

Liens proches dans les réseaux sociaux

La dynamique des commentaires de Flickr

Stéphane Raux^{1,2} et Christophe Prieur²

¹LIP6, Université Pierre et Marie Curie, 104 avenue du Président Kennedy, 75016 Paris

²LIAFA, Université Paris-Diderot, 175 rue du Chevaleret, Case 7014, 75205 Paris Cedex 13

Nous analysons les dynamiques du réseau des commentaires du site Flickr à travers l'évolution de la structure globale du réseau, mais aussi celle des voisinages locaux et la manière dont ils se constituent. Nous montrons que l'essentiel de l'activité concerne des « liens proches », et nous analysons les configurations locales pour montrer les répercussions de ces liens sur la constitution du voisinage : les individus qui ont le plus de voisins choisissent essentiellement ceux-ci parmi les voisins de leurs voisins (dans une proportion bien plus grande que dans le cas d'un choix aléatoire).

Keywords: Complex networks, social networks, graph dynamics, Flickr, ego-centered networks, web 2.0.

Introduction. L'analyse des réseaux sociaux a longtemps été un courant de recherche principalement associé aux sciences sociales [Sco92] avant de connaître un regain d'intérêt dans la décennie précédente : le développement du World-Wide Web et la possibilité de disposer de très grands jeux de données ont alors permis la création d'un nouveau champ de recherche à partir de la découverte des propriétés que partagent un grand nombre de grands réseaux [WS98] qui sont issus de domaines aussi variés que la biologie, l'économie ou la sociologie. Depuis, le succès de nombreuses plateformes web qui tirent parti des interactions de leurs utilisateurs pour leur proposer des services témoigne de la popularité de ces réseaux dits « petit monde » auprès du grand public. Paradoxalement, ce « nouveau » web favorise l'émergence d'un individualisme exacerbé : il s'agit de construire et d'organiser son propre réseau à travers par exemple l'animation d'un blog ou la gestion de ses contacts. Cette question du réseau centré sur l'individu rejoint un courant de l'analyse des réseaux sociaux qui s'attache à décrire les configurations de ces réseaux personnels [Wel07].

A partir du réseau des commentaires du site Flickr.com, nous étudions la manière dont les individus entrent en contact en considérant à la fois l'ensemble du graphe et les configurations des voisinages de chaque sommet. Nous mesurons la distance qui sépare chaque individu de son nouveau voisin lors de l'établissement d'une relation : nous distinguons les voisins « proches », qui ont un contact commun avant l'établissement de la relation, et les voisins « lointains » dans le cas contraire. Cette démarche peut être vue comme une extension dynamique de la notion de coefficient de clustering. Nous nous démarquons ainsi d'autres études qui portent sur Flickr, qu'il s'agisse de mesurer l'évolution de la structure globale du réseau [KNT06] ou les phénomènes de diffusion [CMG09]. Nous travaillons sur les commentaires qui ont été déposés sur les photographies de Flickr entre mars 2004 et août 2006 [BCPP09] : chaque relation signifie qu'un utilisateur a déposé un commentaire sur la photographie d'un autre utilisateur. Nous disposons d'une base de 39 594 157 commentaires datés qui ont été rédigés par 910 454 utilisateurs[†]. De manière plus générale, notre approche fournit une entrée originale pour appréhender la structure et l'évolution de grands réseaux de communication.

Formalisation. Pour prendre en compte la dimension dynamique de nos données, on définit un intervalle de temps discret T (exprimé en pratique en secondes), et pour tout $t \in T$ le graphe non orienté $G_t = (V, E_t)$, où V est l'ensemble des sommets et E_t l'ensemble des relations qui « existent » à l'instant t . Nous considérerons que les commentaires sont cumulatifs : un lien entre u et v est considéré comme existant à l'instant t si les deux sommets ont déjà échangé au moins un commentaire à un instant $t' \leq t$.

[†] Les résultats présentés ici sont détaillés dans [RP09], nous y renvoyons pour une bibliographie plus développée.

Nous notons $N_t(u)$ le *voisinage* d'un sommet u dans G_t et $\deg_t(u)$ son *degré*. La *distance* entre deux sommets u et v , notée $\text{dist}_t(u, v)$ est la longueur du plus court chemin entre u et v dans G_t . S'il n'existe pas de chemin, on notera $\text{dist}_t(u, v) = \infty$. Le sommet v est un *voisin proche* de u dans G_t s'il existe un instant $t' \leq t$ tel que $\text{dist}_{t'-1}(u, v) = 2$ et $\text{dist}_{t'}(u, v) = 1$, et un *voisin lointain* sinon. Nous notons $\tilde{N}_t(u)$ l'ensemble des voisins proches de u , que nous appelons *voisinage proche* de u . On notera $P_t(u)$ la *proportion de voisins proches* de u : $P_t(u) = \frac{|\tilde{N}_t(u)|}{|N_t(u)|}$.

1 L'importance des « liens courts »

Une première approche consiste à mesurer pour l'ensemble des commentaires la proportion de liens répétés et de nouvelles relations. Nous avons choisi par souci de simplification combinatoire de travailler sur un graphe non orienté. Nous montrons dans [RP09] que cela n'introduit pas de biais important : on observe une forte corrélation entre le nombre d'arêtes entrantes et sortantes, en particulier pour les sommets de fort degré, et les relations réciproques, qui représentent presque deux tiers des commentaires (65,2%), constituent une part importante de la structure globale du réseau. Nous parcourons pour cela la liste des commentaires et nous mesurons lors de chaque relation entre u et v la distance qui les séparait juste avant. Nous distinguons trois situations : la répétition d'un commentaire ($\text{dist}_{t-1}(u, v) = 1$), les nouvelles relations avec un voisin proche ($\text{dist}_{t-1}(u, v) = 2$) et les nouvelles relations avec un voisin lointain ($\text{dist}_{t-1}(u, v) \geq 3$). Nous considérons en effet que les membres d'un réseau ont une vision réduite de leur entourage. Ils peuvent avoir conscience d'une partie des personnes qui se situent dans leur entourage à distance 2, mais de leur point de vue il n'y a pas de différence sensible entre des personnes situées à des distances de 3, 4 voire appartenant à une autre composante connexe : il s'agit dans tous les cas de personnes avec lesquelles ils ne partagent aucune connaissance commune.

Kumar *et al.* [KNT06] ont identifié trois phases successives en étudiant l'évolution de la densité du réseau de l'ensemble des utilisateurs de Flickr : d'abord une forte augmentation, puis une baisse et enfin une augmentation soutenue pour le reste de la période.

Nous retrouvons ces trois phases pour les mêmes périodes sur la figure 1, qui représente l'évolution de la proportion de contacts avec des voisins proches parmi l'ensemble des nouveaux contacts. Comme l'apparition d'un nouveau contact entre deux sommets à distance 2 entraîne la création d'un nouveau triangle, la proportion de nouveaux voisins proches est un bon indicateur de la densité locale d'un graphe. Après une phase de mise en place, l'arrivée massive de nouveaux utilisateurs en quelques semaines (de 3 317 à 24 358) entraîne une baisse de la densité dans la deuxième phase, puis la proportion de nouveaux voisins proches croît et atteint 70% à la fin de la période.

On observe aussi que les liens répétés et les nouvelles relations avec des voisins proches représentent respectivement 75,6% et 17,2% de l'ensemble des commentaires. Cela signifie que les commentaires sont échangés dans près de 93% des cas entre des utilisateurs qui sont déjà voisins ou qui ont au moins un voisin en commun : une grande majorité de l'activité du réseau s'effectue sur de très courtes distances. Si les fortes densités locales des réseaux « petit monde » sont un phénomène bien connu, la part très importante de relations répétées nous montre que les utilisateurs laissent en pratique peu de place à la rencontre de parfaits inconnus.

2 Des configurations locales contrastées

Pour comprendre comment les individus construisent et organisent leur « réseau » de contacts, nous mesurons pour chaque sommet du graphe le degré et la proportion de ses voisins proches à la fin de la période observée, que nous notons $P(u)$. La figure 2 indique la distribution des degrés qui est, sans surprise, très hétérogène : on observe un très grand nombre de sommets de très faible degré (438 840 sommets de

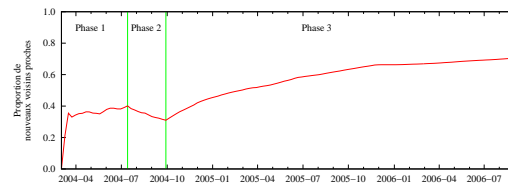


FIG. 1: Évolution de la proportion de voisins proches par rapport à l'ensemble des nouveaux voisins.

degré 1) côtoyant quelques sommets avec un degré très élevé (le degré maximum est de 8 731).

Comme un sommet ne peut pas avoir de voisin proche s'il ne possède pas déjà un voisin pour servir d'intermédiaire, la distribution de la proportion de voisins proches ne concerne que les sommets qui ont au minimum 2 voisins. Cette distribution est difficile à apprécier car elle est en partie biaisée par la distribution des valeurs possibles. Par exemple, les sommets de degré 2 ne peuvent avoir que deux valeurs pour $P(u)$, qui sont 0% s'ils n'ont aucun voisin proche et 50% s'ils en ont un. Comme les sommets de faible degré sont les plus nombreux, cela conduit à une forte proportion de sommets pour lesquels $P(u)$ prend une valeur de 0% ou de 50%.

Si l'on écarte les valeurs liées à ces effets de mesure, on constate que les distributions de $P(u)$ sont homogènes si on regroupe les sommets en fonction de leur degré. La figure 3 montre que plus le degré est important, plus les valeurs moyennes de $P(u)$ sont élevées : près de 90% des sommets dont le degré est supérieur ou égal à 750 ont une proportion de voisins proches supérieure à 80%.

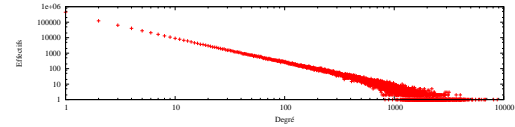


FIG. 2: Distribution des degrés pour l'ensemble des commentaires.

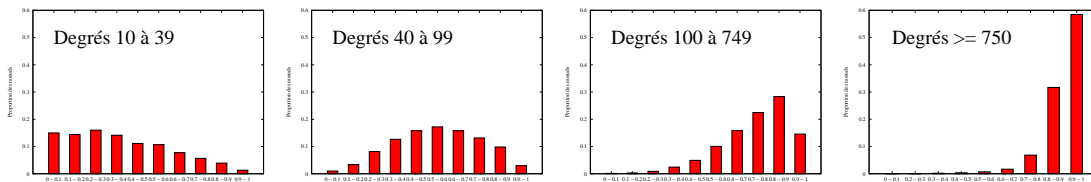


FIG. 3: Distributions des proportions de voisins proches regroupées par classes de degrés. Les abscisses correspondent aux valeurs de $P(u)$ et les ordonnées aux effectifs.

La taille du voisinage à distance 2 augmente très fortement en fonction du degré : on peut supposer que les sommets de fort degré ont simplement plus de chances d'entrer en contact avec un voisin proche mais en pratique, la proportion de voisins proches pour les sommets de fort degré est largement supérieure à celle que l'on obtiendrait pour un choix aléatoire. À titre indicatif, pour chacun des sommets de degré supérieur à 6 000, la proportion de voisins à distance 2 représente plus d'un quart de l'ensemble des sommets alors que la proportion de voisins proches dépasse les 90%.

Ces résultats peuvent bien sûr s'expliquer par les modèles d'attachement préférentiel : les sommets de fort degré jouent un rôle important dans la structuration du « petit monde » en constituant autour d'eux l'essentiel des zones denses, mais les fortes valeurs de $P(u)$ soulignent l'importance de la transitivité dans la construction de leur entourage.

3 Structure et évolution des voisinages locaux

On peut alors imaginer que les variations de la proportion de voisins proches pourraient refléter différentes pratiques dans la construction du réseau personnel : nous comparons l'évolution des liens réciproques de l'entourage d'un sommet qui privilégie les contacts lointains et d'un autre qui privilégie les contacts proches. Pour que la comparaison soit pertinente, nous ne retenons dans le graphe que les sommets qui entretiennent des relations réciproques et nous choisissons deux sommets de degré 80 (la distribution des $P(u)$ pour ce degré est relativement symétrique, voir la figure 3) qui sont tous les deux actifs à partir de fin mars 2006. La figure 4 a été obtenue en parcourant la liste des commentaires et en mesurant pour chaque sommet l'évolution de son entourage ($N_i(u)$ et $\tilde{N}_i(u)$), à intervalles d'une semaine. La première courbe montre l'évolution de la taille des voisinages et la deuxième l'évolution du nombre de nouveaux voisins apparus au cours de la dernière semaine de mesure.

Dans le cas du sommet *A*, qui a une faible proportion de voisins proches ($P(A) = 32,5\%$), le degré augmente par à-coups, avec des paliers successifs. Le sommet *B* favorise au contraire les contacts avec son voisinage proche ($P(B) = 77,5\%$) : son voisinage augmente de façon régulière tout au long de la période. Le

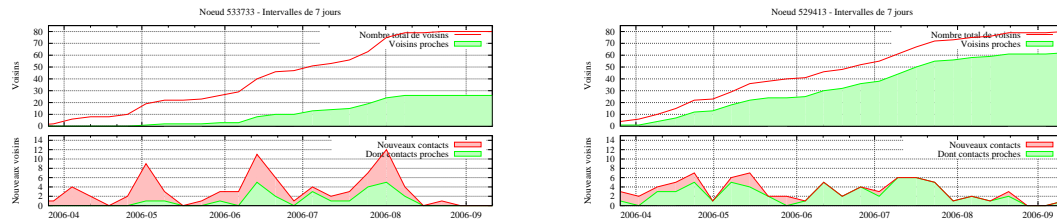


FIG. 4: Évolution de l'entourage des sommets A (à gauche) et B (à droite).

sommet B a une activité plus régulière et plus intense que le sommet A . Le sommet A s'investit moins dans le service, sauf lors de pics d'activité qui peuvent être interprétés comme des périodes d'intérêt ponctuel pendant lesquelles il entre en contact avec de nouveaux voisins.

Ces deux exemples ne sont bien sûr pas représentatifs de la variété des situations que l'on peut observer, mais la multiplication d'observations similaires nous a permis de constater que si les profils d'évolution des sommets à faible $P(u)$ sont très variés, puisqu'ils dépendent d'événements souvent indépendants de la structure du réseau, l'évolution des sommets à $P(u)$ élevé est souvent régulière, en particulier lorsqu'il s'agit de sommets de fort degré. Nos résultats rejoignent sur ce point ceux de [CMG09], qui observent l'évolution du nombre d'utilisateurs ajoutant une photographie parmi leurs « favoris ».

Nous approfondissons par ailleurs ces résultats en mesurant la concentration de l'activité des individus, c'est à dire le nombre de jours pendant lesquels un individu a émis au moins un commentaire par rapport à l'ensemble des jours écoulés entre son premier et son dernier commentaire. Sans nous attarder sur les 62% de relations qui ne sont pas renouvelées au-delà d'une journée, nous observons que la distribution des taux de concentration est très comparable à celle des proportions de voisins proches : plus de 80% des sommets de degré supérieur à 750 ont une concentration de leur activité supérieure à 70%. Les utilisateurs qui parviennent à entretenir un nombre élevé de relations ne sont donc pas ceux qui sont présents sur le service depuis le plus longtemps, mais ceux qui l'utilisent de manière la plus intensive et la plus continue dans le temps, en privilégiant les contacts avec des voisins proches, qui s'établissent plus facilement.

En conclusion, la variété des pratiques des utilisateurs et du rythme de leur activité se répercute sur la structure de leur voisinage local et sur celle de l'ensemble du réseau. Le caractère exceptionnel des relations avec des sommets lointains permet de nuancer la vision idéalisée d'un « petit monde » qui pourrait parfois s'apparenter plutôt à un « monde étriqué ». On pourrait affiner la compréhension de ces mécanismes en prenant en compte l'orientation des relations, ce qui permettrait de déterminer si les fortes proportions de $P(u)$ pour les sommets de fort degré sont fruit de leur activité ou au contraire de celle de leur entourage. Il reste enfin à étudier comment les outils que nous avons décrits peuvent s'appliquer à d'autres types de réseaux de communication.

Références

- [BCPP09] Jean-Samuel Beuscart, Dominique Cardon, Nicolas Pissard, and Christophe Prieur. Pourquoi partager mes photos de vacances avec des inconnus ? Les usages de flickr. *Réseaux*, 154(27), 2009.
- [CMG09] Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th WWW Conference (WWW'09)*, 2009.
- [KNT06] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *KDD '06 : Proceedings of the 12th ACM SIGKDD international conference*, 2006.
- [RP09] Stéphane Raux and Christophe Prieur. Le poids des liens proches – étude de la dynamique d'un grand réseau social. (soumis), <http://hal.archives-ouvertes.fr/hal-00359463/en/>, 2009.
- [Sco92] John Scott. *Social Network Analysis*. Sage, London, 1992.
- [Wel07] Barry Wellman. The network is personal : Introduction to a special issue of social networks. *Social Networks*, 29(3), 2007.
- [WS98] Duncan Watts and Steve Strogatz. Collective dynamics of small-world networks. *Nature*, 393, 1998.