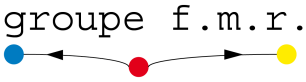
L'analyse des graphes bipartis

Laurent Beauguitte, CNRS, UMR IDEES
beauguittelaurent<at>hotmail.com

Février 2013 - Version 1



Introduction

Un graphe biparti (*bipartite graph*) permet d'étudier et de visualiser les relations entre deux ensembles distincts de sommets, d'où le terme synonyme de réseau 2 modes (*2-mode network*) [1]. Selon la nature des ensembles de sommets considérés, les anglophones distinguent les *affiliations network* (ex. appartenance d'individus à des associations) et les *actors-events network* (ex. participation d'individus à des événements) [2]. Des exemples fameux dans la littérature sociologique portent sur la présence de femmes à des événements sociaux (Davis *et al.*, 1941 [7]), la présence de dirigeants à des conseils d'administration [3] ou encore la présence d'acteurs dans des films [4]. Ces graphes sont également fréquemment utilisés en écologie pour analyser des relations insectes - plantes ou proies - prédateurs et en scientométrie (réseau auteurs - revues et auteurs - mots clés notamment).

Si le graphe biparti étudie les relations entre deux ensembles distincts, il ne prend pas en compte les relations à l'intérieur de ces deux ensembles. Par ailleurs, si les relations au sein d'un graphe biparti sont généralement non orientées [5], elles peuvent par contre être valuées. Ainsi, un auteur peut publier plusieurs fois dans une même revue.

1. Certains auteurs considèrent cependant que l'assimilation graphe biparti - graphe 2-modes est source de confusions.

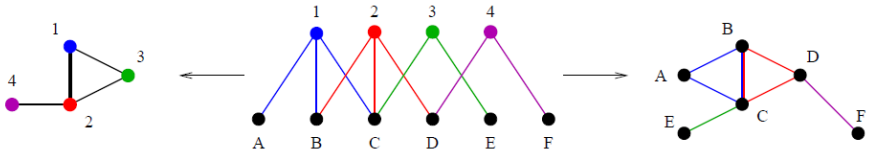
2. Faust cite également les termes de *dual networks*, de *membership networks* et d'*hypernetworks* [11].

3. Tout un champ de recherche est dédié depuis les années 70 à la recherche de ces *Corporate Interlocks*. Pour une discussion sur l'intérêt et les limites de ces études, voir notamment Mizruchi, 1996 [19].

4. *The internet movie database*, http ://www.imdb.com/

5. Dans certains cas, l'orientation serait utile : ainsi, si la plupart des participants à un colloque demandent à y assister (en envoyant un résumé, en payant les frais d'inscription), d'autres sont invités par les organisateurs.

FIGURE 1 – Du graphe biparti aux graphes de co-occurence



La transformation d'un graphe biparti - au centre - entre graphes de co-occurence revient à multiplier la matrice de départ par sa transposée ou la transposée par la matrice de départ. Il n'est pas possible de retrouver le graphe d'origine à partir des matrices de co-occurence obtenues. Le double lien entre B et C sur le graphe de droite signale qu'ils ont deux liens communs dans le graphe de départ (vers les sommets 1 et 2). Figure tirée de Allali, 2011 [1].

L'une des méthodes les plus courantes pour analyser ces graphes est de les transformer en deux graphes distincts de co-occurrence : le graphe acteur - événement est donc transformé en un graphe acteurs - acteurs (un lien entre deux acteurs indique qu'ils étaient tous deux présents à un même événement) et un graphe événement - événement (un lien entre deux événements indique que le même acteur a assisté aux deux). La figure 1 montre comment s'effectue ces transformations. Les deux graphes obtenus peuvent alors être considérés comme deux graphes valués standards. Cette approche a longtemps été critiquée dans la mesure où la transformation implique une perte importante d'informations, point de vue remis récemment en question par Everett et Borgatti [10] : les deux auteurs affirment en effet que l'analyse conjointe des deux graphes valués (acteurs - acteurs ; événements - événements) permet de trouver des résultats similaires et, dans la plupart des cas, de reconstituer le graphe de départ.

Point à souligner, certains auteurs ont affirmé que les réseaux bipartis étaient plus fiables que les réseaux « classiques » dans la mesure où la participation d'un individu à un événement pouvait être connue sans risque d'erreur, tandis que des biais importants existent quand les données portent sur des relations directes[6]. Il n'est pas certain que cet argument soit totalement pertinent : ainsi, toute personne ayant étudié la participation de chercheurs à des colloques sait que la liste des participants ne correspond jamais à la liste des présents... De plus, et il est prudent de le rappeler, co-présence ne signifie pas interaction. Que deux auteurs publient dans une même revue ou utilisent à l'occasion un même mot clé est un indicateur faible d'un quelconque lien entre ces deux auteurs.

6. *"Data on affiliation networks tend to be more reliable than those on other social networks, since membership of a group can often be determined with a precision not available when considering friendship or other types of acquaintance"*, Newman, 2001 [20].

Cette synthèse s'intéresse uniquement aux méthodes d'analyse possibles sur un graphe biparti non transformé [7] en abordant successivement les mesures possibles (globales et locales), la recherche de sous graphes fortement connexes, les adaptations des modèles petits-mondes et sans-échelle et enfin l'étude dynamique des graphes bipartis. Les enjeux posés par la visualisation des graphes bipartis seront abordés dans une synthèse ultérieure.

1 Mesurer un graphe biparti

La plupart des mesures utilisées pour l'analyse des graphes simples peut être adaptée aux graphes bipartis. Mais si le calcul est souvent possible, il arrive que l'interprétation de certaines mesures devienne plus problématique.

La densité d'un graphe biparti symbolisant les relations entre deux groupes distincts d'acteurs a et b se calcule en divisant le nombre de liens par le produit du nombre de sommets dans chacun des deux ensembles. En effet, la densité maximale - celle d'un graphe biparti complet donc - suppose que tous les acteurs assistent à tous les événements.

Les mesures de centralité les plus fréquentes (degré, intermédiarité, proximité) peuvent être calculées pour chacun des deux ensembles de sommets d'un graphe biparti. Si on souhaite obtenir des mesures normalisées, on divisera les résultats obtenus par le maximum possible.

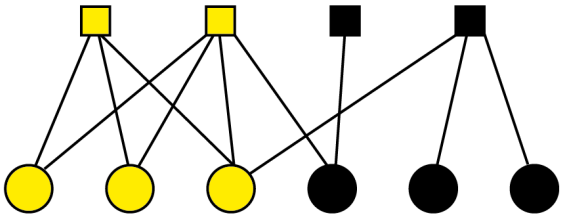
Les degrés sont calculés de façon similaire à celle des graphes simples mais, la direction ici important peu, il n'y a pas de différenciation entre degré entrant et degré sortant. Si le graphe étudie les relations entre deux groupes distincts a et b, le degré maximum d'un sommet du groupe A correspond au nombre d'individus du groupe B. Comparer les distributions respectives des degrés des groupes a et b est possible et peut donner des résultats intéressants mais doit être thématiquement justifié. La corrélation entre les degrés des deux ensembles est également possible. Si dans la plupart des cas, les deux ensembles de sommets sont mesurés séparément, il semblerait pertinent que, par exemple, la centralité d'un événement soit définie à partir des centralités des acteurs qui y participent et non de leur seul nombre [11]. Bonacich a ainsi proposé en 1991 une mesure de la centralité des événements proportionnelle à la centralité des individus qui y assistent et une centralité des acteurs proportionnelle aux événements auxquels ils participent (mesure basée le calcul des valeurs et vecteurs propres) [4] [8].

La distance géodésique - le nombre de liens correspondant au plus court chemin - entre deux sommets du même ensemble est au minimum de 2 et

7. Sur les conséquences méthodologiques de la transformation sur les mesures de centralité, voir par exemple l'article de Billand *et al.* à propos des programmes scientifiques du 6e PCRD [3].

8. La mesure ayant été semble-t-il peu reprise, aucun terme ne s'est imposé pour la nommer.

FIGURE 2 – Les bi-cliques



Les sommets en jaune forment une bi-clique : tous les liens possibles entre les sommets des deux groupes sont en effet présents.

elle est nécessairement paire.

Certaines mesures ne peuvent être calculées dans un graphe biparti. Ainsi, la transitivité ne peut être calculée pour trois sommets : si deux acteurs a et b sont en relation avec l'événement A, il ne peut par définition pas y avoir de lien entre a et b, ces deux acteurs appartenant au même ensemble. Une adaptation est néanmoins possible : si deux acteurs a et b sont en relation avec A, la relation entre b et B entraîne-t-elle une relation entre a et B ? Le même type d'adaptation a été proposé pour mesurer les coefficients de *clustering* locaux et globaux (voir *infra*).

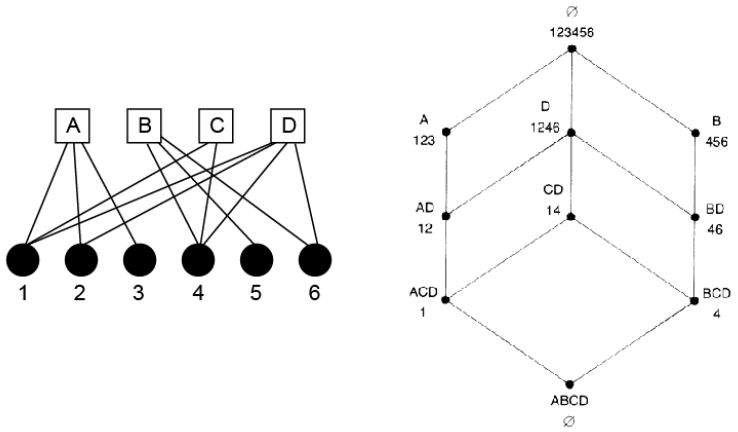
2 Cliques et communautés

Une clique - ensemble maximal d'acteurs entre lesquels tous les liens possibles sont présents - ne peut par définition pas exister dans un réseau biparti, les liens entre acteurs du même ensemble étant exclus. Borgatti et Everett ont proposé la recherche des bi-cliques définies comme l'ensemble maximal d'acteurs entre lesquels tous les liens possibles entre les deux ensembles distincts sont présents [5]. Dans la figure 2, les sommets en jaune forment une bi-clique.

Les *k-plex* ont également été adaptés au graphes bipartis. Dans un graphe simple, le *k-plex* correspond au sous-graphe maximal comprenant n acteurs et où chacun des acteurs est relié à au moins $n-k$ acteurs de ce sous-graphe. Un *(k1,k2)-biplex* peut être défini comme le sous-graphe biparti maximal tel que chaque sommet de A soit adjacent à B - k1 sommet et tel que chaque sommet de B soit adjacent à A - k1 sommets.

D'autres pistes ont été récemment proposées : Zha *et al.* suggèrent de normaliser les liens du graphe avant d'appliquer une variante d'analyse des correspondances [29], Guimerà *et al.* [16] utilisent la modularité sur le graphe biparti puis sur sa variante *one-mode* valuée pour déterminer les sous groupes

FIGURE 3 – Le treillis de Galois



*Le principe du treillis de Galois (*Galois lattice*) est de représenter sous forme de graphe l'ensemble ordonné des relations possibles entre les différents éléments considérés. À gauche, un graphe biparti, à droite le treillis de Galois correspondant : en haut se trouvent les acteurs (avec un ∅ car aucun événement ne réunit les six), en bas les événements (avec un ∅ car aucun acteur n'assiste aux quatre). Chaque point est caractérisé simultanément par les acteurs et les événements qui le caractérise. Parcourir le treillis permet donc de reconstituer le graphe de départ (acteur 1 assiste à ACD etc.), il y a transformation de la visualisation - et des analyses possibles - sans perte d'information.*
Source : Freeman et White, 1993 [15].

fortement connexes, Field *et al.* utilisent un modèle p^* [13]. D'autres auteurs ont mobilisé les treillis de Galois - voir la figure 3 -, suivant ainsi la piste proposée par Freeman [15][14], c'est par exemple le cas de Crampes et Planty [6]. Comparer les approches est délicat dans la mesure où les auteurs utilisent des jeux de données différents.

3 Réseaux bipartis petits-mondes et sans échelle

L'analyse de réseau biparti a connu un renouveau important avec l'intégration de méthodes relatives aux réseaux petits-mondes et sans échelle (voir notamment l'article et la bibliographie de Lapaty *et al.* consacré à l'analyse de grands réseaux [17]).

La piste proposée par Newman *et al.* est de comparer les indicateurs du graphe biparti étudié avec un graphe biparti aléatoire (les spécificités de ces

simulations sont décrites dans [22]), l'objectif étant que la distribution des degrés soit similaire pour chacun des ensembles de sommets. Cette méthode a par exemple été utilisée par deux sociologues étudiant les comédies musicales à Broadway (réseau biparti acteurs - spectacles) [27] et par Robins et Alexander pour étudier les réseaux dirigeants - conseils d'administration en Australie et aux États-Unis [24]. Sur ce thème des *corporate interlocks*, Davis *et al.* ont eux montré le caractère petit-monde des réseaux *interlock* aux États-Unis [8], mais en utilisant les matrices transformées entreprises - entreprises et dirigeants - dirigeants [9] et non la matrice bipartie originelle.

Newman a très tôt (2001) cherché à appliquer ces méthodes aux réseaux de co-publication : les deux ensembles sont les articles et les auteurs, un lien existant entre deux auteurs s'ils ont co-signé un article [21][20]. On peut d'ailleurs se demander s'il ne serait pas judicieux de le considérer comme un graphe simple, la co-signature d'un article supposant *a priori* un minimum d'interactions entre les auteurs... L'intérêt est pourtant réel dans la mesure où cela permet de calculer des indices tant sur les auteurs que sur les articles et de chercher les relations entre ces deux séries d'indices. Les mesures proposées dans ces deux articles mêlent « approches traditionnelles » (degrés, taille des composantes connexes, distance moyenne, intermédiarité) et approches plus récentes (coefficient de *clustering*, effet petit-monde). Cependant, exceptés les degrés, la structure bipartie du graphe ne semble pas considérée, les indicateurs étant apparemment calculés sur le graphe non valué des relations entre auteurs - le poids des liens étant pris en compte dans la troisième partie du deuxième article seulement.

Tore Opsahl a également proposé plusieurs moyens d'adapter les mesures de *clustering coefficient* aux réseaux bipartis booléens et valués : plutôt que de prendre la triade comme élément de référence, il utilise les cycles de longueur 4 [23]. Zhang *et al.*, poursuivant une piste de Lind *et al.* [18], ont proposé deux mesures différentes du *clustering coefficient*, l'une fondée sur les cycles de longueur 3 et l'autre sur ceux de longueur 4, mesures qu'ils utilisent ensuite pour proposer une partition fondée sur la suppression des liens ayant le plus faible *clustering coefficient* [30].

4 Dynamiques et modèles statistiques

L'extension des modèles ERGM à ces réseaux a été proposée à plusieurs reprises avec des nuances méthodologiques dont il n'est pas toujours simple de saisir la portée : dès 1999 par Skvoretz et Faust [26], en 2008 par Wang *et al.* [28], en 2010 par Stadtfled et Geyer-Schulz dans un article non publié

9. Les calculs ont été en fait réalisés sur la plus grande composante connexe de ces deux matrices et, semble-t-il, avec des liens non valués.

qui a reçu le prix du meilleur article d'étudiant de l'INSNA [10]. Le caractère exploratoire de ces articles doit être souligné, aucun consensus n'existant encore sur les critères permettant de déterminer le modèle le plus performant. Comme pour les modèles ERGM *one-mode*, le principe consiste à modéliser les tendances du graphe observé en les comparant à des graphes bipartis aléatoires.

L'un des articles les plus pédagogiques [11] permettant de comprendre comment le modèle est adapté aux réseaux bipartis et quelles sont les structures recherchées est sans doute celui de Faust *et al.* qui étudie les réseaux politiques sous Brejnev [12]. Les auteurs analysent un réseau décrivant la participation à des événements sociaux et à des réunions officielles de membres de l'élite politique soviétique. Dans un premier temps, une analyse factorielle des correspondances permet uniquement de distinguer les responsables politiques en fonction de leur officine d'origine (ministère, Politburo etc.) et aucune différence notable n'apparaît en ce qui concerne la nature des réunions (officielle ou sociale). Le modèle p^* mené ensuite permet lui de mettre en évidence des groupes, parfois similaires à ceux obtenus avec l'ACP, mais la nature différente des réunions apparaît plus nettement.

Une piste intéressante a été proposée par Doreian dans un article déjà ancien [9] : certains réseaux bipartis inclut de façon implicite ou non une dynamique temporelle. Ainsi le *Davis dataset* concerne 14 événements sociaux et successifs. La même remarque pourrait être faite sur les co-publications d'articles ou l'utilisation des mêmes mots clés. Le traitement proposé change alors de nature dans la mesure où les événements deviennent des bornes chronologiques et seules les relations entre individus sont étudiées.

Sharara *et al* [25] ont récemment proposé deux mesures de centralité pour les graphes bipartis « dynamiques » (un graphe mesuré à différents moments) : l'une concerne la stabilité (même individu assistant aux mêmes événements), l'autre la diversité (participants à des événements différents).

Conclusion

De nombreux réseaux issus d'études empiriques peuvent être analysés comme des graphes bipartis, et la transformation en graphe de co-occurence n'est pas toujours nécessaire.

Néanmoins, à la lecture de certains travaux, il est permis de s'interroger sur la plus-value apportée par l'analyse de réseaux. Lorsque le réseau est de type individu - attribut (ex. un auteur utilise certains mots clés), il est possible de penser que d'autres outils d'analyse multivariée seraient tout

10. C Stadtfeld, A Geyer-Schulz, "Analysing event stream dynamics in two mode networks", INSNA The Best Student Paper Award, 2010, accessible en ligne.

11. Il suppose tout de même de connaître ce type de modèle...

indiqués. En effet, l'analyse de réseau consiste dans ce cas à extraire d'un tableau individus - variables une variable et une seule... Borgatti dans une conférence en 2009 s'inquiétait de cette tendance à tout représenter par des réseaux sans la moindre réflexion théorique ou conceptuelle [12]. Si représenter n'importe quel type de phénomène par un graphe est possible, et si la visualisation peut fournir des pistes intéressantes, mener une analyse de réseaux *stricto sensu* n'est pas toujours le plus pertinent.

Autre problème voisin, pourquoi se limiter à des graphes bipartis ? Pourquoi ne pas imaginer des graphes tri ou quadri-partis ? On pourrait ainsi imaginer étudier des graphes comprenant x ensembles d'acteurs distincts. Pour se limiter à trois et donner un exemple, pourquoi ne pas considérer les liens entre les auteurs (groupe 1) publiant dans des revues (groupe 2) appartenant à des éditeurs (groupe 3) ? Des travaux exploratoires existent sur le sujet [2] mais il est possible, là encore, de s'interroger, au-delà de l'enjeu méthodologique, sur l'intérêt thématique de telles formalisations. Dans l'exemple donné à l'instant, le groupe 3 peut ainsi être considéré comme un attribut du groupe 2 plutôt que comme un ensemble indépendant.

Un enjeu peut-être plus stimulant serait de développer les méthodes multi-niveaux permettant l'analyse de graphes bipartis où sont intégrées les relations à l'intérieur de chacun des deux ensembles distincts d'acteurs. On peut, par exemple, considérer le graphe biparti des carnets de recherche et des auteurs sur la plate-forme hypotheses.org : un lien existe si l'auteur a a publié un billet dans le carnet A. Mais les carnets peuvent se citer les uns les autres, et les auteurs écrire des billets ensemble ou se commenter d'un blog à l'autre.

Références

[1] Oussama Allali : *Structure et dynamique des graphes de terrain bipartis : liens internes et prédiction de liens.* Thèse de doctorat, Université Pierre et Marie Curie (UPMC), 2011.

[2] Vladimir Batagelj, Anuška Ferligoj et Patrick Doreian : Indirect Blockmodeling of 3-Way Networks. *Selected Contributions in Data Analysis and Classification*, pages 151–159, 2007.

[3] Pascal Billand, David Frachisse et Nadine Massard : The sixth Framework Program as an affiliation network : Representations and analysis. *13th Coalition Theory Network Workshop*, 2008.

[4] Phillip Bonacich : Simultaneous group and individual centralities. *Social Networks*, 13(2):155–168, 1991.

12. *"Network is the new pie chart - the hot way to display any and all information"*, S. Borgatti, "The analysis of 2-mode networks", *Conference and Workshop on Two-Mode Social Network Analysis*, 1 October, 2009 VU University Amsterdam.

[5] Stephen P. Borgatti et Martin G. Everett : Network analysis of 2-mode data. *Social Networks*, 19(3):243–269, 1997.

[6] Michel Crampes et Michel Plantié : Détection de communautés dans les graphes bipartis. *Actes de la conférence IC 2012*, 2012.

[7] A. Davis, B.B. Gardner et M.R. Gardner : *Deep South*. University of Chicago Press, 1941.

[8] Gerald F. Davis, Mina Yoo et Wayne E. Baker : The Small World of the American Corporate Elite, 1982–2001. *Strategic Organization*, 1(3):301–326, 2003.

[9] Patrick Doreian : On the Evolution of Group and Network Structure. *Social Networks*, 2(3):235–252, 1979/80.

[10] Martin G. Everett et Stephen P. Borgatti : .

[11] Katherine Faust : Centrality in affilliation networks. *Social Networks*, 19(2):157–191, 1997.

[12] Katherine Faust, Karin E. Willert, David D. Rowlee et John Skvoretz : Scaling and statistical models for affiliation networks : patterns of participation among Soviet politicians during the Brezhnev era. *Social Networks*, 24(3):231–259, 2002.

[13] Sam Field, Kenneth A. Frank, Kathryn Schiller et Catherine Riegle-Crumb : Identifyiing positions from affiliation networks : Preserving duality of people and events. *Social Networks*, 28(2):97–123, 2006.

[14] Linton C. Freeman : Cliques, Galois lattices, and the structure of human social groups. *Social Networks*, 18(3):173–187, 1996.

[15] Linton C. Freeman et Douglas R. White : Using Galois Lattices to Represent Network Data. *Sociological Methodology*, 23:127–146, 1993.

[16] R. Guimerà, M. Sales-Pardo et L.A.N. Amaral : Module identification in bipartite and directed networks. *Physical Review E*, 76(3):36102, 2007.

[17] Matthieu Latapy, Clémence Magnien et Nathalie Del Vecchio : Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31–48, 2008.

[18] Pedro G. Lind, Marta C. González et Hans J. Herrmann : Cycles and clustering in bipartite networks. *Physical Review E*, 72(5), 2005.

[19] Mark S. Mizruchi : What Do Interlocks Do ? An Analysis, Critique, and Assessment of Research on Interlocking Directorates. *Annual Review of Sociology*, 22:271–298, 1996.

[20] Mark E.J. Newman : Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64(1):16131, 2001.

[21] Mark E.J. Newman : Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical review E*, 64(1): 16132, 2001.

[22] Mark E.J. Newman, Steven H. Strogatz et Duncan J. Watts : Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):26118, 2001.

[23] Tore Opsahl : Triadic closure in two-mode networks : Redefining the global and local clustering coefficients. *Social Networks*, 2011.

[24] Garry Robins et Malcolm Alexander : Small worlds among interlocking directors : network structure and distance in bipartite graphs. *Computational & Mathematical Organization Theory*, 10(1):69–94, 2004.

[25] Hossam Sharara, Lisa Singh, Lise Getoor et Janet Mann : Finding Prominent Actors in Dynamic Affiliation Networks. *Human Journal*, 1(1), 2012.

[26] John Skvoretz et Katherine Faust : Logit models for affiliation networks. *Sociological Methodology*, 29(1):253–280, 1999.

[27] Brian Uzzi et Jarrett Spiro : Collaboration and Creativity : The Small World Problem. *American Journal of Sociology*, 111(2):447–504, 2005.

[28] Peng Wang, Ken Sharpe, Garry L. Robins et Philippa E. Pattison : Exponential random graph (p*) models for affiliation networks. *Social Networks*, 31(1):12–25, 2009.

[29] Hongyuan Zha, Xiaofeng He, Chris Ding, Horst Simon et Ming Gu : Bipartite graph partitioning and data clustering. *In Proceedings of the tenth international conference on Information and knowledge management*, pages 25–32, 2001.

[30] Peng Zhang, Jinliang Wang, Xiaojia Li, MMenghui Li, Zengru Di et Ying Fan : Clustering coefficient and community structure of bipartite networks. *Physica A*, 387(27):6869–6875, 2008.