

Les groupes capturés forment donc une structure très différente de ce qui peut être trouvé par l'état de l'art. Les méthodes utilisant des séquences de graphes ne peuvent trouver des groupes ayant des durées aussi diverses. Quant aux autres méthodes [RTG14, ELS15], elles ne peuvent trouver qu'un unique groupe de nœuds sur un ou plusieurs intervalles.

4 Conclusion et perspectives

Dans cet article, nous étudions des flots de liens afin d'y trouver des sous-flots qui sont jugé pertinents. Ces sous-flots, c'est-à-dire des sous-ensembles de liens, sont trouvés grâce à la méthode de Louvain appliquée sur une projection du flot sur un graphe statique. Un sous-flot est jugé pertinent s'il est plus dense que les groupes qui lui sont proches temporellement et topologiquement. Nous appliquons cette méthode sur plusieurs jeux de données d'interactions réelles. Nous capturons des groupes pertinents dans chaque jeux de données et mettons en avant leurs structures temporelle et topologique qui sont intéressantes. Ces résultats aident lors de l'étude d'interactions social à cibler des zones importantes.

À terme, il serait intéressant d'avoir ou de générer des jeux de tests ayant une vérité de terrain plus précise afin d'avoir une évaluation quantitative des résultats obtenus.

Références

[BBC+15] Oana Denisa Balalau, Francesco Bonchi, T-H. Hubert Chan, Francesco Gullo, and Mauro Sozio. Finding Subgraphs with Maximum Total Density and Limited Overlap. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, pages 379–388, New York, New York, USA, feb 2015. ACM Press.

[BGLL08] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment*, 2008(10) :P10008, oct 2008.

[ELS15] Alessandro Epasto, Silvio Lattanzi, and Mauro Sozio. Efficient Densest Subgraph Computation in Evolving Graphs. In *Proceedings of the 24th International Conference on World Wide Web*, pages 300–310. International World Wide Web Conferences Steering Committee, may 2015.

[FB14] Julie Fournet and Alain Barrat. Contact patterns among high school students. *PloS one*, 9(9) :e107878, jan 2014.

[HKW14] Tanja Hartmann, Andrea Kappes, and Dorothea Wagner. Clustering evolving networks. *arXiv preprint arXiv :1401.3516*, 2014.

[RTG14] Polina Rozenshtein, Nikolaj Tatti, and Aristides Gionis. Discovering Dynamic Communities in Interaction Networks. In Toon Calders, Floriana Esposito, Eyke Hüllermeier, and Rosa Meo, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8725 of *Lecture Notes in Computer Science*, pages 678–693. Springer Berlin Heidelberg, 2014.

[VL14] Jordan Viard and Matthieu Latapy. Identifying roles in an IP network with temporal and structural density. In *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 801–806. IEEE, apr 2014.