Modèles pour les topologies réalistes

Jean-Loup Guillaume and Matthieu Latapy

LIAFA - Université Paris 7, 2, place Jussieu, 75005 Paris, France. `(guillaume,latapy)@liafa.jussieu.fr` `Matthieu.Latapy@inria.fr`

Résumé : Il est apparu récemment que l'immense majorité des grands graphes rencontrés en pratique ont des propriétés bien particulières. Ceci a conduit la communauté scientifique à essayer d'identifier ces particularités et à les expliquer par l'intermédiaire de modèles de génération de graphes aléatoires ayant certaines propriétés. Aujourd'hui, un grand nombre de tels modèles ont été introduits, suivant des approches variées. Nous présentons ici les principales d'entre elles, et discutons leurs mérites respectifs.

1 Contexte

De nombreux grands graphes jouent un rôle important dans divers contextes. Citons par exemple les graphes issus des liens hypertextes entre pages (ou sites) Web, les graphes d'interconnexion de routeurs (ou d'AS) sur Internet, les graphes d'appels téléphoniques ou d'échanges de courriers électroniques entre personnes, les graphes d'auteurs co-signant des articles, les graphes de citations dans des articles, les graphes d'interactions métaboliques dans des organismes vivants, les graphes d'interconnexions des neurones dans des systèmes nerveux, les graphes de collaboration entre acteurs ou entre dirigeants d'entreprises, les graphes de contacts sexuels, les graphes des réseaux de distribution de l'électricité, les graphes des échanges sur les réseaux pair-à-pair, etc.

En dehors du fait que ces graphes sont peu denses (ou, de façon équivalente, ont un degré moyen très faible), ils n'ont *a priori* aucune raison de partager des propriétés spécifiques. Considérons en particulier les trois notions suivantes sur un graphe $G = (V,E)$:

- La *distance* entre deux sommets u et v de V est la longueur minimale d'un chemin les reliant. La *distance moyenne* de G est la moyenne pour tout couple de sommets de la distance entre eux.
- Le *coefficient de clustering* d'un sommet v de G est la probabilité, étant donnés deux voisins v' et v'' de v, qu'ils soient voisins entre eux. Le *coefficient de clustering* du graphe G est la moyenne du coefficient de clustering de tous ses sommets.
- La *distribution des degrés* de G associe à chaque entier k le nombre de sommets de G ayant degré k. Cette distribution suit une *loi de puissance* si elle est proportionnelle à $k^{-\gamma}$ pour un certain réel γ appelé *exposant* de la loi. Elle suit une *loi de Poisson* si elle est proportionnelle à e^{-k}.

Il a été constaté récemment [3, 8, 22] que dans tous les cas que nous avons cités, et bien d'autres, la *distance moyenne* est faible, que le *coefficient de clustering* est fort, et que la distribution des degrés suit une *loi de puissance*. En d'autres termes, il existe généralement des chemins courts entre les sommets, les voisins d'un sommet donné ont de fortes chances d'être voisins entre eux, et il y a une grande majorité de sommets de faible degré et un nombre faible mais *non négligeable* de sommets de très haut degré.

Comme ces propriétés sont vérifiées par tous les graphes que nous avons cités et de nombreux autres cas pratiques, on peut se demander si elles nous apprennent vraiment quelque chose sur les graphes en question, ou si elles ne sont pas tout simplement le reflet du fait que ce sont des graphes quelconques. En d'autres termes, ces graphes ne sont peut être pas différents de ceux qu'on obtient en mettant des liens au hasard entre des sommets.

Le modèle classique de graphes aléatoires introduit en 1959 par Erdös et Rényi [10] est défini comme suit : étant donnés un entier n et une probab'eilité p, un graphe $G_{n,p}$ est un graphe à n sommet dans lequel chaque arête existe avec probabilité p. Il a donc en moyenne $p \cdot \binom{n}{2}$ arêtes. Ce modèle a été très étudié [7]. Parmi ses propriétés connues, citons le fait que la distance moyenne est courte, ce qui est conforme avec les observations faites sur les graphes que nous avons cités.

Toutefois, la probabilité que deux voisins d'un sommet donné soient voisins entre eux n'est pas plus forte que pour n'importe quelle paire de sommets. Par conséquent, le coefficient de clustering d'un graphe $G_{n,p}$ est égal à p, qui vaut de l'ordre de $\frac{1}{n}$ puisque le graphe est très peu dense. Il est donc très faible. De plus, comme tous les sommets jouent un rôle similaire, les degrés sont concentrés autour de la valeur moyenne : ils suivent une loi de Poisson. Par conséquent, contrairement à ce qu'on a longtemps pensé, ce modèle de graphes ne correspond pas à la réalité des cas rencontrés en pratique.

Cette constatation récente a donné naissance à un grand nombre de travaux. Un des principaux défis posés se formule comme suit : comment générer des graphes ayant les propriétés rencontrées en pratique ? Les objectifs sont multiples : compréhension des phénomènes sous-jacents, construction de topologies réalistes pour la simulation, etc. Plusieurs approches ont été utilisées, suivant les objectifs. Dans certains contextes, le but est de mimer la réalité de certains cas particuliers de façon aussi précise que possible [13, 16]. La multiplication des paramètres rend toutefois l'analyse de ces modèles extrêmement difficile. Nous nous concentrerons ici sur les modèles généraux, basés sur des processus de constructions relativement simples (permettant en particulier certaines analyses formelles) et réalistes (c'est-à-dire s'inspirant de processus effectivement rencontrés en pratique). Certains de ces travaux ce concentrent sur le fort clustering des topologies réalistes, alors que d'autres accordent plus d'importance à la distribution des degrés. Nous présenterons successivement ces deux familles de modèles. Pour des synthèses plus approfondies, mais moins récentes, nous nous référons à [3, 8, 22].

2 Clustering

Une grande caractéristique des graphes rencontrés en pratique est leur structuration en *communautés* : alors que le graphe est peu dense, il est généralement composé de sous-graphes très denses reliés entre eux par quelques arêtes. Un certain nombre de modèles basés sur cette observation ont été introduits. Nous présentons dans cette section deux d'entre eux, représentatifs des propositions faites dans cette direction. Soulignons qu'ils ne produisent pas des graphes dont les degrés suivent une loi de puissance.

2.1 Localité et liens longue distance

Watts et Strogatz ont proposé en 1998 un modèle défini comme suit [23] : on considère un anneau de sommets, chacun étant relié à ses k plus proches voisins sur l'anneau pour un entier k donné. Puis on *recâble* chaque arête avec une probabilité p donnée, c'est-à-dire qu'une extrémité de chaque arête est remplacée, avec probabilité p, par une nouvelle extrémité choisie aléatoirement.

Lorsque $p = 0$ (on ne recâble pas d'arêtes), le graphe a un fort coefficient de clustering mais la distance moyenne est grande. Par contre, si $p = 1$, on a recâblé toutes les arêtes, et donc ce qu'on obtient n'est rien d'autre qu'un graphe aléatoire comme défini par Erdös et Rényi. Il a donc bien une distance moyenne faible mais également un clustering faible. Pour des valeurs de p intermédiaires, toutefois, les distances moyennes sont courtes *et* le clustering est fort, comme dans les cas réels.

Intuitivement, un graphe obtenu par ce modèle rend compte du fait que chaque personne connaît les personnes proches de lui (géographiquement, par exemple) et *quelques* personnes éloignées. D'autres idées du même type ont été étudiées [6, 15] sous un angle différent : ces travaux montrent que, non seulement la distance moyenne est courte, mais qu'en plus il existe des stratégies *locales*, sous certaines conditions, pour effectivement trouver des chemins courts entre tous couples de sommets.

2.2 Ajouts de cliques

On peut remarquer qu'un certain nombre des graphes que nous avons cités ci-dessus sont obtenus en superposant successivement des cliques (Figure 1). C'est par exemple le cas du graphe des acteurs apparaissant dans un même film, ou du graphe des scientifiques ayant co-signé un article (suivant le cas, chaque film ou chaque article correspond à une clique). Dans les cas rencontrés en pratique, les tailles des cliques suivent une loi de puissance.

Il a été montré [12] que la distribution de la taille des cliques et leur recouvrement moyen (moyenne pour toutes les cliques de la proportion de chacune contenue dans d'autres cliques) suffisent à expliquer la faible distance moyenne et le fort clustering de ces graphes. De plus, on peut généralement décomposer les autres

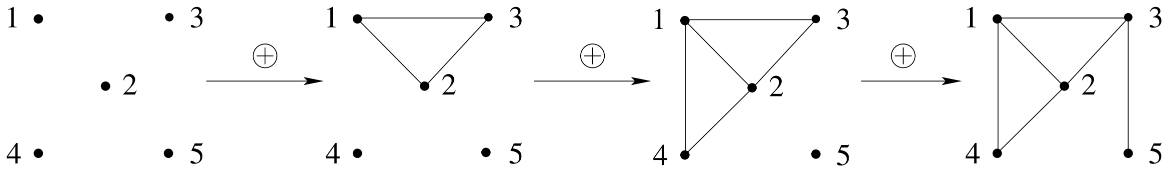

FIG. 1: *Obtention d'un graphe par ajouts successifs de cliques se recouvrant partiellement. Ici, on ajoute* $\{1,2,3\}$, $\{1,2,4\}$ *puis* $\{3,5\}$.

graphes cités en cliques se recouvrant partiellement. Les propriétés de l'ensemble des graphes rencontrés en pratique peuvent alors être vues comme conséquences des propriétés de ce modèle.

3 Distributions de degrés

Le fait que les graphes rencontrés en pratique aient une distribution de degrés en loi de puissance les différencie fortement des graphes d'Erdös et Rényi : certains sommets, souvent appelés *hubs*, ont un très haut degrés. Il a été montré que ces sommets jouent un rôle central pour divers problèmes (par exemple la robustesse du réseau sous-jacent [4, 20]). C'est pourquoi un grand nombre de travaux se sont consacrés à l'étude de cette propriété spécifique. Nous présentons ci-dessous deux travaux représentatifs de deux approches bien distinctes.

3.1 Propriétés générales

Étant donnée une distribution, on peut construire un graphe dont les degrés suivent cette distribution comme suit : on duplique chaque sommet un nombre de fois tiré aléatoirement en conformité avec la distribution, puis on apparie aléatoirement les sommets dans le multi-ensemble ainsi obtenu. Ces paires forment les arêtes du graphe final, dont les degrés suivent exactement la distribution fixée au départ [1]. On peut alors étudier les propriétés attendues d'un tel graphe, qui est un tirage aléatoire uniforme parmi les graphes ayant la distribution de degrés spécifiée.

La connexité de ces graphes a été étudiée dans [17, 18, 19]. Puis Aiello, Chung et Lu [1] ont montré que, pour des distributions en loi de puissance avec exposants proches de ceux rencontrés en pratique (typiquement entre 2 et 3), ces graphes ont une distance moyenne faible. Ce résultat montre que la distance moyenne faible est une propriété naturelle pour les graphes ayant une distribution des degrés en loi de puissance, comme pour les graphes d'Erdös et Rényi.

3.2 Attachement préférentiel

L'autre approche utilisée pour l'étude des distributions de degrés des graphes rencontrés en pratique est de tenter de les reproduire par des processus de construction aussi simples et naturels que possible. Le principal modèle atteignant ces objectifs est celui d'Albert et Barabási [2, 9] : les sommets sont ajoutés un à un et reliés aléatoirement aux sommets préexistants. Toutefois, les sommets auxquels le nouveau sommet est relié sont choisis avec une probabilité qui croît en fonction de leur degré. En d'autres termes, un nouveau sommet se lie plus facilement à un sommet déja beaucoup lié qu'à un sommet peu lié. Cette procédure a reçu le nom d'*attachement préférentiel*.

L'attachement préférentiel correspond par exemple au fait qu'une personne créant une nouvelle page Web aura tendance à pointer vers une page qui est déja bien connue (donc bien référencée). En effet, si une page est peu connue, la probabilité que le créateur de la nouvelle page la connaisse est faible. De plus, les pages beaucoup référencées sont des autorités de leur domaine, et on aura donc tendance à pointer vers elles.

On obtient de cette façon des graphes dont la distribution des degrés suit une loi de puissance, comme souhaité. De plus, la distance moyenne entre les sommets est faible. Par contre, les graphes obtenus de cette façon ont un coefficient de clustering relativement faible. Ce modèle ne rend donc pas parfaitement compte des propriétés des topologies réalistes. Il propose toutefois une explication simple et naturelle à la présence de distributions des degrés en lois de puissances.

De nombreux autres modèles basés sur l'attachement préférentiel (parfois sous une forme différente) ont été introduits afin de proposer des processus de construction plus réalistes ou de permettre une analyse formelle des propriétés des graphes engendrés (voir par exemple [11]). Pour une présentation plus détaillée, voir [3, 8, 22].

4 Aller plus loin

Nous avons vu qu'un certain nombre de modèles, désormais classiques, arrivent à rendre compte du coefficient de clustering ou de la distribution des degrés en loi de puissance (tout en conservant la distance moyenne faible). Toutefois, ces deux propriétés ne coexistent pas, à moins de modifier les modèles avec des critères *ad-hoc*. L'étape suivante est donc la proposition de modèles générant des graphes ayant l'ensemble des trois propriétés citées. Des premiers travaux vont dans cette direction [5, 21, 14], mais restent encore embryonnaires et peu satisfaisants.

Une direction prometteuse semble être la combinaison de méthodes ayant déjà fait leurs preuves pour une des deux propriétés. On peut par exemple combiner le modèle par ajouts de cliques et l'attachement préférentiel en affectant aux sommets de hauts degrés une plus forte probabilité de figurer dans les recouvrements. Ceci traduirait par exemple l'intuition selon laquelle un scientifique ayant déjà co-signé avec beaucoup de collègues est susceptible de publier plus d'articles avec des co-auteurs différents.

On peut par ailleurs imaginer que les topologies réalistes partagent d'autres propriétés qui n'ont pas encore été identifiées. De plus, la plupart des modèles introduits jusqu'ici ne tiennent pas compte de l'*orientation* des arcs ou de la *dynamique* des topologies concernées. Ces deux aspects peuvent pourtant jouer un rôle important dans certains cas réels, et restent très largement à explorer.

Références

[1] W. Aiello, F. Chung and L. Lu. *A random graph model for massive graphs.* Proceedings of the Thirtysecond Annual ACM Symposium on Theory of Computing. 2000.

[2] Réka Albert and Albert-László Barabási. *Emergence of scaling in random networks.* Science 286 :509-512. 1999.

[3] Réka Albert and Albert-László Barabási. *Statistical Mechanics of Complex Networks.* Reviews of Modern Physics 74, 47. 2002.

[4] Réka Albert, Hawoong Jeong and Albert-László Barabási. *Error and attack tolerance in complex networks.* Nature 406 :378. 2000.

[5] A.L. Barabasi, Z. Dezso, E. Ravasz, S. H. Yook, and Z. Oltvai. *Scale-free and hierarchical structures in complex networks.* To appear in Sitges Proceedings on Complex Networks. 2003.

[6] Lali Barrière, Pierre Fraigniaud, Evangelos Kranakis and Danny Krizanc. *Efficient Routing in Networks with Long Range Contacts.* In 15th International Symposium on Distributed Computing (DISC). LNCS 2180, 270-284. Springer. 2001.

[7] Bela Bollobas. Random Graphs. Cambridge Studies in Advanced Mathematics, Cambridge University Press. August 2001.

[8] S.N. Dorogovtsev and J.F.F. Mendes. *Evolution of networks.* Adv. Phys. 51, 1079. 2002. cond-mat/0106144

[9] S. Dorogovtsev, J. Mendes and A. Samukhin. *Structure of growing networks with preferential linking.* Phys. Rev. Lett. 85 :4633-4636. 2000.

[10] P. Erdös and A. Rényi. *On random graphs.* Publ. Math. Debrecen 6 : 290-297. 1959.

[11] A. Fabrikant, E. Koutsoupias and C. H. Papadimitriou. *Heuristically Optimized Trade-offs : A new paradigm for power laws in the Internet.* ICALP 2002.

[12] J.-L. Guillaume, M. Latapy and Ada Suk-Fung Ng. *A realistic model for complex networks.* Submitted.

[13] E.M. Jin, M. Girvam and M.E.J. Newman. *The structure of growing social networks.* Phys. Rev. E 64, 046132. 2001.

[14] K. Klemm and V.M. Eguiluz. *Highly clustered scale-free networks.* Phys. Rev. E 65, 036123. 2002. cond-mat/0107606

[15] Jon Kleinberg. *The Small-World Phenomenon : An Algorithmic Perspective.* Proceedings of the 32nd ACM Symposium on Theory of Computing (STOC). 2000.

[16] R. Kumar, P. Raghavan, S. Rajagopalan and D. Sivakumar. *Stochastic models for the Web graph.* FOCS : IEEE Symposium on Foundations of Computer Science (FOCS).

[17] Tomasz Luczak. *Sparse random graphs with a given degree sequence.* Random Graphs, vol 2, 165-182, Wiley, New York. 1992.

[18] M. Molloy and B. Reed. *A critical point for random graphs with a given degree sequence.* Random Structures and Algorithms, Vol. 6, no. 2 and 3 161-179. 1995.

[19] M. Molloy and B. Reed. *The size of the giant component of a random graph with a given degree sequence.* Combin. Probab. Comput. 7, no., 295-305. 1998.

[20] Adilson E. Motter and Ying-Cheng Lai. *Cascade-based attacks on complex networks.* Physical Review E 66. 2002.

[21] E. Ravasz and A.-L. Barabási *Hierarchical organization in complex networks.* Physical Review E (in press). cond-mat/0206130

[22] Steven H. Strogatz. *Exploring complex networks.* Nature 410. March 2001.

[23] D. Watts and S. Strogatz. *Collective dynamics of small-world networks.* Nature 393 :440-442. 1998.