L'analyse de réseau en sciences sociales. Petit guide pratique

Laurent Beauguitte

L'analyse de réseau en sciences sociales

Petit guide pratique

Laurent Beauguitte

UMR Géographie-cités

Groupe fmr (flux, matrices, réseaux)

Espaces et radicalités

Version 0.1 - Été 2023

Version 0.1 (été 2023)

Deux personnes oubliées des remerciements dans la version zéro sont désormais présentes. Une vingtaine de coquilles a été corrigée. Un lien a été ajouté vers la thèse de Sampson page 77 (merci à François Briatte). Une version html est de ce petit guide est désormais accessible à l'adresse https://beauguitte.github.io/analyse-de-reseau-en-shs/.

Remerciements plus des bricoles

J'avais prévu de faire relire ce guide pratique avant de le mettre en ligne et je n'ai pas osé déranger les collègues dont la majorité est sous l'eau, débordée et/ou en *burn-out* chronique - j'ai tout de même demandé une relecture pour les annexes mathématiques à mon frère, un grand merci Pierre ! J'espère n'avoir pas écrit trop de bêtises ni laissé traîner trop de coquilles. Ce guide doit beaucoup aux différentes collègues avec lesquelles j'ai eu le plaisir d'animer des formations ces dernières années, notamment, par ordre alphabétique, Fabien Eloire, Violaine Jurie, Karine Karila-Cohen, Claire Lemercier, Rosemonde Letricot, Marion Maisonobe, Silvia Marzagalli, Pierre Mercklé, Hugues Pecout et Isabelle Rosé. Ce guide n'existerait pas si nous n'avions pas fondé en 2010 le groupe fmr (flux, matrices, réseaux) avec César Ducruet : un grand merci à toi et au « noyau dur historique » (Françoise, Marion, Matthieu et Serge). Merci aussi à François Briatte pour le *Awesome Network Analysis* et pour les découvertes musicales (Olhava !). Iels ne sont évidemment pas responsables des approximations et des formules à l'emporte-pièce présentes dans ce texte.

En 2023, on continue à lire des appels pour des journées/salons/forums des « doctorants » ou des calendriers concernant les concours « chercheurs ». Ça m'agace. Donc ce texte, comme plusieurs de mes textes récents, est exclusivement écrit au féminin ; la seule exception concerne les communautés scientifiques exclusivement masculines évoquées dans le chapitre 2.

Ce guide pratique est entièrement libre et gratuit. Fonctionnaire payé par l'argent public, j'estime que mon travail doit accessible à toute personne intéressée. Et que je ne suis pas là pour rapporter des sous à quelque éditeur que ce soit. Ce qui ne vous dispense pas, si ce guide vous est utile, de le citer dans vos travaux. Comme ça par exemple : Laurent Beauguitte, 2023, *L'analyse de réseau en sciences sociales. Petit guide pratique*. Groupe fmr, url : et là vous mettez l'url d'HAL.

Si vous repérez des erreurs, n'hésitez pas à me les signaler, soit par mail[1], soit en commentant le billet dédié sur le carnet de recherche du groupe fmr (L'AR en sciences sociales. Petit guide pratique - SAV). Si la lecture de certains chapitres ne vous satisfait pas totalement (ce qui est mon cas...), si vous avez l'impression qu'il manque des éléments, que je passe trop vite sur certains aspects, si vous aimeriez que telle méthode soit abordée, etc. etc. surtout, surtout, n'hésitez pas à me le signaler. Une version Un verra sans doute le jour prochainement.

1. laurent<dot>beauguitte<at>cnrs<dot>fr

Sommaire

Introduction

Ce qu'est ce guide et ce qu'il n'est pas

Ce guide n'est pas un manuel : il ne donne pas de recettes, il n'explique pas ce qu'il faut faire pour mener à bien une analyse de réseau de A à Z. Ce guide tente de présenter ce qu'il est possible de faire - souvent beaucoup de choses différentes, il faudra donc choisir, et argumenter ses choix - et donne des conseils aussi génériques que possible pour réaliser des analyses pouvant s'avérer pertinentes.

Ce guide s'adresse à toute personne 1. se demandant ce qu'est l'analyse de réseau et si elle en a besoin, 2. formée à l'analyse de réseau dans une discipline et désirant savoir ce qui se fait dans d'autres disciplines et 3. souhaitant approfondir les approches concernant des réseaux spécifiques (réseau bimodal, multiplexe, dynamique notamment). Affirmer qu'aucun pré-requis mathématique n'est demandé serait absurde, il est nécessaire de savoir compter par exemple, mais si vous n'avez pas fait le moindre exercice de maths depuis la première, ça ne devrait poser aucun problème de compréhension. Les notations et les formules mathématiques sont regroupées et commentées en fin de volume dans les annexes A et B.

Ce guide n'est pas exhaustif : toutes les mesures ne sont pas abordées, toutes les analyses possibles sur tous les types possibles de réseaux ne sont pas présentées et les modèles statistiques ne sont que brièvement évoqués. Les aspects algorithmiques liés à l'analyse de réseau ne sont pas abordés dans la mesure où rares sont les personnes en sciences sociales qui créent de toute pièce leurs outils [1]. L'objectif de ce guide n'est pas d'être complet, ce qui est devenu compliqué en analyse de réseau, mais de vous aider à vous poser quelques questions utiles : que représente mon réseau ? que sont mes liens ? mes sommets ? quelles sont les mesures pertinentes pour mes questions de recherche et celles qui le sont moins ? quelles précautions prendre quand je crée puis que je commente une visualisation de réseaux ? quels logiciels utiliser ? quelles revues lire pour me tenir à jour ?

Les exemples évoqués ainsi que les jeux de données mobilisés sont d'une part des jeux de données devenus classiques en analyse de réseau et d'autre part des exemples issus de travaux plus ou moins récents et issus de différentes disciplines. Étant, comme toute chercheuse, située dans un champ académique et disciplinaire donné, il est logique qu'un certain nombre de ces derniers soient issus de mon réseau professionnel personnel.

1. Je partage l'opinion de Newman qui affirme en substance dans son manuel que, tôt ou tard, on a besoin de faire des choses que les logiciels courants ne savent pas faire et qu'il est alors nécessaire de créer ses propres fonctions ou programmes. Il ajoute également à juste titre que se limiter aux mesures et aux méthodes déjà existantes limite la production de connaissances (2013, p. 277-278). Les compétences en programmation étant ce qu'elles sont en sciences sociales, cette perspective me paraît malheureusement trop lointaine pour pouvoir être abordée ici.

Ce guide s'appuie sur une décennie de formations en analyse de réseau, formations de durée variable (de 3 à 40 heures) et destinées à des publics de statuts, de niveaux et de disciplines variées [1] (élèves ingénieures statisticiennes, masterantes, doctorantes en sciences sociales, etc.) et sur plus d'une décennie de rédaction de tutoriels [2]. Géographe, je présente ici des méthodes issues de plusieurs disciplines (sociologie, physique, géographie, etc.). J'ai cherché à être aussi générique que possible afin que ce guide puisse vous être utile, que vous soyez anthropologue, archéologue, géographe, historienne, sociologue ou autre ; je ne prétends pas avoir réussi, à vous de me le faire savoir si vous le souhaitez.

Les encadrés intitulés *À l'intention des formatrices* donnent un certain nombre de conseils pratiques aux personnes souhaitant animer des formations d'analyse de réseau. N'hésitez surtout pas à m'envoyer vos retours pour signaler ce qui marche et ce qui ne marche pas.

Enfin, ce texte est écrit par un agnostique. Je ne considère pas que l'analyse de réseau soit toujours pertinente ni qu'elle permette systématiquement d'obtenir des résultats intéressants : l'analyse de réseau est un outil disponible parmi d'autres, rien de plus, rien de moins. Mieux vaut une analyse de correspondance maîtrisée qu'une analyse de réseau mal fichue. Je ne considère pas non plus que telle ou telle approche soit systématiquement supérieure à telle autre : parfois il est pertinent de faire de la détection de communautés, parfois étudier la distribution des degrés est utile ; parfois ces deux approches sont inadaptées. J'indique quand telle ou telle approche paraît pertinente et quand elle l'est moins. Ceci explique qu'il n'y ait pas dans ce guide d'exemple d'analyse de réseau menée du début (construction des données et de la problématique) à la fin (interprétation des résultats) : il n'existe pas, à mon avis, de déroulé standard et passe-partout en analyse de réseau.

Structure du guide

Le chapitre 1 tente de répondre à une question simple : pourquoi faire de l'analyse de réseau ? Le chapitre 2 présente brièvement quelques traditions disciplinaires d'analyse de réseau, sa lecture n'est pas essentielle à un public débutant. Le chapitre 3 présente les termes indispensables, en partie issus de la théorie des graphes, pour décrire les réseaux étudiés. L'équivalent anglais de chaque terme est donné en italiques et entre parenthèses, les logiciels disponibles et une grande partie de la littérature étant dans cette langue. Le chapitre 4 propose quelques règles simples et de nombreux exemples permettant de mettre en forme des données relationnelles.

Le chapitre 5 est consacré aux mesures ; seules les plus courantes sont présentées. La compréhension des différentes mesures et de leur utilité permet la présentation de trois grands modèles théoriques de réseau (section 5.4). Le chapitre 6 s'intéresse aux méthodes permettant de simplifier les réseaux d'une part et aux méthodes permettant de partitionner les individus d'autre part. Là encore, les méthodes sont nombreuses et seules les plus fréquemment implémentées dans les logiciels courants d'analyse de réseau sont présentées. La description des indicateurs et des méthodes mobilise peu les symboles et formules mathématiques, celles-ci étant expliquées dans les annexes A et B.

1. J'écris au féminin et j'utilise l'accord de voisinage.
2. Voir notamment les carnets de recherche groupe fmr et GDR Analyse de réseaux en SHS ainsi que la collection du groupe fmr sur HAL.

Question de vocabulaire

Dans les pages qui suivent, les termes **individu** et **population** sont employés au sens statistique du terme : l'individu est l'unité d'observation, l'ensemble des individus étudiés constitue la population. Un individu peut donc être une personne, une plante, une entreprise, un mot, un article, une ville etc. etc.
Le terme **graphe** est utilisé pour désigner l'objet mathématique constitué d'un ensemble de points et d'un ensemble de liens entre ces points. J'utilise le terme **réseau** dès qu'un attribut est donné aux points (un nom par exemple) et/ou aux liens.
J'appelle **analyse de réseau** toute démarche quantitative visant à étudier les propriétés relationnelles d'un ensemble d'individus. Le fait d'analyser *un* réseau donné (un réseau ferré, un réseau migratoire, un réseau social numérique) ne signifie pas qu'on pratique l'analyse de réseau. Inversement, des objets qui ne sont pas *a priori* des « réseaux » (un corpus de textes, de votes, de participations à des événements) peuvent être étudiés à l'aide de l'analyse de réseau.
Les sociologues anglophones pratiquant l'**analyse de réseaux sociaux** (*Social Network Analysis - SNA*) emploient parfois le terme qu'elles considèrent équivalent de *structural analysis*. Pour éviter la confusion possible avec l'analyse structurale issue des travaux de Claude Lévi-Strauss, certains sociologues francophones emploient l'expression « analyse néo-structurale ». Seule l'expression *SNA* est utilisée dans ce texte.
Les **réseaux sociaux** étudiés en SNA sont essentiellement des réseaux de relations entre individus. Les outils et platerformes permettant des échanges en ligne (Twitter, Instagram, etc.) sont appelés **réseaux sociaux numériques** dans ce texte.
Les termes de **science des réseaux** (*network science*) et de **réseau complexe** (*complex network*) ont pris une importance croissante ces dernières années/décennies. Ils sont très peu utilisés dans ce texte.
Le terme **modélisation** utilisé seul désigne tout processus visant à transformer un aspect du monde social en données pouvant être analysées pour répondre à une question de recherche. Les termes **modélisation graphique** et **modélisation statistique** désignent les tentatives de décrire et/ou d'expliquer ces aspects du monde social soit à l'aide de schémas soit à l'aide de calculs.

Les chapitres suivants présentent des méthodes adaptées à des réseaux particuliers : réseau bimodal (liens entre deux ensembles différents d'individus - chapitre 7), réseau multiplexe (existence de plusieurs relations entre les individus - chapitre 8), réseaux dynamiques (chapitre 9) et réseaux personnels (chapitre 10). Les indicateurs et les méthodes sont moins stabilisées pour ces trois premiers types de réseaux, ce qui explique sans doute pourquoi elles sont généralement moins traitées dans les manuels. J'ai tenté de sélectionner les indicateurs et méthodes les plus faciles à utiliser et cette sélection est donc largement dépendantes des logiciels disponibles. Si votre réseau n'appartient pas à ces catégories, vous pouvez allègrement sauter ces chapitres.

Le chapitre 11 présente deux types de modélisation : les modélisations graphiques et les modélisations statistiques. Les premières sont peu couramment employées - alors que leur usage peut s'avérer précieux - et les secondes supposent des compétences mathématiques supérieures aux méthodes évoquées précédemment. Ceci explique tant la position du chapitre que sa brièveté ; sa lecture n'est sans doute pas indispensable en première approche. Plus développé, le chapitre 12 est consacré à la visualisation des données relationnelles. Il présente quelques règles basiques de sémiologie graphique, des conseils pratiques permettant de créer des images de réseaux lisibles et enfin différents modes de représentation. Le

nombre de praticiennes en analyse de réseau ne proposant aucune forme de visualisation étant à ma connaissance très restreint, la lecture du chapitre 12 est recommandée.

Les deux derniers chapitres ont un intérêt pratique plus immédiat : le chapitre 13 vise à vous guider dans le choix d'un ou de plusieurs logiciels pour mener à bien vos analyses de réseaux ; le chapitre 14 présente différentes ressources permettant d'actualiser et/ou d'approfondir vos connaissances.

L'annexe A est découpée en deux parties : une introduction expliquant comment lire une équation ; un point sur le rôle du calcul matriciel en analyse de réseau. L'annexe B détaille les formules mathématiques de quelques indicateurs usuels en analyse de réseau.

L'index des noms propres est avant tout une lubie personnelle, facilitée par l'utilisation de LaTeX, me permettant de contrôler les surreprésentations de genre ainsi que les poids disciplinaires respectifs des autrices citées. Il est logique que plus d'hommes soient cités avant 1960-1970 (en géographie et en *SNA* notamment) ; il est logique que les sociologues quantitativistes et les physiciennes soient citées plus souvent (elles ont écrit les manuels de référence, je m'en suis copieusement nourri et je paye mes dettes) ; j'ai cependant tenté d'équilibrer les genres et de varier les disciplines évoquées.

Bonne lecture et à bientôt j'espère pour une version révisée et augmentée de ce petit guide pratique.

Chapitre 1

Pourquoi faire de l'analyse de réseau ?

Se former à l'analyse de réseau prend du temps, de l'énergie et les résultats peuvent être décevants voire triviaux. Répondre à la question posée ici n'est donc pas tout à fait inutile. J'examine d'abord la question des données avant d'aborder celle des questions de recherche.

1.1 Des méthodes quantitatives adaptées aux données relationnelles

L'analyse de réseau, quelle que soit votre discipline, est utile si et seulement si vous souhaitez étudier de manière *quantitative* un *ensemble de données* que vous considérez comme *relationnelles*. Le terme *quantitatif* ne doit pas effrayer : il signifie qu'une analyse de réseau suppose à un moment ou à un autre des mesures portant sur vos données, mesures qui permettront notamment de qualifier votre réseau et de classer vos individus (au sens statistique du terme, voir encadré *supra*). Pourquoi parler de données « considérées comme relationnelles » ? Tout simplement parce que les données que vous souhaitez étudier pour répondre à vos questions de recherche sont des objets que vous avez construits ou que vous avez récupérés et, dans ce cas, d'autres que vous les ont construits. Il n'existe pas de données qui soient « par nature » relationnelles [1].

Imaginons que je souhaite travailler sur la production scientifique mobilisant l'analyse de réseau. Je construis un corpus d'articles et, pour chacun des articles, je récupère un ensemble de données sur les autrices (genre, institution, discipline) et sur les articles (date, résumé, mots clés) (figure 1.1). Il existe de très nombreux moyens d'analyser un tel corpus : je pourrais calculer des indicateurs statistiques univariés pour les autrices, les revues, caractériser les articles (combien d'articles écrits seule ? combien d'articles à plusieurs ? combien d'articles co-signés par des personnes de disciplines différentes ? etc.) ; je pourrais mobiliser des méthodes d'analyse lexicale pour analyser l'ensemble des résumés (liste non exhaustive).

1. Il n'existe pas dans le monde social des choses qui pourraient être considérées comme des données « brutes » qu'il suffirait de « nettoyer » pour produire de la connaissance ; il est toujours nécessaire de *construire* ces données. L'illusion de transparence liée à certains types de données, notamment les données numériques issues d'activités en ligne, est abordée dans le chapitre 4.

Si je décide d'utiliser l'analyse de réseau, je dois définir un type de relations entre certains des éléments présents. La liste qui suit n'est pas exhaustive et vise simplement à montrer qu'un même corpus peut se prêter à différentes modélisations, le choix d'une modélisation étant lié aux thématiques que je souhaite traiter :

- mes individus sont les autrices : je crée un lien entre deux autrices quand elles ont publié un article ensemble ;
- mes individus sont les mots clés : je crée un lien quand deux mots clés sont utilisés pour le même article ;
- mes individus sont les institutions : je crée un lien entre deux institutions quand des autrices issues d'institutions différentes co-signent un article ;
- je choisis d'étudier les liens entre deux populations : celle des autrices et celle des revues, celle des mots clés et celle des revues, etc. (chapitre 7)

Dans la mesure où je construis une liste de liens entre des individus, il est possible de mener ensuite une analyse de réseau.

FIGURE 1.1 – Un article scientifique, x réseaux potentiels

Analysing Personal Networks in Geographical Space Beyond the Question of Distance

Claire Bidart
LEST, CNRS, Aix Marseille Univ, France
Marion Maisonobe
Géographie-cités, CNRS, France
Gil Viry
School of Social and Political Science, University of Edinburgh, UK

Abstract: Recent literature recognises the importance of situating social networks in spatial context. Yet, the spatial analysis of personal networks has often been limited to examining residential distances between actors. While distance is a central characteristic of social relationships, it is a poor indicator for understanding the intricacies of the geographical space, places and personal networks. This study develops an original approach for mapping and analysing personal networks based on their geographical scope and the distribution of the residential locations of network members in relevant geographical areas. We perform a factor and cluster analysis to identify the major geographical patterns of personal networks using two samples of egocentric networks from France and Switzerland. We validate the approach first by interpreting the patterns both quantitatively and qualitatively, and second by examining how these patterns relate to important social characteristics of respondents and their personal networks. We conclude by discussing the significance of this approach for integrating geographical information into the analysis of personal networks and for rethinking networks and the geographical space as co-constituted.

Keywords: distance; geographical space; mixed methods; personal networks; place; social network analysis

Published: 20 September 2022

Capture d'écran du site de la revue Social Inclusion faite le 24 février 2023.

Il n'existe pas *une* façon pertinente de modéliser ses données, un même corpus de départ peut donc donner lieu à des formalisations multiples. Ce n'est pas parce que tel aspect du monde social est *toujours* modélisé de la même façon dans une discipline donnée qu'il s'agit de la seule modélisation pertinente : la recherche académique est parfois d'une paresse et d'un conformisme terribles. Mais répéter ce qui a déjà été fait x fois dans nos disciplines n'est peut-être pas le moyen le plus adapté pour produire des résultats intéressants.

Les choix opérés lors de la construction des données relationnelles sont fonction des questions de recherche que l'on souhaite explorer. Travailler sur les seuls mots clés du corpus peut permettre d'identifier les thématiques d'une discipline donnée, leur évolution ou les effets de mode scientifique ; travailler sur les liens entre laboratoires des autrices et revues

permet de construire une géographie de l'activité scientifique ; travailler sur les liens entre autrices et revues peut permettre de mettre en évidence des communautés au sein d'une discipline. Comme toujours avec les méthodes, qu'elles soient qualitatives, quantitatives ou mixtes, les questions de recherche sont premières et guident la construction des données.

> *À l'intention des formatrices*
>
> Partir de données les plus « brutes » possibles est intéressant dans la mesure où il est possible de montrer comment les questions de recherche permettent de construire les données. Inversement, partir d'un jeu de données classique de l'analyse de réseau escamote la construction des questions de recherche et donc des données. Il peut être intéressant au niveau pédagogique de prendre des exemples connus de l'ensemble des participantes (réseau social numérique, corpus de livres ou d'articles).
> Il me semble important de présenter rapidement ce dont on a besoin pour mener à bien une analyse de réseau : une question, une liste d'individus, une relation entre ces individus. Présenter dès le départ les formats liste et matrice me paraît également une pratique souhaitable dans la mesure où les participantes peuvent utiliser toute la durée de la formation pour réfléchir aux mises en forme possibles de leurs données (quand elles en ont déjà évidemment).

Quelle que soit la formalisation choisie, vous obtenez une liste de liens entre des individus. Si je prends la première des options listées plus haut (« mes individus sont les autrices et je crée un lien entre deux autrices quand elles ont publié un article ensemble »), j'obtiens les liens suivants : CB-MM, CB-GV et MM-GV. La liste de liens peut être transformée en une tableau carré rempli de 1 (lien présent entre les individus) et de 0 (lien absent) : on parle de matrice d'adjacence (*adjacency matrix*). Par convention, si le sens de la relation importe, ce qui n'est pas le cas dans cet exemple, les origines sont disposées en lignes et les destinations en colonnes. La petite liste donnée plus haut devient alors :

	CB	MM	GV
CB	0	1	1
MM	1	0	1
GV	1	1	0

Ce type de tableau permet de mener à bien une analyse de réseau mais également d'autres méthodes statistiques type analyse des correspondances multiples (ACM). Si une liste de liens entre des individus est nécessaire pour faire de l'analyse de réseau, elle n'impose pas d'en faire et d'autres traitements sont possibles.

Il est d'ailleurs tout à fait possible de combiner les méthodes d'analyse : mener à bien une analyse de réseau peut permettre de générer de nouveaux attributs portant sur les individus, attributs pouvant ensuite servir pour une analyse statistique multivariée. Si les individus possèdent des coordonnées géographiques, il est possible d'utiliser des méthodes d'analyse spatiale et/ou de cartographier les relations étudiées. Une autre possibilité est de partir de la matrice d'adjacence pour mener à bien d'un côté une analyse de réseau et de l'autre une ACM pour comparer les résultats obtenus. L'analyse de réseau permet aussi de produire des visualisations devenues très prisées ces dernières années, ce dernier aspect est traité dans le chapitre 12.

Enfin, l'analyse de réseau peut favoriser la curiosité extra-disciplinaire dans la mesure où de nombreuses disciplines mobilisent ces méthodes. Il n'est pas rare lorsqu'on cherche à maîtriser une approche de parcourir des travaux de sociologues, de biologistes, de physi-

ciennes et/ou de géographes. Si cette ouverture disciplinaire est stimulante d'un point de vue intellectuel, elle peut être compliquée [1] dans la mesure où les traditions disciplinaires d'analyse de réseau se sont développées de manière relativement autonomes les unes par rapport aux autres (voir le chapitre suivant).

À l'intention des formatrices

L'un des principaux objectifs d'une formation d'initiation à l'analyse de réseau devrait être de permettre aux participantes de répondre à la question suivante : ai-je vraiment besoin de faire de l'analyse de réseau ? Il existe des milliers de jeux de données et de questions de recherche qui ne nécessitent pas cette approche. La question est particulièrement importante pour les personnes ayant de fortes contraintes de temps (doctorantes en milieu de thèse par exemple).

Il est important de préciser que d'autres types de méthodes quantitatives existent pour analyser des données relationnelles. Ainsi, la famille des modèles gravitaires développés en économie régionale et en géographie quantitative pour analyser des flux entre lieux peut permettre d'étudier certains types de liens. Nombre de modèles portant sur des relations entre espèces animales et/ou végétales développés en écologie permettent d'étudier des données relationnelles, notamment entre deux populations différentes. Les modèles épidémiologiques et plus généralement de diffusion sont également des outils permettant l'analyse de données relationnelles. On peut donc étudier des données relationnelles sans nécessairement mobiliser l'analyse de réseau. Inversement, on peut mobiliser l'analyse de réseau pour étudier des phénomènes qui ne traduisent qu'indirectement une relation (co-participation à un événement par exemple).

1.2 Quelques grandes questions de recherche

Vous avez votre population d'individus - au sens statistique du terme donc ça peut être n'importe quoi, des plantes, des personnes, des films, des mots, des entreprises, des pays, des licornes, etc. etc. - et vous avez défini un type de relations entre ces individus (ex. un lien entre deux mots clés signifie que ces deux mots clés ont été utilisé ensemble pour un article paru entre telle date et telle date dans un corpus de revues construit de telle et telle façon). L'analyse de réseau vous permet de répondre à des questions du type :

- comment qualifier le réseau ?
- certains individus occupent-ils une place particulière dans le réseau ?
- certaines relations sont-elles spécifiques ?
- les individus partageant certaines caractéristiques sont-ils plus susceptibles d'être en relation les uns avec les autres ?
- peut-on créer des sous-ensembles pertinents au sein du réseau ?
- quels sont les mécanismes susceptibles d'expliquer la structure du réseau étudié ?
- l'ajout ou la suppression de certains liens ou de certains sommets est-elle susceptible de modifier la structure du réseau ?

Les questions dépendent évidemment du type de données étudiées. Si je m'intéresse à des relations entre des enfants dans une classe de maternelle, je ne vais pas imaginer qu'un

1. La destruction en cours de l'enseignement supérieur et de la recherche est évidemment un frein majeur à la curiosité extra-disciplinaire. Si vous êtes précaire et que vous aspirez à entrer dans le monde de la recherche, il est essentiel d'être identifiée *d'abord* par les praticiennes de votre discipline : ce sont elles qui vous qualifient (CNU), ce sont elles qui éventuellement vous recruteront (comités de sélection, sections CNRS, etc.).

des enfants disparaisse subitement pour étudier l'impact sur mon réseau. Si j'étudie un réseau d'infrastructures ferroviaires, étudier l'impact de l'ouverture d'une nouvelle ligne à grande vitesse et de la fermetures de x lignes locales [1] serait au contraire bienvenu. Malgré ces différences en partie liées aux données et en partie disciplinaires, certaines questions restent classiques.

Qualifier le réseau étudié est une étape incontournable en analyse de réseau : certains indicateurs sont basiques (nombre d'individus, nombre de liens entre ces individus), d'autres plus élaborés. Les mesures permettent généralement de qualifier le réseau mais également de le rapprocher de modèles de réseau largement étudiés dans la littérature (ces modèles sont rapidement évoqués dans le chapitre suivant et détaillés dans la section 5.4). Le fait d'étudier tel ou tel type de réseau peut faciliter votre travail dans la mesure où il restreint de fait les mesures pertinentes [2].

Une des questions essentielles en analyse de réseau concerne la notion de centralité : y-a-t'il dans le réseau étudié des individus particulièrement importants ? Des individus qui à eux seuls créent en grande partie la structure du réseau (imaginez le réseau aérien français sans l'aéroport Roissy Charles de Gaulle, la structure du réseau internet sans Google ou un réseau de citations en sociologie sans Bourdieu) ? La notion de centralité a une telle importance que différents indicateurs ont été construits par la mesurer (voir la section Centralités du chapitre 5). Elle prend parfois un nom différent dans telle ou telle discipline mais le principe général reste le même.

Si la centralité permet de hiérarchiser les individus, certaines mesures visent à qualifier l'efficacité et la robustesse du réseau dans son ensemble. Qu'il s'agisse d'un réseau d'infrastructures (circulation des personnes, des biens et des services) ou d'un réseau de personnes (circulation de l'information), la structure du réseau permet-elle de remplir efficacement son rôle ? Si non, est-il possible d'identifier des micro-configurations qui expliquent la faible efficacité du réseau ?

Des notions et des concepts développés dans la sociologie quantitative nord-américaine ont eu un fort impact dans leur discipline et dans les disciplines voisines. Distinguer liens forts et liens faibles dans un réseau ou identifier des individus en position de *broker* [3] est fréquent quelles que soient les données étudiées ; ces deux notions sont développées dans la section 5.3.

Depuis les années 1930, la recherche de sous-groupes comprenant des individus fortement inter-connectés entre eux fait partie des questions de recherche fréquentes en analyse de réeau. Là encore, quelles que soient les données et la discipline, mettre en évidence différentes « communautés » est fréquent, ce qui ne signifie pas que ce soit toujours pertinent. D'autres méthodes de partition, essentiellement utilisée en *Social Network Analysis* (*SNA*) visent à rassembler les individus en fonction de leur position dans le réseau. Ces différentes méthodes sont présentées au chapitre 6.

La majeure partie des questions évoquées dans les lignes qui précèdent est d'ordre *descriptive*. La description des propriétés du réseau et des individus est indispensable mais elle est rarement suffisante : l'étape suivante consiste généralement à identifier les processus

1. Cet exemple est évidemment tout à fait fictif.

2. Certaines mesures n'ont pas de sens avec certains types de réseaux. Cela ne signifie pas qu'un type donné de réseau suppose *obligatoirement* telle ou telle démarche d'analyse de réseau. Il existe des canons disciplinaires, il n'est pas nécessairement utile de les appliquer au pied de la lettre.

3. Personne occupant une position d'intermédiaire entre deux autres personnes et susceptible de faire circuler ou de bloquer une information, une ressource, etc.

expliquant la structure du réseau. Si l'apparition des liens entre individus peut s'expliquer par une caractéristique commune, on vise alors à mettre en évidence des effets d'homophilie (*homophily*). Les réseaux amicaux chez les enfants montrent une forte homophilie de genre ; les réseaux amicaux chez les adultes une forte homophilie liée à la catégorie socio-professionnelle. On observe par ailleurs dans certains réseaux une très forte hiérarchie entre les individus : certains ont des milliers de liens et la grande majorité en ont très peu [1]. La mise en évidence des logiques d'avantage cumulatif, renommé attachement préférentiel (*preferential attachment*) en analyse de réseau, est devenue une démarche fréquente. Enfin, des familles de modèles statistiques visent à mettre en évidence et à expliquer les caractéristiques les plus saillantes des réseaux étudiés. Ces modèles étant plutôt adaptés à un public expert, ils ne seront que brièvement évoqués dans ce guide (chapitre 11) tout comme les modélisations graphiques qui sont rarement utilisées mais présentent pourtant des intérêts heuristiques évidents.

1. Les réseaux de citations scientifiques en sont un exemple caricatural, la majorité des articles n'étant *jamais* cités.

Chapitre 2

Histoires & disciplines

La lecture de ce chapitre n'a pas d'intérêt pratique immédiat et peut donc être considérée comme optionnelle. Elle peut cependant s'avérer utile si vous souhaitez franchir les limites méthodologiques de votre discipline à moindre coût. Ce chapitre est une esquisse et mériterait une approche historique plus rigoureuse.

Ce chapitre pourrait donner lieu à une jolie série d'ouvrages. L'objectif ici est plus modeste : il s'agit d'indiquer, pour une poignée de disciplines seulement (sociologie, géographie, écologie et physique), comment l'analyse de réseau s'est développée, quelles ont été les données utilisées ainsi que les principales questions de recherche. Les références bibliographiques, plus nombreuses que dans les autres chapitres, permettront aux personnes curieuses de creuser ces questions. La dernière section évoque les questions de traduction disciplinaire afin de vous permettre de naviguer plus facilement d'une discipline à l'autre.

2.1 Théorie des graphes et analyse de réseau

Dans tous les manuels d'analyse de réseau ou presque, vous trouverez la jolie histoire du mathématicien Euler qui, en 1736, se demande s'il est possible de se promener dans Kaliningrad (Königsberg à l'époque) en passant une fois et une seule sur chacun des ponts de la ville (la réponse est non). Euler pour résoudre son problème ayant symbolisé les rives par des points et les ponts par des lignes, il est souvent présenté comme l'inventeur de la théorie des graphes. La réalité est un petit peu plus compliquée et parler de théorie des graphes comme programme de recherche autonome avant 1930 [1] voire 1950 [2] est un anachronisme.

La théorie des graphes étudie des objets constitués d'un ensemble d'entités appelées sommets et d'un ensemble de liens entre ces entités. Il n'existe pas de notation standard mais l'une des plus courantes est $G = \{V, E\}$ où G désigne le graphe, V l'ensemble de sommets (V comme *vertex*, pluriel *vertices*) et E l'ensemble des liens (E comme *edges*). L'ensemble de sommets est fini et non vide, l'ensemble de liens est fini et éventuellement

1. Parution du premier manuel de Dénes König en allemand en 1936, *Theorie der endlichen und undendlichen Graphen*. Le contexte international a limité sa diffusion.

2. Réédition à New-York mais toujours en allemand du manuel de König et surtout parution en français en 1958 et traduction en anglais en 1962 de l'ouvrage de Claude Berge, *La théorie des graphes et ses applications*. Le manuel de König est traduit en anglais en 1990 seulement.

vide [1]. À la différence des approches disciplinaires évoquées par la suite, la notion d'attribut n'est pas pertinente et la représentation visuelle des relations n'est pas un enjeu.

Lorsqu'on étudie les premiers articles d'analyse de réseau et ce quelle que soit la discipline, on se rend vite compte que les emprunts à la théorie des graphes sont en fait réduits : le vocabulaire de base est généralement repris (graphe, liens, sommets) mais souvent adapté ou modifié, une poignée de mesures peuvent être recyclées si elles permettent de répondre à des questions de recherche et c'est en général à peu près tout [2]. Ceci n'est d'ailleurs pas étonnant : là où la théorie des graphes cherche à produire des énoncés aussi généraux que possible, les approches en sciences sociales visent au contraire à étudier un ou plusieurs réseaux spécifiques liés à un aspect précis du monde social. Pour le formuler autrement, il n'est guère utile de parcourir des manuels de théorie des graphes pour pratiquer l'analyse de réseau. Dans certains cas, il n'existe d'ailleurs pas le moindre lien avec la théorie des graphes et c'est notamment le cas de la sociométrie apparue aux États-Unis dans les années 1930.

2.2 Quand Jennings et Moreno créent la sociométrie

L'ouvrage de Jacob Moreno, écrit avec l'aide d'Helen Jennings (cette dernière n'étant pas créditée), *Who shall survive ?* paru en 1934 et disponible en ligne [3] est sans doute le premier ouvrage [4] proposant une démarche complète d'analyse de réseau, évoquant tant les concepts que la récolte des données, leur visualisation et les mesures possibles.

Les données relationnelles étudiées par Moreno et Jennings sont exclusivement interpersonnelles, au sein d'un groupe donnée, et la démarche privilégiée pour recueillir ces données est le questionnaire. Contrairement à ce qui sera privilégié par la suite, la sociométrie recueille systématiquement deux types de liens : les liens positifs (ex. j'apprécie cette personne, j'aime travailler avec cette personne) et les les liens négatifs (je n'apprécie pas cette personne, je ne souhaite pas partager le même dortoir).

Le recueil systématique de ce type de liens au sein d'un groupe permet notamment de mettre en évidence les logiques de groupes, d'identifier les personnes populaires et les personnes rejetées, de révéler les mécanismes d'homophilie (fait que des personnes partageant certaines caractéristiques communes s'apprécient davantage). Deux moyens principaux sont utilisés pour étudier les relations : le dessin des relations à l'aide de ce que Moreno nomme des sociogrammes, le calcul d'indicateurs statistiques [5].

Dans *Who Shall Survive ?* puis dans la revue *Sociometry* qu'il crée, Moreno accorde un importance extrême à la lisibilité des sociogrammes : tous ont la même légende (variation de forme, de couleur et de taille des sommets, variation de forme des liens), règles explicites de

1. Les premiers manuels de théorie des graphes portaient sur les graphes finis et infinis ; en sciences sociales, on ne travaille pas à ma connaissance sur des graphes infinis.

2. J'ai étudié les liens faibles entre analyse de réseau en géographie et théorie des graphes dans un article disponible en ligne (2022). Des approches comparables dans d'autres disciplines seraient utiles.

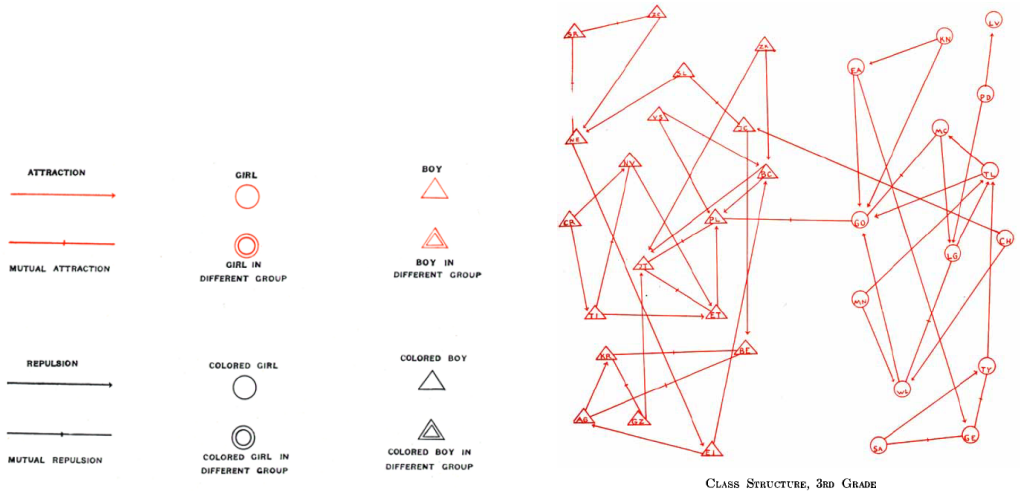
3. La réédition largement modifiée de 1953 est également disponible sur archive.org.

4. L'importance d'étudier les relations entre entités a été affirmée par de nombreuses autrices avant cette date mais sans que soit proposée une méthode d'analyse spécifique.

5. Moreno et Jennings présentent dans un formidable article paru en 1938 un ensemble de mesures permettant de caractériser ces relations ; une version française et commentée par Françoise Bahoken et moi de cet article est disponible dans la collection « textes » du groupe fmr.

construction[1]. La figure 2.2 montre un extrait de la légende et un extrait de sociogramme montrant une forte homophilie de genre à l'école primaire.

FIGURE 2.1 – Le sociogramme de Moreno (1936)



Extrait de Jacob L. Moreno, 1934, Who Shall Survive ?, *pp. 30 et 37.*

Cette approche, après une popularité impressionnante dans les années 1940[2], est tombée dans un relatif oubli. Certaines autrices invoquent des facteurs personnels, faisant de Moreno une personnalité pathologique qui peu à peu fait fuir toutes les chercheuses sérieuses. Il est possible de supposer que des contraintes institutionnelles et techniques ont également joué : Moreno n'a pas de poste à l'université et ne peut donc pas former d'étudiantes ou diriger de thèses ; la quantification souhaitée par Jennings et Moreno repose en partie sur du calcul matriciel très contraignant à réaliser sans ordinateur.

Il est intéressant de noter que, contrairement à d'autres disciplines où la domination masculine est flagrante[3], la sociométrie comptait une proportion notable d'autrices. Les thèmes de recherche (sociabilité des enfants, sociabilité familiale) expliquent sans doute en partie cette relative mixité.

2.3 L'analyse de réseaux sociaux : Manchester *vs* Harvard

L'analyse de réseaux sociaux (*Social network analysis*) apparaît dans les années 1950-1960 en Angleterre puis aux États-Unis. Si l'on suit la grille de lecture proposé par Michael Eve (2002), on peut distinguer deux approches sensiblement différentes.

Les autrices anglaises basées à Manchester ont généralement des formations d'anthropologues et favorisent donc les approches fondées sur un travail de terrain long et minutieux, souvent dans des terrains coloniaux ou post-coloniaux. Parler d'école de Manchester est peu approprié dans la mesure où les démarches proposées par les différentes autrices ne

1. À l'époque, les figures sont évidemment dessinées manuellement mais les règles proposées par Moreno (placer les individus les plus connectés au centre, éviter le chevauchement de liens, etc.) restent aujourd'hui encore des règles utilisées par la plupart des algorithmes de visualisation de réseaux.

2. Consulter l'index des autrices de la revue *Sociometry* permet de repérer la plupart des grands noms de la sociologie étasunienne de l'époque.

3. Dans la géographie quantitative ou l'école de Harvard, les rares femmes apparaissent uniquement comme épouses et/ou secrétaires.

donneront pas lieu à la création d'une méthode stabilisée. Deux ouvrages au moins méritent d'être lus : Elisabeth Bott, *Family and Social Network*[1] et l'ouvrage dirigé par J. Clyde Mitchell, *Social networks in urban situations*.

Dans les deux cas, la démarche ethnographique prime et les autrices cherchent à étudier le réseau d'une personne ou d'une famille en multipliant les observations et les entretiens. Contrairement à l'approche sociométrique, il est rarement question d'observer un type de relations au sein d'un groupe donné mais plutôt de mettre en évidence les types de relations entretenues par un individu ou une famille. Dans le premier cas, on parle d'approche par réseau complet[2], dans le second, d'approche égo-centrée ; cette distinction est développée au chapitre 4.

FIGURE 2.2 – « École de Manchester » et multiplexité des liens

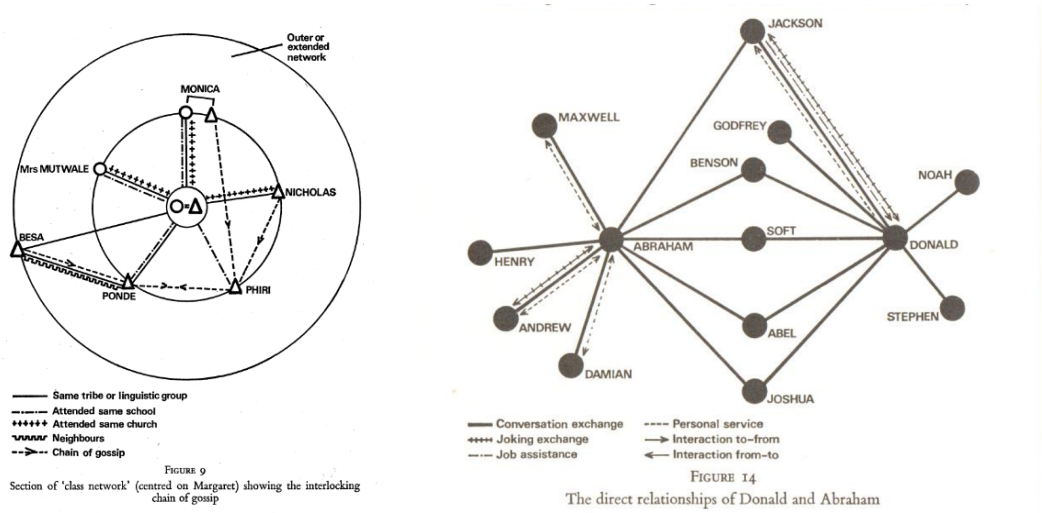


FIGURE 9
Section of 'class network' (centred on Margaret) showing the interlocking chain of gossip

FIGURE 14
The direct relationships of Donald and Abraham

*Ces deux figures sont extraites de l'ouvrage dirigé par J. Clyde Mitchell (1969). À gauche, A.L. Epstein montre les liens multiples révélées par les commérages (*Gossip, Norms and Social Networks*, pp.117-127. La figure se trouve page 127). À droite, B. Kapferer montre les liens directs et indirects entre deux ouvriers (*Norms and Manipulations of Relationships*, pp. 181-244 ; la figure se trouve page 216).*

L'analyse de réseaux sociaux développée par Harrison White (thèse de physique au MIT en 1955, thèse de sociologie à Princeton en 1960) à Harvard à partir des années 1960 a elle permis la création d'un domaine de recherche au sein de la sociologie quantitative, la *Social network analysis* (*SNA*). White crée en 1965 le premier cours de licence d'analyse de réseau et dirige une série de thèses (Bonacich, Granovetter, Levine, Wellman, etc.[3]). Un élément crucial explique le succès de cette approche : la création de logiciels qui uniformisent tant les traitements que le type de données pouvant être analysées.

La plupart des travaux issus de l'école de Harvard étudient des réseaux complets (*i.e.* ensemble de relations au sein d'une population donnée) relatifs soit à des groupes d'individus, soit à des liens entre personnes et conseils d'administration (ce qui est appelé les

1. L'ouvrage paraît en 1957, est réédité en 1971 et régulièrement réimprimé depuis. La version de 1971 est conseillée car on y trouve une passionnante postface intitulée *Reconsiderations* où l'autrice discute les choix méthodologiques et terminologiques de l'ouvrage.
2. Le terme réseau complet peut également désigner le résultat d'une mesure, voir chapitre 5.
3. Pour plus d'informations sur White, voir l'ouvrage de Freeman, 2004, p. 121-128.

interlocks). Les principales questions de recherche concernent la recherche de partitions d'un réseau, la centralité des individus, leur position stratégique ; la plupart des mesures mises au point par l'école de Harvard restent utilisées aujourd'hui et ont été importées dans plusieurs autres disciplines (degré, intermédiarité, voir chapitre 5), notamment en géographie, en archéologie (Collar *et al.*, 2014) et en histoire (Lemercier, 2005). Inversement, des méthodes fréquemment utilisées dans les années 1970-1980 (*blockmodel* notamment) ont moins percolé en dehors de la sociologie quantitative.

Outre ces deux approches devenues classiques, on peut considérer avec Michel Grossetti que l'étude des chaînes relationnelles est une forme d'analyse de réseau. Dans ce cas, ce qui importe est le nombre de liens nécessaires pour faire circuler une ressource d'un sommet à un autre [1].

2.4 Réseaux d'infrastructures et analyse de flux

Au début des années 1960, les géographes quantitativistes nord-américains développent leur programme de recherche d'analyse de réseau, notamment autour de William Garrison à l'université de Washington [2]. Les réseaux étudiés sont essentiellement des réseaux d'infrastructures pouvant être modélisés par des graphes dits planaires - un graphe est planaire quand il peut être projeté sur un plan sans qu'aucun lien ne se croise.

Piochant dans la théorie des graphes, dans les études de communication et, dans une moindre mesure, dans les travaux d'optimisation des flux, les géographes construisent en quelques années un programme de recherche original et n'ayant à peu près aucun lien ni avec la sociométrie ni avec l'analyse de réseaux sociaux. Les questions de recherche sont également différentes dans la mesure où l'efficacité, l'accessibilité des sommets et l'évolution du réseau sont au cœur des préoccupations. Enfin, étudiant des objets mathématiques légèrement différents (graphe planaire ; sommets et liens ayant des coordonnées géographiques), les outils développés pour les analyser sont également différents.

En ce qui concerne l'étude des flux (liens orientés et d'intensité variable entre les sommets), deux types d'approche se développent : d'une part des méthodes permettant de simplifier les flux (Nystuen et Dacey, 1961 [3]), d'autre part des modèles dits gravitaires issus de la *social physics* - ces derniers ne sont pas abordés dans ce manuel.

2.5 L'analyse des réseaux écologiques

La vision relationnelle des écosystèmes est ancienne et, dès la fin du XIX^{e} siècle, les représentations graphiques de systèmes trophiques (qui mange qui) sont relativement courantes ; il faut cependant attendre la parution de *Animal Ecology* d'Elton en 1927 pour que l'analyse intègre ces relations (Bersier, 2007). Si les figures évoquent des réseaux, peu de relations peuvent être modélisées sous forme de graphe. L'une des rares relations où cette modélisation est possible concerne les relations insectes pollinisateurs - plantes et une poignée d'indicateurs est créée pour étudier ce type de relations.

1. Un dossier sur les chaînes relationnelles est en cours de publication dans la revue *ARCS - Analyse de réseaux pour les sciences sociales*.

2. J'ai abordé plus longuement cet aspect dans l'article déjà mentionné plus haut, cette section en est une version très réduite.

3. La collection « textes » du groupe fmr propose des versions bilingues et commentées de plusieurs articles fondateurs de l'analyse de réseau, en géographie et dans d'autres disciplines. L'article de Nystuen et Dacey est disponible dans cette collection (Bahoken et Beauguitte, 2021).

Les modèles mathématiques utilisés, comme le modèle proie-prédateur de Lotka-Volterra, sont fondés sur des équations différentielles (Ings et Hawes, 2018). Si les figures évoquent des réseaux et si la formalisation mathématique est importante, il est pourtant difficile de parler d'analyse de réseau en écologie avant les années 2000 et l'impact des travaux des physiciennes [1]. Il existe cependant des exceptions et l'on peut citer par exemple les travaux de l'écologue Sade portant sur les interactions entre grands singes : si ses premiers articles montrent les interactions (figure 2.3) sans guère les mesurer, sa production montre une utilisation croissante des méthodes et mesures issues de la *SNA*.

FIGURE 2.3 – Qui toilette qui (Sade, 1965)

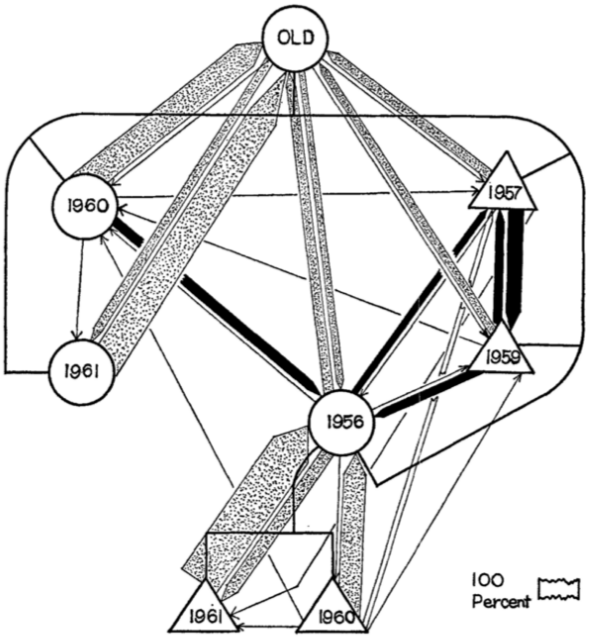


Fig. 3 Grooming relations in genealogy 4 from March 21, 1961 through December 25, 1961. Drawn from the data in table 3. Explanation in text.

Extrait de D.S. Sade, 1965, Some aspects of parent-offspring and sibling relations in a group of rhesus monkeys, American Journal of Physical Anthropology*, 23(1) :1-17. La forme des sommets permet de différencier le sexe (cercle, femelle ; triangle, mâle), l'année indiquée est l'année de naissance, les liens sont teintés selon la proximité familiale du couple de singes (filiation, cousinage, autre) et leur épaisseur varie selon la fréquence des contacts.*

2.6 Quand les physiciennes bouleversent le paysage

Pour une raison qui n'est pas tout à fait claire [2], les physiciennes qui s'intéressaient peu à l'analyse de réseau jusque-là ont surgi brutalement dans le paysage à la fin des années

1. Pourquoi alors cette section ? Tout simplement parce que j'ai prétendu le contraire durant des années, me laissant abuser par le caractère relationnel des données, les représentations graphiques et la formalisation mathématique de certains articles. Il m'a fallu quelques années pour comprendre que la combinaison des trois peut produire autre chose que de l'analyse de réseau (de l'optimisation linéaire par exemple).

2. Contraintes budgétaires et nécessité de trouver de nouveaux marchés académiques ? Augmentation des capacités informatiques ? Disponibilité et taille croissantes des données relationnelles ? Popularité croissante de l'idéologie saint-simonienne, reprise par une star internationale de la sociologie contemporaine dans un pensum en trois volumes que je ne citerai pas, dans une société où tout serait réseau ? La thèse de Li Vigni sur le développement des instituts de *complex studies* fournit quelques éléments de réponse (Li Vigni, 2018).

1990. Et, qu'elles le veuillent ou non, toutes les disciplines ont dû se positionner vis-à-vis de ces nouvelles arrivantes.

Deux articles parus à un an d'intervalle ont eu un impact majeur sur la production académique relative aux réseaux : l'article des mathématiciens Watts et Strogatz [1] paru dans *Nature* en 1998 et l'article des physicien.ne.s Barabási et Albert [2] paru l'année suivante dans *Science*.

Ces deux articles ont fait date pour plusieurs raisons. Ils proposaient de nouveaux modèles de réseaux, l'un dit des réseaux petit-monde (*small-world networks*), l'autre dit des réseaux sans échelle (*scale-free networks*) - les caractéristiques de ces modèles sont détaillées au chapitre 5. Ils portaient ensuite sur des données de réseaux beaucoup plus volumineuses que celles étudiées auparavant. Ils ouvraient enfin des pistes d'analyse stimulantes et les travaux de physiciennes concernant les réseaux se sont multipliés très rapidement sans s'intéresser le moins du monde aux décennies de travaux existant en analyse de réseau dans d'autres disciplines : lire à ce sujet les pages très drôles de Watts intitulées *Here come the physicists...* dans son ouvrage grand public [3] *Six Degrees* (2003, pp. 37-41).

N'hésitant pas à envahir tous les terrains et à manipuler toutes les données possibles, les physiciennes ont souvent suscité des réactions de défiance voire d'hostilité déclarée chez les analystes de réseau en sciences sociales [4]. Certaines personnes ont au contraire utilisé ces approches pour renouveler leurs travaux : on compte dans x disciplines des centaines voire des milliers d'articles où l'autrice se demande si son réseau est *small-world* ou *scale-free* - l'article étant publié, la réponse est généralement oui, au moins pour le *small-world*. Et le réseau dit sans échelle, réseau où quelques individus monopolisent la majorité des liens, n'a pas eu de mal à trouver sa place dans toutes les disciplines où certains phénomènes présentent des distribution très hiérarchisées (lexicométrie, géographie urbaine, économie, bibliométrie, etc.).

Les physiciennes ont apporté beaucoup à l'analyse de réseau : une rigueur mathématique certaine, deux nouveaux modèles, des méthodes permettant de traiter de gros volumes de données et le dynamisme de l'analyse de réseau ces vingt dernières années leur doit énormément. Et contrairement à des disciplines où les chercheuses s'obstinent à publier dans des revues où accéder au moindre article coûte quelques dizaines de dollars, les physiciennes ont l'intelligence de mettre *a minima* tous leurs *preprints* en accès libre.

2.7 Des traditions à la traduction disciplinaire

Jusqu'à la fin des années 1990, le paysage de l'analyse de réseau est relativement stable : les sociologues étudient les réseaux sociaux, les géographes les réseaux spatiaux, les archéologues piochent chez les unes ou les autres en fonction de leurs besoins, chaque communauté disciplinaire utilise son vocabulaire, ses méthodes et ses logiciels propres pour résoudre ses

1. Watts est actuellement *computational social scientist* à l'université de Pennsylvanie ; Strogatz est toujours professeur de mathématiques appliquée à la *Cornell University* où il a dirigé la thèse de Watts, thèse soutenue en 1997.

2. Le premier est le directeur de thèse de la seconde : Albert Réka est désormais professeure de physique et de biologie à l'université d'État de Pennsylvanie, Barabási dirige le *Center for Complex Network Research* à la *Northeastern University* de Boston.

3. Non seulement les physiciens publient beaucoup mais ils produisent ensuite rapidement la version grand public - Barabási a fait exactement la même chose avec l'ouvrage *Linked* paru en 2002. L'ouvrage de Watts est à la fois honnête et stimulant ; celui de Barabási est embarrassant.

4. Ces réactions étant plus souvent exprimées oralement qu'à l'écrit, elles laissent peu de traces.

questions de recherche spécifiques puis publie dans ses revues disciplinaires[1] et le dialogue entre praticiennes de disciplines différentes est rare. L'une des conséquences de ce relatif isolement est une forte complexité terminologique : un même indicateur peut avoir différents noms, un même terme peut désigner des indicateurs différents. Par exemple, le rapport entre le nombre de liens présents et le nombre de liens possibles dans un réseau est appelé densité en analyse des réseaux sociaux, indice gamma en géographie et connectance en écologie (chapitre 5). Le terme de densité en géographie qualifiera lui le rapport entre la longueur des réseaux d'infrastructures et la surface du territoire étudié.

L'arrivée des physiciennes a dans un premier temps accru cette confusion terminologique dans la mesure où elles n'ont pas pris la peine de regarder ce qui était produit depuis des décennies en sciences sociales[2] et ont pu proposer de nouveaux termes pour des indicateurs déjà bien connus, appeler *clustering coefficient* ce que les praticiennes de la *SNA* nommaient depuis des décennies la transitivité par exemple.

Il est délicat de décrire le paysage scientifique contemporain de l'analyse de réseau et les lignes qui suivent au mieux sont une proposition de lecture. L'intérêt des physiciennes pour l'analyse de réseau ne s'est pas démenti ces vingt dernières années et leur production reste pléthorique. Les collaborations avec des thématiciennes issues d'autres disciplines existent mais donnent lieu à peu de publications. La *SNA*, sûre de ses méthodes et ayant ses propres espaces (conférences, revues), semble avoir été peu impactée par les travaux des physiciennes. La géographie quantitative semble avoir emprunté davantage de méthodes, tant à la *SNA* qu'aux physiciennes, mais la majorité des travaux de géographes en analyse de réseau continue à porter sur des réseaux d'infrastructures généralement planaires. L'histoire a été peu impactée par les approches des physiciennes, la taille des corpus et le problème des données manquantes étant des freins méthodologiques forts[3]. Et je me prononcerai peut-être sur l'écologie[4] ou l'archéologie quand j'aurais pris le temps de lire un peu plus...

À l'intention des formatrices

Il est rare d'aborder ces sujets lors d'une formation courte à l'analyse de réseau. Prendre quelques minutes pour expliquer ce que sont les réseaux petit-monde et sans-échelle me paraît pourtant indispensable dans la mesure où ces modèles ont été employés ensuite - parfois à tort et à travers mais ce n'est pas le sujet - dans la plupart des disciplines. Il est également utile de signaler que presque tous les manuels francophones et anglophones sont monodisciplinaires (sociologie le plus souvent).
Contrairement à ce qui est fait ici, il est préférable de présenter réseaux petit-monde et réseaux sans échelle *après* avoir abordé les principales mesures...

1. Il y a bien entendu des exceptions, ce panorama est volontairement simplifié.
2. Dans *Six degrees*, Watts explique très honnêtement qu'avec son directeur de thèse, ils avaient deux options : lire ce qui existait déjà et donc perdre des mois ou publier de suite ; ils choisissent la deuxième option et leur article déjà cité a un impact immédiat.
3. La faible formation aux méthodes quantitatives est sans doute un facteur explicatif supplémentaire non négligeable.
4. L'ouvrage coordonné par Cédric Sueur (2015) est passionnant car il rassemble trois types de contributions : les contributions orthodoxes d'analyse de réseau en écologie, celles influencées par la *SNA* et celles influencées par la physique.

Chapitre 3

Graphe et réseau : principes et vocabulaire de base

Ce chapitre tente d'être simultanément complet et synthétique. Il n'est pas utile de retenir toutes les définitions mais le consulter à l'occasion pour vérifier que vous savez qualifier correctement un réseau donné peut être utile, notamment pour échanger avec vos collègues. La quasi totalité des logiciels et des très bons manuels [1] étant en anglais, les termes anglais sont systématiquement indiqués en italiques et entre parenthèses.

> *À l'intention des formatrices*
>
> Il est sans doute temps d'abandonner le diaporama (si vous en utilisez un) et le cours magistral, de vous munir d'un feutre et de dessiner, de faire dessiner, de faire remplir des matrices avec des 0 et des 1, etc. Dessiner des réseaux très simples (orientés ou non, avec boucle ou simple, avec isolés) puis faire écrire les listes de liens et les matrices d'adjacence correspondantes est un bon moyen pour faire assimiler le vocabulaire et faire comprendre les enjeux liés à la mise en forme des données. Vous gagnerez du temps quand vous expliquerez ce qu'attendent les logiciels ou quand vous aurez des questions des participantes à propos de leurs propres données.

3.1 Des points, des lignes et des chemins

Un graphe est un objet mathématique formé d'un ensemble fini et non vide de points et d'un ensemble fini et éventuellement vide de liens entre ces points ($G = \{V, E\}$ [2]). Le nombre de sommets est appelé l'ordre (*order*) du graphe ; le nombre de liens la taille (*size*) du graphe. Les sommets (*vertex*, pluriel *vertices*) sont également appelés nœuds, points ou acteurs (*node*, *point*, *actor*). Les liens (*edges*) sont aussi appelés relations, arcs ou arêtes (*relation*, *arc*, *edge*). Les deux derniers termes sont employés lorsqu'on souhaite différencier les liens dits orientés (un lien du sommet v_1 vers le sommet v_2 n'implique pas l'existence

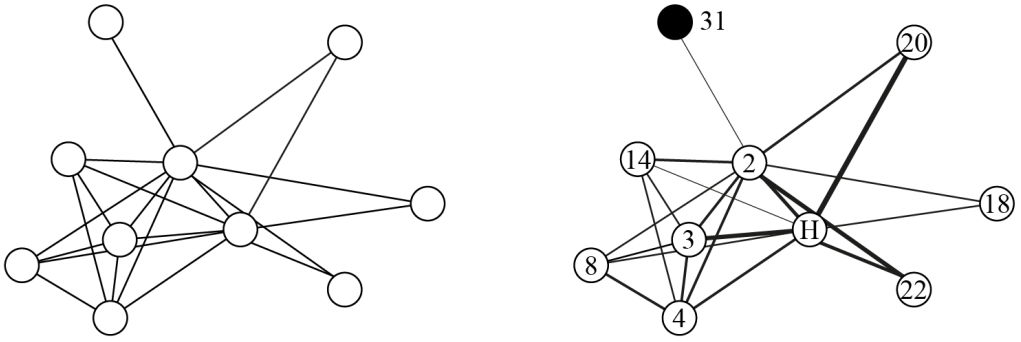
1. Les manuels francophones d'analyse de réseaux sociaux sont de bonne qualité et se complètent bien les uns les autres, je pense notamment à la paire Lazega (2014) - Mercklé (2011), mais leur format réduit ne leur permet pas de lutter contre le manuel de Wasserman et Faust (1994, une mise à jour s'impose) ni contre celui de Newman (2018). Ces deux ouvrages sont indispensables à qui souhaite utiliser sérieusement l'analyse de réseau ; le manuel de Newman est nécessaire si l'on souhaite échapper à la seule *Social Network Analysis*. On les trouve gratuitement en ligne sans problème (zlibrary était votre ami, libgen l'est encore).

2. Dans l'ensemble de ce texte, les graphes sont notés G, les ensembles de sommets V (comme *vertices*) et les ensembles de liens E (*edges*). Un sommet donné sera noté v, un lien précis e.

d'un lien de v_2 vers v_1) - on parle alors d'arcs - et les liens non orientés (tout lien entre v_1 et v_2 suppose l'existence d'un lien entre v_2 et v_1) - on parle alors d'arêtes.

L'analyse de réseau pour être utile en sciences sociales nécessite des éléments supplémentaires : les sommets sont systématiquement pourvus d'attributs (*a minima* un nom) et les liens peuvent également être pourvus d'attributs. La figure 3.1 montre un graphe (à gauche) et un réseau (à droite).

FIGURE 3.1 – Graphe *vs* réseau



L'objet mathématique à gauche est constitué d'un ensemble de sommets et d'un ensemble de liens entre ces sommets, il s'agit d'un graphe et il n'y a pas besoin d'attributs pour l'étudier ; le réseau à droite a exactement la même forme mais les sommets sont nommés et de deux couleurs différentes. Les liens sont plus ou moins épais, ce qui suggère des relations plus ou moins intenses.
Extrait du fameux jeu de données dit Zachary Karate Club, *tiré de l'article de l'anthropologue Wayne Zachary (1977) sur les relations au sein d'un club de karaté.*

Un chemin (*path*) est une suite de liens entre deux sommets. Le plus court chemin (*shortest path*) désigne le ou les chemins entre deux sommets comprenant le minimum de liens. La longueur d'un chemin entre deux sommets correspond au nombre de liens sur ce chemin ; s'il s'agit d'un plus court chemin, on parle parfois de distance géodésique (*geodesic distance*, voir figure 3.2). Un chemin comprenant au moins trois liens et dont le sommet de départ et le sommet d'arrivée sont identiques est appelé un cycle [1] (terme identique en anglais).

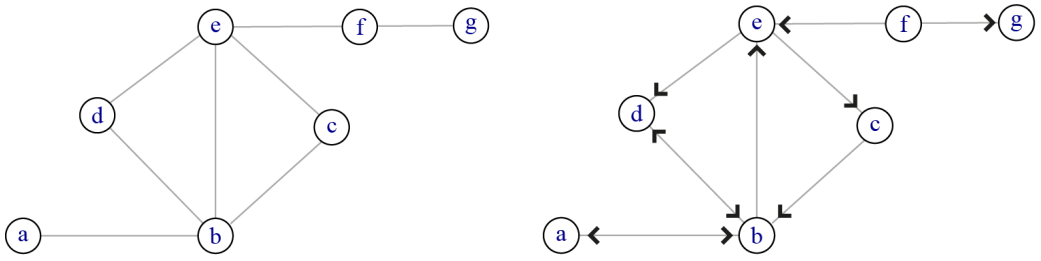
L'ensemble des sommets liés par au moins un chemin s'appelle une composante ou composante connexe (*component*). Si tous les sommets d'un graphe appartiennent à la même composante, on parle alors de graphe connexe (*connected graph*). S'il n'existe pas de chemin entre certains sommets, le graphe est alors non connexe et il est composé de deux composantes ou plus. Il arrive qu'un sommet ne soit connecté à aucun autre, on parle alors d'isolé (*isolate*). Lorsqu'un réseau est connexe si et seulement si on ne tient pas compte de l'orientation des liens (figure 3.2, réseau de droite), on parle de réseau faiblement connexe (*weakly connected*) ; si un réseau est connexe, y compris lorsqu'on prend en compte l'orientation des liens, on parle de réseau fortement connexe [2] (*strongly connected*). La distance entre deux composantes connexes est considérée par convention comme infinie.

1. La théorie des graphes distingue les cycles dans les graphes non orientés et les circuits dans les graphes orientés ; la distinction a peu d'intérêt en sciences sociales.

2. Wasserman et Faust (1994) distinguent également les *uniterally* et les *recursively connected graphs* selon que les mêmes liens sont empruntés dans un sens et dans l'autre dans les graphes orientés (p. 132).

Le caractère connexe ou non du réseau étudié a des conséquences importantes d'un point de vue méthodologique : les indicateurs basés sur des calculs de plus courts chemins (cf chapitre 5) doivent être interprétés avec précaution lorsque le réseau étudié comprend plusieurs composantes connexes.

FIGURE 3.2 – Chemin, distance et connexité



Le réseau de gauche est connexe : il est possible en partant d'un sommet de rejoindre n'importe quel autre sommet. On peut par exemple rejoindre a et e avec les liens {ab,bd,de}, ce qui donne un chemin de longueur 3. Le plus court chemin entre ces deux sommets est de longueur 2 {ab,be}. Il peut exister plusieurs plus courts chemins entre deux sommets : entre d et c, les deux chemins de longueur 2 {de,ec} et {db,be} sont des plus courts chemins.

Si les liens sont orientés, comme c'est le cas à droite, il est beaucoup plus difficile d'obtenir un réseau fortement connexe : le nombre de composantes peut devenir élevé, même avec un nombre de sommets réduit, et la longueur des chemins augmente. Ici, les sommets {abcde} forment une composante connexe, f et g sont isolés. Le plus court chemin entre e et c est de longueur 1 ({ec}) ; le plus court chemin entre c et e est lui de longueur 2 ({cb,be}). La plupart des logiciels ne prennent pas en compte l'orientation éventuelle des liens pour calculer les indicateurs basés sur les plus courts chemins.

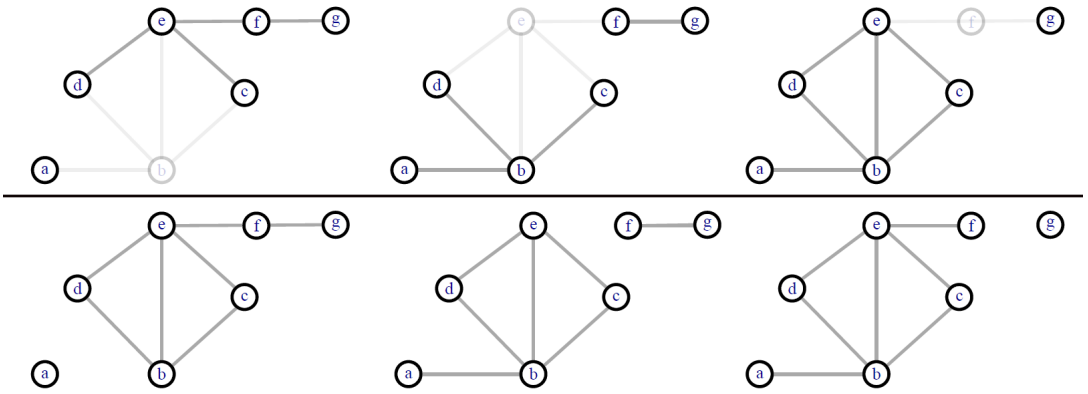
Certains sommets et certains liens jouent un rôle particulier dans la connexité des réseaux : leur suppression augmente le nombre de composantes. Quand il s'agit d'un sommet, on parle de point d'articulation (*articulation point*, *cutpoint* ou *cut-vertex*) ; quand il s'agit d'un lien, on parle d'isthme (*bridge*). Dans le réseau à gauche de la figure 3.2, b, e et f sont des points d'articulation (la suppression d'un sommet entraîne la suppression des liens adjacents à ce sommet). Les liens $\{ab\}$, $\{ef\}$ et $\{fg\}$ sont des isthmes (un sommet isolé est une composante d'ordre 1 - cf figure 3.3). Un nombre élevé de points d'articulation et/ou d'isthmes signale un réseau potentiellement vulnérable.

3.2 Qualifier un réseau d'après ses liens

Le type de liens dans un réseau peut permettre de catégoriser ce dernier. Si les liens ont une direction et qu'un lien de v_1 vers v_2 n'implique pas nécessairement un lien de v_2 vers v_1, on parle de graphe orienté (*directed graph* ou *digraph*) ; dans le cas contraire, on parle de graphe non orienté (*non directed graph*). Il est cependant possible de ne pas prendre en compte la direction des liens et de traiter un graphe orienté comme s'il était non orienté [1]. Il arrive également que les données disponibles ne permettent pas de connaître la direction

1. Il est *toujours* possible de transformer ses données, et donc ses réseaux, pour les adapter à ses questions de recherche et/ou aux méthodes disponibles, il suffit simplement de documenter les transformations effectuées.

FIGURE 3.3 – Points d'articulation et isthmes



En haut, suppression des points d'articulation (grisé) et conséquences sur la connexité ; en bas, suppression des isthmes. Lorsqu'un sommet est adjacent à un seul lien, ce lien est nécessairement un isthme.

---

initiale. Enfin, certains logiciels ignorent l'éventuelle orientation des liens pour le calcul de certains indicateurs (chapitre 5).

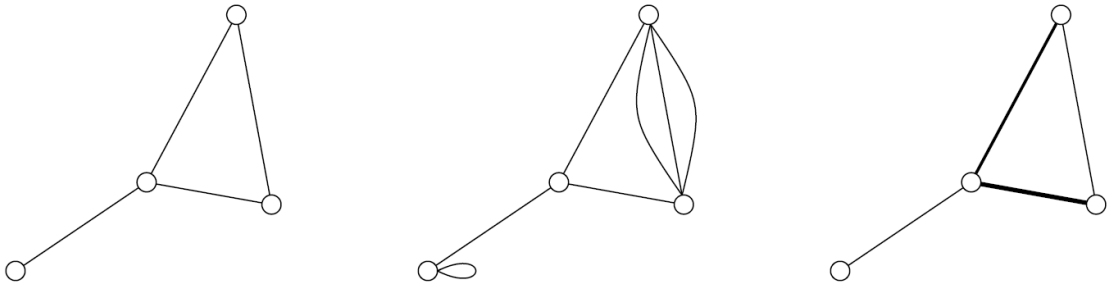
> **Exemple 1** : les liens d'amitié sur le réseau social Facebook sont au départ orientés : une personne demande à une autre si elle peut être ajoutée comme amie. Cette information est cependant perdue dès que la demande est acceptée : vous êtes amie avec telle personne mais il n'est pas possible de retrouver qui a fait la demande. Dans ce cas, la structure même des données récoltées impose l'absence d'orientation des liens.

> **Exemple 2** : certains liens familiaux sont non orientés. Si je suis mariée à une personne, cette personne est nécessairement mariée avec moi (ou l'une des deux ment). À l'inverse, un lien de filiation est nécessairement orienté. En fonction de mes questions de recherche, je peux cependant supprimer cette orientation si elle n'est pas utile dans ma démarche.

Le lien entre deux sommets dans le réseau peut être de type présence (1) - absence (0) ; on a alors affaire à un réseau dit binaire ou booléen (*binary* ou *boolean*). Il arrive, notamment dans les enquêtes sociométriques, que les liens soient porteurs de signes positifs, indiquant généralement une préférence (j'aime telle personne), ou négatifs, exprimant généralement un rejet (je n'aime pas telle personne). On parle dans ce cas de réseau signé (*signed graph*). Les logiciels n'étant qu'exceptionnellement conçus pour analyser ce type de réseau, il est rare de les garder sous cette forme et les réseaux signés sont à peine évoqués dans ce petit guide pratique. Enfin, le lien peut être porteur d'une intensité ; on parle dans ce cas de réseau valué ou pondéré (*weighted graph*).

> **Exemple** : j'étudie les données du commerce international. Je peux choisir de les modéliser sous forme de réseau entre États. Les liens sont orientés (la France exporte tant vers l'Allemagne, l'Allemagne exporte tant vers la France) et valués. Ce réseau peut être transformé de multiples façons afin d'obtenir un réseau non orienté (si je m'intéresse au volume des échanges et non à leur

FIGURE 3.4 – Types de graphes



Le graphe de gauche est un graphe simple. Celui du centre présente à la fois une boucle et des liens multiples. À droite, l'épaisseur des liens suggère une intensité de relations différentes entre les sommets ; il s'agit d'un graphe valué. Ces trois graphes sont non orientés.

---

direction) et/ou un réseau binaire (je justifie un seuil permettant de conserver la plus grande partie du flux tout en simplifiant l'analyse).

On appelle graphe simple (*simple graph*) un graphe qui ne comprend ni boucle (lien d'un sommet vers lui-même, *self-tie* ou *loop* en anglais) ni liens multiples (présence de plusieurs liens entre les mêmes sommets). S'il existe des liens multiples, on parlera le plus souvent de graphe multiplexe (*multigraph*) ou multi-couches[1] (*multi-layers*) (figure 3.4). Le terme de réseau complexe (*complex network*), souvent utilisé dans les travaux des informaticiennes et physiciennes, désigne dans l'immense majorité des cas un réseau simple de grande taille (*a minima* plusieurs milliers de liens et de sommets).

Quel que soit l'aspect du monde social, passé ou présent, que vous étudiez, il est rare qu'un seul type de relations existe entre vos individus : choisir de modéliser ces relations par un réseau simple est une simplification, souvent nécessaire et utile, de la réalité et il est prudent dans s'en rappeler lorsqu'on commente ses résultats. Les méthodes d'analyse des réseaux multiplexes étant encore aujourd'hui en grande partie exploratoires, elles sont abordées dans le chapitre 8.

Exemple 1 : j'étudie les pratiques de citations dans une discipline donnée. Je crée un lien de l'autrice a vers l'autrice b quand un travail de a cite un travail de b. Comme tous les réseaux de citations, mon réseau est orienté : a cite sa directrice de thèse mais cette dernière ne cite pas a, a cite des autrices décédées, etc. La probabilité que les autrices se citent elles-mêmes est forte, ce qui créent des boucles dans mon réseau. Il est également probable que je puisse valuer mon réseau dans la mesure où a cite sans doute plusieurs travaux d'une chercheuse de sa discipline. Je peux bien entendu choisir d'ignorer les boucles si la pratique de l'auto-citation ne fait pas partie de mes questions de recherche. Je peux aussi négliger la valuation des liens si je considère que le critère pertinent est la citation *vs* la non-citation. Aucune solution n'est meilleure qu'une autre : tout dépend de ce que j'étudie. Et tant que cette dernière question n'est pas

1. La terminologie est en pratique beaucoup plus confuse : Kivelä *et al.* dans un article de 2014 répertorient 26 termes différents désignant peu ou prou le même type d'objet. Le terme de pseudographe désignant les graphes avec boucles et liens multiples n'est quasiment jamais utilisé en sciences sociales.

totalement fermée, j'ai tout intérêt à garder les données les plus complètes possibles (chapitre 4).

Exemple 2 : dans un article paru en 2017 dans la revue *Cybergeo*, Sandrine Berroir *et al.* étudient les relations entre les villes françaises (sommets) et elles prennent en compte sept types de liens différents (navettes domicile-travail, offre aérienne et/ou tgv, partenariats scientifiques, etc.) ; elles obtiennent un réseau multiplexe où certaines relations sont orientées et d'autres non, toutes sont valuées. Il est bien entendu possible de transformer ce réseau multiplexe en partie orienté afin de créer un réseau simple. Les autrices de l'article créent un lien synthétique entre deux villes en trois étapes : discrétiser chaque flux en cinq classes et attribuer un score de 0 (20% des flux les plus faibles) à 4 (20% des flux les plus intenses) ; sommer ces scores pour créer un indice synthétique mesurant « l'intensité globale des liens interurbains » ; sélectionner les paires de villes où le score est maximal (4) pour au moins 3 des 7 flux. D'autres critères pouvaient tout à fait être choisis pour cette transformation ; il était également possible de garder la structure multiplexe du réseau pour l'analyser (chapitre 8).

Il est possible, à partir des données relationnelles que vous avez récoltées et mises en forme, de construire un réseau et le type de liens permet de qualifier ce réseau. Gardez cependant à l'esprit que ce réseau peut être modifié. On a tout à fait le droit de ne pas considérer l'orientation ou l'intensité des liens si ces deux critères n'ont pas d'importance pour nos questions de recherche. Il arrive également que la transformation soit liée à nos lacunes (je ne sais pas comment analyser un réseau multiplexe) ou au logiciel que nous utilisons (ce logiciel ne permet pas d'analyser un réseau multiplexe). Ce qui importe est d'expliciter et de justifier les transformations réalisées afin que vos lectrices puissent comprendre votre démarche.

3.3 Qualifier un réseau d'après ses propriétés

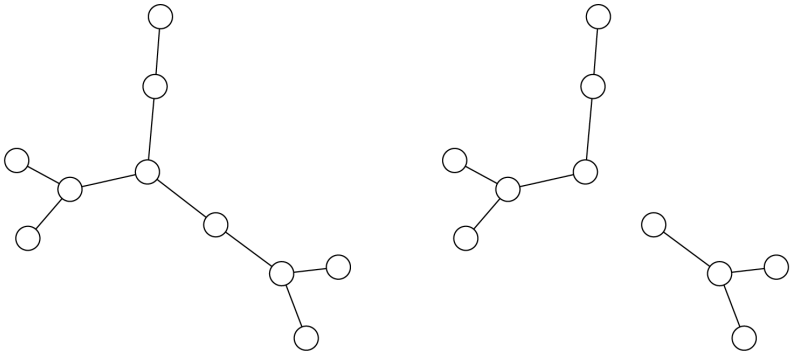
Les manuels d'analyse de réseau utilisent souvent pour expliquer les variations d'un indicateur des idéaux-types de réseaux à des fins pédagogiques. Ainsi on peut, pour évoquer la densité (nombre de liens présents divisé par le nombre de liens possibles), distinguer les graphes vides (aucun lien n'existe) et les graphes complets [1] (tous les liens possibles sont présents). Il va de soi que de tels réseaux ne présentent guère d'intérêt pour l'analyse.

Plus proche de certains réseaux, comme beaucoup de réseaux hydrographiques ou d'arbres généalogiques, les réseaux où aucun cycle n'est présent sont appelés des arbres s'ils sont connexes. Un arbre (*tree*) possède des propriétés intéressantes pour certaines méthodes d'analyse de réseau, notamment parce qu'un arbre de V sommets contient au minimum et au maximum V - 1 liens. Si le réseau acyclique n'est pas connexe, on parle de forêt (*forest* - figure 3.5).

Certains idéaux-types peuvent avoir des structures proches de celles des réseaux étudiés, par exemple les réseaux linéaires, les réseaux circulaires ou les réseaux en étoile (*star-graph*).

1. Attention à ne pas confondre les deux emplois du terme « réseau complet » : il peut s'agir de la démarche de recueil de données (ensemble des liens au sein d'un groupe donné, par opposition aux approches par réseau personnel) ou d'un réseau où tous les liens possibles sont présents.

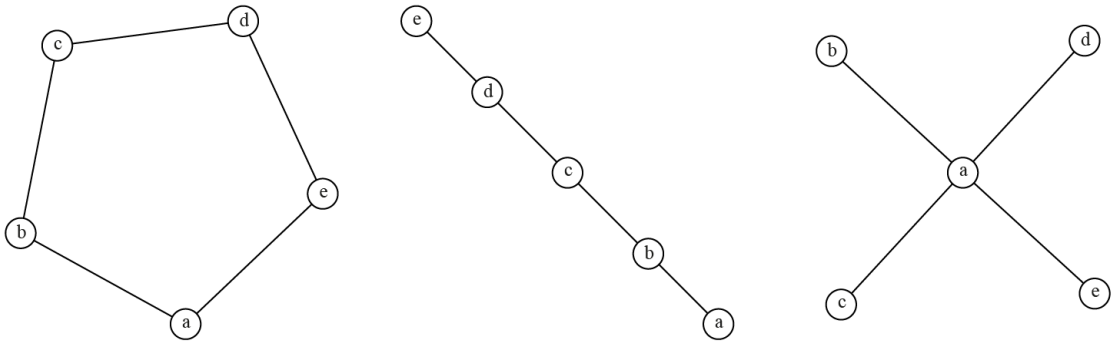
FIGURE 3.5 – Arbre et forêt



Le réseau à gauche est connexe (on peut trouver un chemin entre toute paire de sommets) et acyclique (il n'est pas possible de partir d'un sommet et d'y revenir en empruntant des liens différents). Il comprend 10 sommets et, pour être connexe, il doit avoir au minimum 9 liens. Ce minimum est également le maximum : tout ajout d'un lien sur la figure créerait un cycle. La suppression d'un seul lien crée un graphe non connexe, composé de 2 arbres, appelé forêt (à droite) : tout lien dans un arbre est un isthme.

La figure 3.6 présente ces trois formes : chacun des graphes possède le même nombre de sommets, ligne et étoile ont également le même nombre de liens.

FIGURE 3.6 – Cercle, ligne et étoile



5 sommets, 5 liens 5 sommets, 4 liens 5 sommets, 4 liens

Les réseaux étudiés peuvent avoir des similitudes formelles avec ces trois idéaux-types. Si l'on imagine que les sommets représentent des personnes et les liens des échanges d'information, on note que la situation devient de plus en plus inégalitaire à mesure que l'on se rapproche de l'étoile où le sommet central est un point obligé pour tout échange d'information. Par ailleurs, il est important de se rappeler que des réseaux ayant le même nombre de sommets et le même nombre de liens peuvent avoir des structures très différentes.

Il existe des structures de réseaux, issues des mathématiques ou de la physique, qui servent régulièrement de points de repère pour étudier nos propres réseaux (réseau aléatoire, réseau petit-monde, réseau sans échelle). Dans la mesure où ces modèles nécessitent de connaître certaines mesures, ils seront évoqués dans le chapitre 5.

À l'intention des formatrices

Débuter une formation par une liste de définitions de vingt ou trente termes n'est sans doute pas nécessaire. Il me semble pourtant utile que les personnes sachent rapidement comment qualifier et transformer un réseau. N'hésitez pas à multiplier les dessins et les exemples pour illustrer chaque cas. Il me paraît également essentiel d'insister sur le fait que le réseau est un objet construit pour la recherche : il n'est pas « par nature » orienté, valué ou que sais-je encore mais peut être transformé sans aucun problème, à condition évidemment d'indiquer explicitement les transformations opérées. Il faut cependant être claire avec les participantes : choisir de conserver certaines propriétés originelles du réseau (notamment pour les réseaux bimodaux et multiplexes) restreint le choix des logiciels adaptés pour leur analyse (chapitre 13).

Graphe unimodal, bimodal et hypergraphe

On parle de réseau uniparti, unimodal ou *one-mode* quand le réseau est constitué d'un ensemble de sommets et d'un ensemble de liens entre ces sommets. Lorsqu'aucun adjectif n'est présent, le réseau considéré est dans l'immense majorité des cas unimodal. La grande majorité des articles publiés en analyse de réseau concerne ce type d'objet et la majeure partie des mesures et des méthodes d'analyse a été construite pour des réseaux unimodaux.

Un graphe est dit biparti (*bipartite graph*) si son ensemble de sommets V peut être divisé en deux sous-ensembles disjoints V_1 et V_2 tel que chaque arête ait une extrémité dans V_1 et l'autre dans $V2$; on le note alors $G = \{V_1, V_2, E\}$. En sciences sociales, les termes de réseau biparti, bimodal ou *2-mode* sont souvent considérés comme synonymes [1]. J'utiliserai dans la suite de ce guide le terme bimodal qui me parait le plus adapté pour désigner les relations entre deux ensembles différents de sommets.

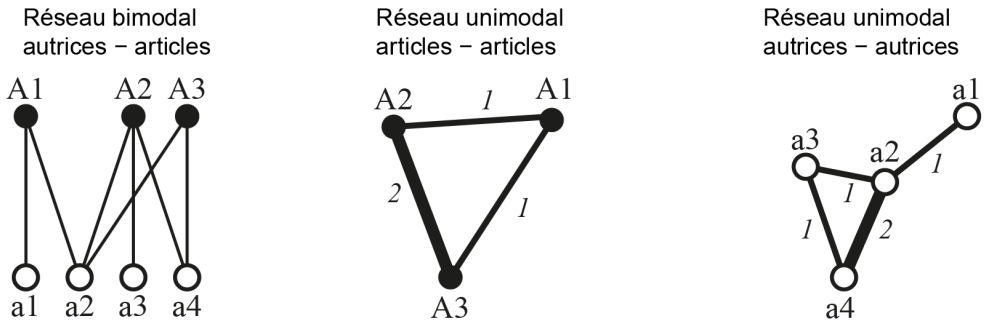
Les liens dans un réseau bimodal sont généralement considérés comme non orientés ; ils peuvent par contre être binaires ou valués. Une opération mathématique simple, illustrée dans la figure 3.7, permet de transformer un réseau bimodal en deux réseaux unimodaux valués. On pourrait tout à fait imaginer des réseaux tripartis ou quadripartis mais aucune méthode n'a su s'imposer pour les analyser [2].

Les données modélisées sous forme de réseau bimodal pourraient souvent être modélisées sous forme d'hypergraphe. Un hypergraphe est un graphe où un lien peut contenir de 1 à n sommets, n étant le nombre de sommets du graphe. Peu de logiciels permettent de les analyser et l'utilisation des hypergraphes reste rare en sciences sociales.

> **Exemple 1** : je m'intéresse aux pratiques de co-signatures d'articles scientifiques. Je constitue un corpus d'articles (ensemble de sommets V_1) et un corpus d'autrices ayant signé ces articles (ensemble de sommets V_2). Le réseau est bimodal car les liens existants sont entre autrices et articles. Si je le transforme en réseau unimodal autrices - autrices, un lien entre deux autrices signifie qu'elles ont signé ensemble un nombre d'articles au moins égal à 1. Si je le transforme en réseau articles - articles, un lien entre deux articles signale la présence d'au moins une autrice commune aux deux articles (figure 3.7).

1. En théorie des graphes, le caractère biparti d'un graphe est lié à sa structure ; ainsi tout arbre est un graphe biparti (Beauguitte, 2023) ; figure 7.1 page 67.
2. Je pourrais par exemple avoir un corpus d'autrices, un corpus de revues où publient ces autrices et un corpus d'éditrices propriétaires de ces revues et créer ainsi un réseau triparti.

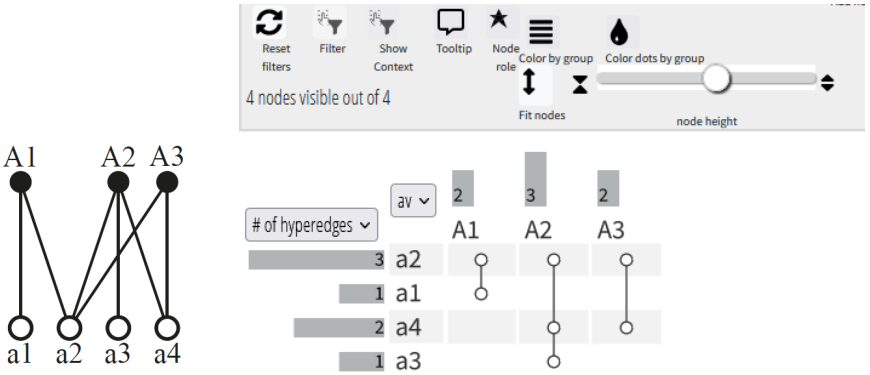
FIGURE 3.7 – Du réseau bimodal aux réseaux unimodaux



Le réseau bimodal de gauche peut être transformé en deux réseaux unimodaux valués : le réseau articles - articles au centre, le réseau autrices - autrices à droite. Dans le réseau central, l'intensité du lien A2-A3 est de 2, cela signifie que ces deux articles partagent deux mêmes autrices. Dans le réseau de droite, l'intensité du lien a2-a4 est de 2, cela signifie qu'elles ont co-signé deux articles. La transformation en réseau unimodal valué s'accompagne d'une perte d'informations dans la mesure où il n'est plus possible, une fois la transformation réalisée, de retrouver la structure de départ du graphe bimodal. Cette transformation peut s'avérer nécessaire pour l'analyse, de nombreux logiciels n'étant pas conçus pour analyser des réseaux bimodaux.

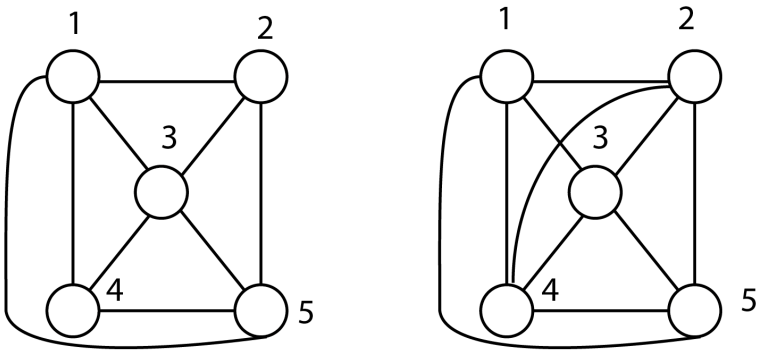
Exemple 1bis : je m'intéresse aux pratiques de cosignatures d'articles scientifiques. Je peux choisir une modélisation des données sous forme de réseau bimodal ; je peux également choisir d'utiliser des hypergraphes où un lien équivaut à un article et les x sommets de ce lien aux autrices (figure 3.8).

FIGURE 3.8 – Deux modélisations d'un même corpus : réseau bimodal et hypergraphe



La représentation sous forme de réseau bimodal est courante, celle par hypergraphe l'est moins. Cette dernière présente pourtant des atouts intéressants pour l'exploration visuelle des données (chapitre 12). La figure de droite a été réalisée avec le logiciel Paohvis (Valdivia et al.*, 2018).*

FIGURE 3.9 – Réseau planaire et non planaire



Le réseau de gauche est planaire, aucune arête ne se croise. Il est également complet ; le seul lien manquant (2-4) ne peut être ajouté sans provoquer une intersection de liens. Le réseau de droite est non planaire et complet.

Le réseau unimodal valué issu du réseau bimodal de départ indique une cooccurence et cette relation n'est pas nécessairement synonyme d'interactions ou de relations [1]. Si je remplace mon corpus d'articles par un corpus de conférences scientifiques, je peux transformer mon réseau pour obtenir un réseau chercheuses - chercheuses où le lien indique qu'elles ont assisté à 1, 2 ou x conférences communes ; rien ne me permet de supposer que ces deux personnes se connaissent, connaissent leurs travaux respectifs ou se sont adressées la parole.

Le réseau multi-niveaux (*multilevel network*) peut être considéré comme un mélange entre réseaux unimodaux et bimodaux. Le réseau est en partie bimodal : on étudie les liens entre deux catégories de sommets, par exemple entre des personnes et des organisations. Il est également unimodal dans la mesure où on cherche à étudier dans le même temps les liens à l'intérieur des deux niveaux considérés. J'aurai donc d'une part les liens d'appartenance entre individus et organisations, les liens entre les individus et enfin les liens entre organisations. Ce type d'approche nécessite un dispositif méthodologique de recueil des données plus lourd que les précédents.

Réseau planaire et réseau non planaire

Un graphe planaire est un graphe pouvant être projeté sur un plan sans qu'aucun lien ne se croise. Dans un graphe non planaire, deux liens peuvent se croiser sans problème. Lorsqu'on passe du graphe au réseau, le choix est en partie lié au type de sommets et au type de liens étudiés. Un réseau routier est généralement considéré comme un réseau planaire : les routes forment les liens et les intersections forment les sommets. Certes, il peut y avoir des ponts et des tunnels qui perturbent légèrement la structure mais, dans l'ensemble, le réseau peut être considéré comme planaire. Certains réseaux de métro sont planaires (Marseille), d'autres ne le sont pas (Paris). Le réseau aérien peut lui être considéré comme non planaire, les couloirs aériens pouvant emprunter les trois dimensions.

Le fait de considérer qu'un réseau est planaire a des conséquences méthodologiques fortes dans la mesure où les formules permettant de calculer les indicateurs doivent être

1. On connaît en général les autrices avec lesquelles on signe un article. Mais ce n'est pas toujours le cas...

adaptées (chapitre 5), cela impacte également le choix du logiciel utilisé. Le choix planaire/non planaire est moins fonction des liens présents et de l'existence d'éventuels croisements que de la possibilité d'*imaginer* des liens entre tous les sommets. Par exemple, si je considère les relations ferroviaires entre les villes françaises de plus de 20 000 habitants, imaginer une voie ferrée directe entre chacune de ces villes n'a pas de sens, que ce soit au niveau économique ou environnemental, et le réseau sera considéré comme planaire. Si je considère les relations au sein d'une classe de 25 élèves, je peux très bien imaginer que chaque élève connaisse les noms et prénoms de tous les autres élèves et le réseau sera considéré comme un réseau non planaire, quelles que soient les relations existant réellement dans cette classe.

3.4 Vos réseaux, vos choix

Au risque de me répéter, un réseau, qu'il s'agisse d'un réseau d'infrastructures, d'un réseau social (au sens sociologique et non numérique du terme) ou d'un réseau écologique, n'existe pas en soi, c'est un construit comme le sont les données relationnelles ayant servi à créer ce réseau. Il est toujours nécessaire de définir son ensemble de sommets et de définir ce que représentent les liens entre ces sommets.

Un réseau peut être transformé afin d'être plus adapté aux questions que vous vous posez et/ou aux outils et aux méthodes disponibles. La seule règle est d'expliquer comment on passe des données de départ au réseau analysé et, c'est quand même préférable, de justifier les transformations ou sélections opérées avant de mener à bien l'analyse.

Ce qui est vrai pour la construction du réseau reste bien entendu vrai pour l'analyse et la visualisation des données relationnelles : chaque fois que je fais un choix, que je décide de supprimer telle ou telle information, il est important d'expliquer comment et pourquoi je le fais.

À l'intention des formatrices

Les six termes de vocabulaire que les participantes sont censées connaître à l'issue d'une initiation à l'analyse de réseau sont sommet, lien, connexe, réseau orienté, réseau valué, réseau bimodal. Si on connaît la signification de ces termes, on doit pouvoir retrouver les autres (ex : sommets + liens = réseau ; inverse de réseau orienté = réseau non orienté, etc.).
Si les participantes oublient les boucles ou les liens multiples, ce n'est pas très grave : 95 % de la littérature en sciences sociales analysent des réseaux simples (pourcentage non validé par une étude empirique digne de ce nom). Les réseaux planaires intéressent les seules géographes des transports, et encore, réseaux fluvial, routier et ferré uniquement.

Chapitre 4

Construire ses données

Mener à bien une analyse de réseau suppose de construire ses données relationnelles afin d'obtenir un ensemble de sommets et un ensemble de liens symbolisant une relation entre ces sommets. Définir ces deux ensembles est l'une des tâches essentielles : que représentent mes sommets ? que représentent mes liens ? Passer du temps pour répondre aussi précisément que possible à ces deux questions est indispensable pour tenter d'obtenir des résultats intéressants.

4.1 Construire l'objet

Les données relationnelles que vous choisissez d'analyser ont deux origines possibles : vous les avez construites, vous les avez récupérées et dans ce cas d'autres que vous les ont construites. Le terme construction est essentiel : il n'existe nul part dans le monde social, y compris le monde social numérique, des objets qui pourraient être considérées comme des données qu'il suffirait de « récolter » puis de « nettoyer [1] » avant de les analyser.

Quel que soit le soin avec lequel vous récoltez et mettez en forme vos données relationnelles, il est important de garder à l'esprit qu'un réseau est un modèle de la réalité, une version simplifiée et incomplète d'un type donné de relations entre certaines entités. Un réseau n'existe pas en soi et ce quelle que soit sa nature. Ceci est vrai des réseaux personnels des individus mais c'est également vrai pour des réseaux type infrastructures de transport. Tout recueil et mise en forme de données supposent des choix plus ou moins arbitraires, des décisions plus ou moins bien informées.

> **Exemple 1** : je m'intéresse aux relations sur Twitter. Ma première étape est de définir ma population et de justifier ses limites : pourquoi ces comptes et pas d'autres ? La deuxième étage est de définir ce que j'appelle une relation sur twitter : un abonnement réciproque (v_1 est abonnée à v_2 et v_2 est abonnée à v_1) ? le fait de retwitter ce que poste un compte ? le fait de répondre à un tweet ? le fait de nommer l'utilisatrice du compte dans les tweets ? Il n'existe pas une solution, il en existe des dizaines et toutes peuvent avoir leur pertinence en fonction des questions que je me pose.

1. Les données ne sont pas « sales » au départ et « propres » à l'arrivée : les données initiales présentent un certain nombre de caractéristiques qui peuvent être intéressantes à analyser, les données que vous avez transformées de telle ou telle façon permettent de mieux répondre à vos questions de recherche. Oui, ceci est l'hypothèse optimiste.

Exemple 2 : soit le réseau ferroviaire en Île-de-France que je souhaite modéliser sous forme de réseau planaire. Quelles lignes dois-je prendre en compte ? Si je prends le Francilien, il serait peut-être utile de considérer aussi les lignes RER. Est-ce que j'inclus les correspondances entre les gares parisiennes et si oui, lesquelles (RER, métro) ? Est-ce que je prends en compte les travaux et l'inaccessibilité temporaire de certaines stations ? Est-ce que je cherche à prendre en compte la fréquence des dessertes et pas la seule présence d'une gare sur le plan ?

À l'intention des formatrices

Il est relativement facile de montrer le caractère construit des données relationnelles, y compris avec des données qui paraissent aussi intangibles qu'un réseau ferré, autoroutier ou viaire. Imprimer un extrait de plan OpenStreetMap et demander à trois groupes de tracer le réseau viaire (en donnant un minimum de règles simples et une question de recherche générale) correspondant à ce plan, vous obtiendrez sans aucun doute des résultats différents. L'objectif n'est pas de conclure qu'on peut faire ce qu'on veut et que dans tous les cas on triche mais au contraire d'insister sur la nécessité de documenter les choix effectués.

Les lignes qui suivent sont triviales et s'appliquent à toute démarche de recherche, qu'elle se prétende qualitative, quantitative ou mixte. Si je m'amuse à récolter des données, c'est pour répondre à des questions. Lorsque ces données sont relationnelles, cela suppose que je suis capable de délimiter la population que j'étudie et que je suis capable de définir la ou les relations entre les individus de cette population - je le répète une dernière fois, population et individus sont entendus au sens statistique, ce peut être des villes, des plantes, des personnes, des journaux, etc. etc. Plus les critères utilisés, tant pour délimiter la population que pour définir la ou les relations, sont précis et plus vous serez capable d'interpréter les résultats obtenus. Inversement, récupérer un corpus dont on ne sait pas trop comment il a été construit est un bon moyen de perdre son temps.

4.2 Recueillir les données

Réseau complet *vs* réseau personnel

En ce qui concerne le recueil des données, deux grandes approches d'analyse de réseau existent : l'approche par réseau complet et l'approche par réseau personnel (*ego-network*, *personal network*[1]). Dans le premier cas, je délimite une population et je cherche à recueillir un ou plusieurs types de relations existant à l'intérieur de cette population. L'une des difficultés principales consiste à s'assurer de la congruence entre ses questions de recherche et ce que m'apportera une analyse en réseau complet. Il faudra également anticiper les fluctuations possibles dans la composition du groupe.

Exemple : je m'intéresse aux relations amicales chez les jeunes enfants et je prends pour terrain d'étude les relations au sein d'une classe de CP. Il n'est pas du tout certain que les données que je récolte soient pertinentes : je suppose plus ou moins implicitement que les enfants nouent leur amitié à l'école et dans leur classe. Il faudra *a minima* que je pose des questions sur leurs meilleures

1. Hennig *et al.* distinguent dans leur manuel les réseaux personnels (ego et ses alters) et les réseaux égocentrés (liens entre ego et alters et entre alters) (2012, p. 54). Les autres manuels consultés, et ce guide pratique, ne font pas cette distinction.

amies pour vérifier si c'est le cas et dans quelles proportions. Si par contre, je travaille sur les dynamiques de groupe au sein d'une classe, le dispositif est sans doute plus directement utile.

La collection des données par réseau personnel vise à collecter l'ensemble des liens d'un groupe d'individus choisis pour une raison donnée. Si je reprends l'exemple de la salle de classe, je ne cherche pas à déterminer les relations entre les élèves de cette classe mais les relations que chaque élève (ego) entretient avec ses amies, ses voisines, sa famille, etc (alters). Dans la mesure du possible, je cherche également à déterminer si les relations de l'individu se connaissent entre elles, ce qui est parfois difficile à déterminer. En effet, il est rarement possible d'interroger les egos puis l'ensemble des alters cités. Les liens entre alters sont donc ceux qui sont connus, identifiés par les egos. Ceci n'est pas un problème dans la mesure où ce qui importe est bien la façon dont ego voit son entourage et non les relations « réelles » au sein de ce dernier.

Avec la démarche par réseau complet, j'obtiens un réseau des relations entre les V individus du groupe étudié ; avec la démarche en réseau personnel, j'obtiens V réseaux personnels en forme d'étoile centrés sur les différents individus enquêtés.

Exemple : la sociologue Claire Bidart réalise en 1995 la première vague d'enquête dite du panel de Caen (Bidart *et al.*, 2011 ; carnet de recherche Panel de Caen). Elle interroge trois groupes d'élèves (terminales SES, bac pro et stage d'insertion) sur leurs relations amicales, amoureuses, professionnelles, associatives, etc. Une distinction est faite entre liens faibles (liens occasionnels ou se déroulant dans un seul contexte) et liens forts (liens existant dans plusieurs contextes ou considérés comme importants par la personne[1]).

Générer des noms : questionnaires et entretiens

Lorsque l'on souhaite étudier des liens interpersonnels, quatre dispositifs de recueil de données au moins sont possibles :

- l'utilisation de questionnaire ;
- la passation d'entretien ;
- l'examen de traces laissées par les personnes (numériques ou non) ;
- l'observation.

La première étape, quelle que soit l'option retenue, est d'aller voir ce qui a déjà été fait, non pour le répéter mais pour s'en inspirer. Il n'est évidemment pas utile de tout lire mais piocher dans de vieux articles de sociométrie, des thèses[2] récentes et d'autres moins récentes, des carnets de recherche, etc. permet d'avoir une vision large de ce qu'il est possible de faire.

Il existe de nombreux manuels sur ces questions (chercher « méthodes qualitatives en ... » en remplaçant ... par votre discipline et les références devraient apparaître) et il n'est pas question pour moi ici de les plagier. Il s'agit simplement ici d'attirer l'attention sur des obstacles récurrents et des scrupules tout aussi récurrents[3]. En règle générale, il

1. Pour les définitions précises du lien fort dans cette enquête, voir la page méthodologie du carnet de recherche.

2. Les précisions méthodologiques sont souvent omises, par manque de place, dans les articles ou dans les ouvrages tirés des thèses ; regarder les annexes de ces dernières est plus efficace.

3. Un grand merci à mes collègues historiennes avec lesquelles j'ai souvent échangé sur le problème des données manquantes.

est plus facile de questionner les personnes sur des actions que sur des sentiments ou des concepts abstraits. Nous n'avons pas toutes la même définition de l'amitié, des personnes importantes dans notre vie, etc. Par contre, nous sommes toutes capables de dire si nous appelons cette personne toutes les semaines, si nous la voyons une fois par mois ou plus, si nous partons en vacances ensemble, etc. Poser des questions sur des faits (et pas des faits anciens, nos mémoires sont peu fiables) et sur des actions (récentes elles aussi) est un bon moyen d'obtenir des réponses précises.

Le conseil est également valable pour les démarches basées sur l'observation. Si j'étudie les interactions entre enfants dans une crèche, j'ai tout intérêt à avoir des items précis décrivant des actions (tire les cheveux, prend un jouet des mains, griffe [1]).

> **Exemple** : dans l'ouvrage coordonné par Gribaudi (1998), l'annexe 2 reproduit le carnet que doivent remplir pendant deux semaines les personnes enquêtées. Les extraits ci-dessous montrent comment certains problèmes ont été anticipés par l'équipe.
>
> « 1. *À quelle heure a lieu la rencontre que vous enregistrez ?* Si vous ne vous rappelez pas de l'heure exacte, donner une indication du moment de la journée. [...]
>
> 3. *Identité de la personne ?*
>
> - Donnez le prénom exact de la personne et, pour garantir l'anonymat, l'initiale du nom.
>
> - Si vous ne connaissez pas son nom ou si vous ne vous en souvenez pas, attribuez un code de trois lettres identiques. Attention, ce code devra être unique. Il doit être attribué à une seule personne, et réutilisé dans le cas où cette personne réapparaît dans les jours suivants. [...]
>
> - Si, lors des rencontres successives, vous parvenez à connaître le vrai nom, indiquez-le tout en continuant à utiliser son code que vous mettrez maintenant (entre parenthèses).
>
> Ex. : Georges_S (AAA) » [...]
>
> 10. *Profession principale*. Attribuez la dénomination que vous utiliseriez sans réfléchir. » (Gridaudi (dir.), 1998, pp. 328-329).

La période durant laquelle les personnes enquêtées sont censées remplir le carnet des rencontres est court (15 jours non fériés) ; des questions supplémentaires visent à nommer des personnes « importantes » absentes durant cette période. Les personnes sont censées noter les « rencontres significatives » (annexe 1, pp. 313-325), terme flou utilisé afin d'éliminer les contacts routiniers liés à la sphère professionnelle (les personnes enquêtées sont professeures du secondaire, ingénieures ou médecins).

Quelle que soit la méthode retenue, il vous faudra anticiper un certain nombre de problèmes. Il n'est bien sûr pas très pertinent d'imaginer *tous* les obstacles pouvant se dresser entre vous et le recueil de données mais quelques situations se rencontrent fréquemment. La liste qui suit est non exhaustive et une solution possible est systématiquement indiquée. Vous avez défini votre population, que la méthode soit de type réseau complet ou réseau personnel, et vous savez quel type de relation vous souhaitez étudier. Si ce n'est pas le cas, inutile à ce stade de faire passer un entretien ou un questionnaire. Les obstacles les plus classiques que vous risquez de rencontrer sont les suivants :

1. Oui, je sais, j'ai une vision sombre de la petite enfance.

1. Absence(s). Vous avez choisi de travailler sur un réseau complet (de personnes, d'entreprises, d'animaux, etc.) or il y a des absentes, ce qui pose problème quel que soit le type de méthode choisie. Un moyen de limiter les dégâts est de prévoir des périodes courtes pour récolter les informations. La date de la période de recueil compte également ; éviter le mois de juin pour des données en milieu scolaire par exemple. Mais quelles que soient les précautions prises, il risque d'y avoir des données manquantes et ce n'est pas grave, c'est *toujours* le cas (y compris avec des sources numériques, j'y reviendrai). Si la personne absente s'avère centrale dans le réseau étudié (elle est souvent citée), il peut être nécessaire de revenir sur le terrain pour la rencontrer.

2. Instabilité. Vous avez choisi de travailler sur un réseau complet ou sur des réseaux personnels et votre population, soigneusement définie, ne cesse de varier. Des individus partent, d'autres arrivent, certains fusionnent ou se séparent : la variabilité de l'échantillon fait partie de ses caractéristiques et il serait illusoire de prétendre le contraire. Là encore, une ou plusieurs périodes aussi brèves que possible de recueil des données permet de limiter les variations. Quel que soit le réseau observé, il s'agit d'un système en mouvement, vous parviendrez au mieux, et c'est déjà bien, à capturer de façon un peu précise l'état de la plupart des relations à un temps t.

Des problèmes apparaîtront ensuite lorsque vous saisirez vos données et heureusement, les problèmes sont l'un des meilleurs moyens pour faire surgir des questions inattendues.

L'un des problèmes récurrents est la variabilité des réponses. Les réponses obtenues peuvent être décevantes : les enquêtées ont le droit de faire preuve de mauvaise volonté et de répondre à la moitié (au mieux) des questions posées. Ça fait partie du jeu. C'est exactement pareil quand on travaille avec des statistiques au niveau international, il est des États statistiquement plus fiables que d'autres. Les réponses peuvent également présenter des variations très fortes : une personne déclarera avoir 50 amies proches, une autre en déclarera 2 ou 0. Certaines chercheuses limitent ce biais en imposant un nombre maximum de réponses ; il n'est pas certain que cette solution soit totalement satisfaisante dans la mesure où elle empêche de prendre en compte la variabilité inter-individuelle des phénomènes sociaux étudiés.

Générer des liens : observer les interactions

Le recours à l'observation des interactions peut être motivé par différentes raisons non exclusives les unes des autres. Il est des cas où l'utilisation de l'entretien ou du questionnaire est impossible (animaux, plantes, enfants très jeunes). Il est des terrains où la chercheuse peut considérer que l'observation ou la participation est une méthode moins intrusive et plus éthique que l'entretien (populations marginalisées, se livrant à des activités illégales, activisme politique, etc.). Certaines disciplines ont par ailleurs recours à ces méthodes (anthropologie notamment) quand d'autres ne peuvent les adopter (histoire, archéologie).

Exemple : « la plupart des données de cet article proviennent d'observations d'environ une heure par jour effectuées pendant 275 des 459 jours entre le 21 mars 1961 et le 24 juin 1962 et d'observations d'environ six heures par jour effectuées pendant 47 des 48 jours entre le 14 juin et le 31 juillet 1963. [...] J'ai chronométré certaines séquences de comportement. J'ai noté chaque événement dans un carnet quelques secondes après l'avoir vu se produire.[...] Chaque fois

que j'ai vu un singe de ce groupe s'asseoir ou se coucher en touchant un autre, j'ai identifié les individus et enregistré l'événement. » (Sade, 1965, pp. 2-4).

L'observation n'est possible qu'avec des groupes restreints dont il est possible de différencier les différents membres. C'est par ailleurs une méthode relativement coûteuse en temps : les périodes d'observation doivent être suffisamment longues pour que l'observatrice puisse distinguer les régularités de l'action observée. Des dispositifs techniques comme les puces RFID sont parfois utilisés pour pallier ces différents inconvénients : le projet *SocioPatterns* a par exemple voulu étudier les interactions entre participantes à des colloques [1]. Chaque personne est munie d'une puce et un lien est créé lorsque deux personnes se trouvent face à face à moins d'un mètre l'une de l'autre. L'information est alors supposée complète - elle ne l'est pas, des personnes posent leurs puces, la retirent, le dispositif connaît des ratés, etc. - mais il peut être difficile de différencier interactions volontaires et interactions fortuites.

Générer des réseaux textuels

La construction des réseaux lexicométriques (liens entre des mots) obéit presque aux mêmes règles que les réseaux personnels mais nécessitent des outils différents. Ce qui ne change pas, c'est qu'il faut commencer par définir sa population de sommets et définir ce qu'est une relation entre ces sommets. Une étape préalable supplémentaire consiste à définir un corpus de textes. Et travailler sur des corpus de textes n'est pas plus facile ni plus difficile que travailler sur des données d'enquêtes.

J'ai défini mon corpus de textes (un ensemble de romans, d'articles ,de chansons, de tweets, etc., etc.). J'ai défini les termes qui correspondent à mes sommets : cela suppose d'adopter des règles de lemmatisation [2] claires. Est-ce que singulier et pluriel sont équivalents (la politique et les politiques par exemple) ? Est-ce que les formes nominales et adjectives sont équivalentes ? L'une des difficultés consiste à identifier les multiples formes lexicales qu'un même sommet peut prendre. Si le corpus est volumineux, suis-je capable de mettre en œuvre des méthodes semi-automatiques efficaces ?

Exemple : je m'intéresse à la place des pays de l'Union européenne (plus le Royaume-Uni) dans les quotidiens français. J'ai un certain nombre d'hypothèses (prime aux voisins et aux puissances économiques, indifférence vis-à-vis des pays scandinaves et baltiques). Je m'intéresse notamment aux pays cités ensemble dans un titre. Je constitue mon corpus de titres et, à l'aide d'une collègue, nous mettons au point un outil qui capture les titres des articles classés « Actualités internationales [3] ».

J'ai mon corpus (liste de quotidiens d'information français avec une date de début et une date de fin), mes sommets (États membres de l'UE en 2022 plus le Royaume-Uni) et mes liens (coprésence d'États dans un titre de quotidien). Le 26 juillet 2022, nous capturons les titres suivants sur lefigaro.fr [4] : « Berlin prépare sa stratégie de sécurité nationale », « Pays Baltes : la ruée vers

1. Voir le site http://www.sociopatterns.org/. Le jeu de données évoqué ici est le *SFHH conference data set*.
2. Il s'agit de regrouper les différentes formes pouvant être prises par un mot.
3. Ce qui n'est pas une tâche si facile qu'elle en a l'air.
4. Je le précise pour mes jeunes lectrices : avant de devenir un média réactionnaire spécialisé dans les paniques morales absurdes, ce fut un excellent quotidien d'information (cf notamment les articles de Patrick de Saint-Exupéry sur le Rwanda en 1994).

l'ouest »ou encore « Démission de Draghi : Macron salue un grand homme d'État italien et un partenaire de confiance ». Ces trois exemples, pas du tout sélectionnés au hasard, montrent que le dictionnaire des homonymes qu'il va falloir construire devra inclure les capitales (Berlin = Allemagne), les chefs d'États (présidentes, premières ministres), les adjectifs et qu'il faudra également créer des règles *ad hoc* (noter 1/3 pour chacun des pays baltes ? est-ce que l'expression pays baltes justifie la création d'un lien entre les trois pays ?).

Des données numériques au réseau

Les données nativement numériques sont des données textuelles un peu particulières dans la mesure où leur production est en grande partie liée au fonctionnement des plateformes numériques étudiées et qu'elle est généralement continue [1] ; leur volume peut donc rapidement devenir important. Mais il est tout à fait possible de mener des analyses mixtes sur de petits corpus de données numériques.

> **Exemple** : dans un article de 2021, Mehdi Arfaoui collecte les données de 262 comptes de sociologues présentes sur Twitter et retient les 11 278 liens entre ces comptes (méthode en réseau complet). Ces liens sont orientés. Il récolte des attributs sur les personnes (rang, statut, genre), ajoute un attribut taille à chaque sommet correspondant au nombre d'abonnées total sur Twitter et caractérise qualitativement sa population en étudiant notamment le ratio nombre d'abonnements / nombre d'abonnées. Si l'article porte sur la présence de comptes de sociologues sur un réseau social numérique, la partie analyse de réseau *stricto sensu* est réduite et consiste essentiellement en une visualisation du réseau obtenu [2].

Il est amusant de comparer les scrupules des personnes travaillant avec des données historiques et l'absence totale de scrupules de certaines personnes travaillant avec des données numériques qui sont sensées être fines, exhaustives, actualisées en temps réel et donc capables de générer des portraits supposés fidèles d'un aspect du monde social. Les données numériques sont certes généralement horodatées et volumineuses mais elles servent à renseigner une activité sur un outil. Si j'étudie des liens d'amitié sur Facebook, les amitiés que j'observe sont les amitiés Facebook et non un reflet numérique des amitiés des personnes ayant un (ou plusieurs) comptes Facebook. Les réseaux sociaux numériques ajoutent des couches qu'il peut être passionnant d'étudier au monde social, elles ne contribuent absolument pas à le rendre plus transparent ou plus facilement compréhensible.

Si les données historiques sont quasi systématiquement parcellaires, les données nativement numériques sont rarement exhaustives dans la pratique de la recherche : les outils de capture connaissent des ratés, les utilisatrices utilisent un VPN perturbant leur identification, la même personne utilise différents comptes sur un même réseau social numérique etc. etc. Si les personnes travaillant sur ces données emploient sans cesse le terme (impropre) de « nettoyage » des données, ce n'est pas un hasard. Les données nativement numériques nécessitent, comme toutes les données, d'être construites pour répondre aux questions de recherche.

1. Que je sois active ou non sur le réseau étudié, le simple fait de me connecter à la plateforme génère des logs (*a minima* identifiant, date de début, date de fin, adresse IP).

2. La visualisation en question fait partie de la famille très très nombreuse des « grosse patates avec des couleurs », j'y reviendrai au chapitre 12.

Des infrastructures au réseau

A priori, l'existence dans le monde social d'un réseau routier ou ferré ne se questionne pas. Les questions pourtant apparaissent lorsqu'il s'agit de modéliser ce réseau sous forme de graphe. Des règles précises doivent être construites pour définir l'emprise spatiale du réseau considéré (quelles limites choisir ?), les sommets (quels types d'intersection ?) et les liens (prendre seulement les axes principaux et si oui comment les définir ?).

> **Exemple** : dans un article de 1972, le géographe étatsunien William Muraco cherche à comparer l'accessibilité intra-urbaine dans deux villes. Il se base sur un « plan de transport et d'aménagement du territoire d'Indianapolis », plan qui catégorise les voies en « autoroutes, voies rapides, artères primaires, artères secondaires, collecteurs et dessertes locales. » Il ne garde que les quatre premières catégories en justifiant ainsi sa sélection : « cette étude porte principalement sur les sites industriels et commerciaux, les artères principales sont supposées refléter les liens pertinents du réseau de transport. »

Lorsqu'on étudie des réseaux de transport où seuls les sommets sont localisés (port, aéroports, antennes de téléphonie mobile), les règles sont les mêmes : définir les sommets, les liens et les limites du réseau étudié.

> **Exemple** : dans un article de 1990, Nadine Cattan s'intéresse au réseau des villes européennes en étudiant deux types de relations aériennes, le transport de personnes et celui de marchandises. Deux types de distance sont considérées : la distance à vol d'oiseau et la distance temps (temps de vol + temps d'attente + temps pour relier les centres-villes aux aéroports). Les villes (sommets) sont sélectionnées en fonction de la population (plus de 200 000 habitant.e.s), les liens sont les « échanges réguliers, internationaux » fournis par l'Organisation de l'Aviation Civile Internationale.

Qu'elles soient personnelles, textuelles, nativement numériques ou relatives à des infrastructures, les données relationnelles supposent toujours la construction d'une population de sommets et la définition de ce que signifie un lien entre ces sommets. Les choix effectués sont en partie fonction des données disponibles mais ils dépendent surtout des questions que vous vous posez. Une fois les données collectées, elles peuvent être présentées sous trois formats différents.

4.3 Trois formats pour un même objet : liste, matrice et graphique

Vous avez récolté vos données, géré au mieux les lacunes et les approximations, vous avez régulièrement dû faire des choix et vous les avez documentés aussi précisément que possible sinon vous allez vite les oublier et vos données perdront beaucoup de leur intérêt, reste maintenant à les mettre en forme afin qu'elles puissent être lues et analysées par un logiciel.

Il existe trois formats courants pour représenter un réseau :
- la liste de liens (*edgelist*) et, éventuellement, la liste de sommets (*nodelist*) ;
- la matrice dite d'adjacence (*adjacency matrix*) ;
- la représentation graphique.

À l'intention des formatrices

Ce petit guide pratique tente d'être aussi peu directif que possible et vise à présenter différentes options possibles en soulignant à chaque fois les avantages et les inconvénients. Les paragraphes qui suivent sont un peu plus normatifs que les précédents pour une raison pratique : si l'on souhaite faire de l'analyse de réseau, on a besoin de logiciels et ces logiciels attendent des formats de données précis.
Un bon moyen d'attirer l'attention sur l'importance de correctement mettre en forme ses données est de fournir des mini jeux de données plus ou moins corrompus et de faire identifier les problèmes (séparateurs multiples, encodage exotique et caractères spéciaux, nombre de colonnes variable et présence de noms de colonnes tout aussi variable, répétition des informations, etc.). Il ne s'agit pas de former des spécialistes des bases de données relationnelles mais de donner les outils nécessaires pour que les participantes prennent un minimum de bonnes habitudes afin de gagner du temps, d'être plus efficaces et, avec un peu de chance, qu'elles puissent ensuite s'inscrire dans des démarches de science ouverte.

Reprenons les réseaux d'exemple du chapitre 3 (figure 4.1). À gauche, la liste de liens correspondante est la suivante : $\{ab, bc, bd, be, ce, de, ef, fg\}$. Dans la mesure où le réseau est non orienté, il n'est pas nécessaire de préciser systématiquement le lien opposé (ba, cb etc.). À droite par contre, le réseau étant orienté, tous les liens doivent être indiqués et on obtient la liste $\{ab, ba, bd, db, be, cb, ec, fe, fg\}$.

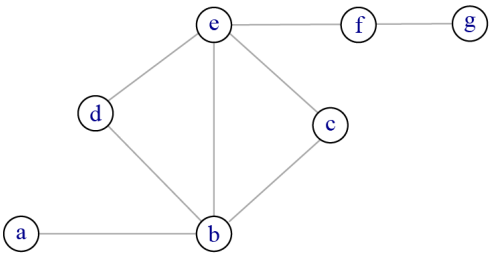
Il est rare que l'on liste les liens en ligne comme cela est fait dans le paragraphe précédent : contrôler les données serait pénible et ajouter des attributs aux liens serait compliqué. La liste de liens aura une forme de type sommet de départ - séparateur - sommet d'arrivée - séparateur - attribut numéro 1 - séparateur - attribut numéro 2, etc. Ce qui donne avec les deux exemples vus à l'instant les listes de liens suivantes :

Réseau non orienté	Réseau orienté
a,b	a,b
b,c	b,a
b,d	b,d
b,e	d,b
c,e	b,e
d,e	c,b
e,f	e,c
f,g	f,e
	f,g

Le séparateur ici est une virgule mais j'aurais pu choisir un point-virgule ou une tabulation, la seule règle est de conserver le même séparateur tout au long de la liste de liens. Si mes liens ont des attributs, une intensité dans le cas d'un réseau valué par exemple, il suffit d'ajouter une colonne ; si mes liens ont x attributs (date de début et de fin, nature du lien, etc.), j'ajoute x colonnes.

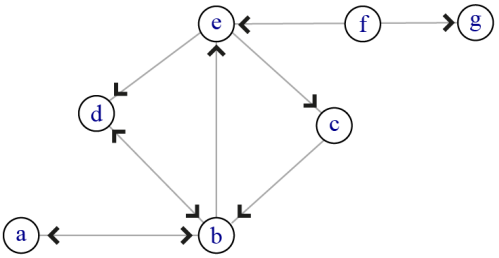
La matrice d'adjacence d'un réseau unimodal simple de V sommets est une matrice carrée de taille $V \times V$; la présence d'un lien est signalée par un 1, l'absence de liens par un 0. Par convention, dans le cas d'un réseau orienté, l'origine est en ligne et la destination en colonne. Il n'y a pas de boucle dans un réseau simple donc la diagonale principale est vide. Si le réseau est non orienté, la matrice est symétrique par rapport à la diagonale principale.

FIGURE 4.1 – Graphe, liste de liens, matrice d'adjacence



$\{ab, bc, bd, be, ce, de, ef, fg\}$

	a	b	c	d	e	f	g
a	0	1	0	0	0	0	0
b	1	0	1	1	1	0	0
c	0	1	0	0	1	0	0
d	0	1	0	0	1	0	0
e	0	1	1	1	0	1	0
f	0	0	0	0	1	0	1
g	0	0	0	0	0	1	0



$\{ab, ba, bd, db, be, cb, ec, fe, fg\}$

	a	b	c	d	e	f	g
a	0	1	0	0	0	0	0
b	1	0	0	1	0	0	0
c	0	1	0	0	0	0	0
d	0	1	0	0	1	0	0
e	0	0	1	0	0	0	0
f	0	0	0	0	1	0	1
g	0	0	0	0	0	0	0

Chacune de ces représentations contient exactement la même information mais leur utilité est différente.

La matrice d'adjacence peut prendre en charge les pseudographes (présence de boucles) : la diagonale ne sera alors pas nécessairement vide. Elle peut également prendre en charge les réseaux valués : on remplit les cases avec une valeur numérique correspondant à l'intensité de la relation. Par contre, dans le cas de réseaux multiplexes (présence de plusieurs relations entre les sommets), chaque relation nécessite sa propre matrice. Enfin, lorsqu'on a affaire à un réseau bimodal avec deux ensembles V_1 et V_2 de sommets, la matrice d'adjacence correspondante est de taille $V_1 \times V_2$.

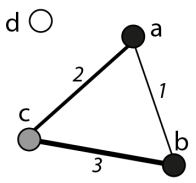
Quant à la représentation graphique, ici sous la forme très majoritairement utilisée nœuds-liens, elle est utile tant en phase exploratoire que pour présenter ses résultats (chapitre 12).

Liste de liens, matrice et représentation graphique fournissent les mêmes informations mais ces objets ont des utilités différentes. La liste de liens est le moyen le plus simple pour organiser et stocker ses données ; c'est également le format privilégié par les logiciels d'analyse de réseau (chapitre 13). Si l'on souhaite ajouter des attributs aux liens, il suffit d'ajouter une colonne. Si l'on souhaite ajouter des attributs aux sommets, on crée la table des sommets et on ajoute une colonne. Ce format est aussi le plus économique pour stocker ses données [1].

Le format matriciel peut être utilisé pour visualiser les données (chapitre 12), il sert également à calculer un certain nombre d'indicateurs courants en analyse de réseau. Par contre, l'objet est encombrant : les matrices données en exemple contiennent un grand nombre de zéros et, si le réseau est non orienté, elle présente deux fois la même observation.

1. Ce n'est pas tout à fait vrai : on pourrait stocker ses données sous une forme plus compacte avec une seule ligne par individu où seraient listés ses voisins (donc les liens) et ses attributs. Ce format est peu employé en pratique.

FIGURE 4.2 – Isolés et listes



Liste de sommets	Liste de liens
a, noir	a,b,1
b, noir	a,c,2
c, gris	b,c,3
d, blanc	

Le réseau simple, valué et non orienté à gauche présente un sommet isolé. Les sommets ont un attribut, les liens également. Les deux listes (sommets et liens) correspondent au réseau représenté à gauche. Là encore, le séparateur utilisé est la virgule.

Enfin, la forme graphique est utile tant pour explorer ses données que pour communiquer ses résultats ; elle ne permet évidemment pas d'automatiser le moindre calcul.

Petite astuce pratique : si le réseau comprend des sommets isolés, la liste de liens ne suffit pas. Deux options existent alors : créer la liste de sommets correspondants ; créer une boucle fictive du sommet isolé vers lui-même et la supprimer une fois les données importées dans le logiciel[1]. Dans la pratique, il est rare que l'on ait des sommets sans attribut et on a généralement une liste de liens d'une part, une liste de sommets d'autre part (figure 4.2).

4.4 Documenter, enrichir, partager

Vous avez maintenant tout ce qu'il vous faut pour mener à bien une analyse de réseau : une liste de sommets et des attributs, une liste de liens et des attributs éventuels. En fonction de vos habitudes de travail, vous stockez ces listes dans deux feuilles d'un classeur LibreOffice ou Excel ou vous créez un fichier texte par liste. Le format .csv a ma préférence car il est lisible à la fois par les tableurs et par les éditeurs de texte type Notepad++ ; ce dernier outil est pratique notamment quand on doit gérer des caractères spéciaux et qu'on souhaite contrôler l'encodage utilisé. Par contre, il oblige à stocker dans des fichier différents les liens, les sommets et le troisième fichier indispensable, à savoir les métadonnées.

Ce dernier fichier documente votre jeu de données, il précise quand et comment vous l'avez construit, quelles sont les règles utilisées pour définir sommets et liens, quels individus ont éventuellement été supprimés, quels sont les attributs et leurs modalités, etc. etc.

Ce dernier fichier doit être aussi détaillé et précis que possible. Ok, c'est long et pénible mais prendre le temps et l'énergie de le faire correctement vous fera gagner un temps fou ensuite ; c'est à peu près le seul moyen permettant de construire un jeu de données et de l'analyser six mois ou deux ans après. Ce qui nous paraît évident quand on construit des données cesse de l'être quelques semaines plus tard et, sans métadonnées précises, on est le plus souvent obligé de reconstruire tout son jeu de données de A à Z. Ce qui est encore plus long et plus pénible.

Comment savoir si vos métadonnées sont assez détaillées ? Imaginez que vous devez à l'issue de votre travail mettre vos données en accès ouvert afin qu'elles puissent être réutilisées de façon pertinente, ce qui est par ailleurs de plus en plus souvent exigé par

1. J'utilise parfois cette astuce en phase exploratoire, quand je souhaite jeter un œil aux données que je suis en train de construire.

les financeurs de la recherche [1]. En dehors des contraintes croissantes en vigueur dans nos milieux professionnels, se poser la question de l'ouverture des données dès leur construction oblige à documenter soigneusement l'ensemble des choix effectués (est-ce que je supprime certains sommets ? est-ce que je néglige certains liens ? si oui, pourquoi et comment est-ce que j'opère la sélection ?) et l'interprétation que vous ferez ensuite des résultats n'en sera que meilleure.

Dernier point avant de voir comment analyser ces données relationnelles : il n'existe aucun lien entre le volume des données et la qualité des analyses possibles. On peut produire des connaissances passionnantes avec des données relatives à une douzaine de familles londoniennes (Bott, 1971) ; on peut produire des résultats triviaux avec d'énormes volumes de données. L'inverse est tout aussi vrai : je peux réaliser des analyses médiocres sur de petits corpus et des analyses passionnantes sur de gros corpus. Il est vrai cependant que toutes les méthodes ne sont pas adaptées pour toutes les tailles de corpus mais ceci sera systématiquement signalé dans les chapitres qui suivent.

1. Il est possible de tenir des discours euphoriques sur l'ouverture des données ; on peut aussi considérer que cela revient à faire un cadeau supplémentaire au secteur privé sans la moindre contrepartie.

Chapitre 5

Quelques mesures possibles

Ce chapitre pourrait être deux, trois voire dix fois plus épais sans trop de problèmes : il existe de nombreuses mesures couramment utilisées en analyse de réseau. J'ai choisi d'en présenter une poignée en insistant à chaque fois sur ce que ces mesures permettent et sur ce qu'elles ne permettent pas. L'adaptation de ces mesures pour les réseaux valués et les réseaux planaires est systématiquement indiquée. Par contre, les mesures des réseaux bimodaux et multiplexes sont évoquées dans des chapitres indépendants. L'aspect mathématique a été réduit au minimum mais quelques formules sont cependant présentes ; les détails mathématiques se trouvent en fin d'ouvrage dans les annexes A et B.

Dans les synthèses du groupe fmr (flux, matrices, réseaux) rédigées il y a presque dix ans avec des collègues géographes [1], nous distinguions les mesures locales, portant sur un sommet et plus rarement sur un lien, et les mesures globales portant sur le réseau dans son ensemble. Cette distinction commode, généralement absente des manuels d'analyse de réseau [2], n'est pas reprise ici : si les premières mesures présentées (densité, diamètre) peuvent être considérées comme globales, celles qui suivent, notamment le degré, sont utiles tant pour hiérarchiser les sommets que pour caractériser le réseau dans son ensemble (distribution des degrés).

Deux rappels basiques avant de commencer cette courte exploration :

- si je choisis de mesurer quelque chose, c'est pour répondre à une question et non parce que l'indicateur est disponible dans le menu déroulant d'un logiciel. Si je ne comprends pas ce que je mesure, je ne saurai pas interpréter les résultats. Cela est encore pire s'il existe des paramètres que je peux modifier (voir *infra* les différentes centralités de vecteur propre) ;

- il n'existe pas d'indicateur adapté à tous les types de données et à toutes les questions que l'on se pose. Mesurer l'intermédiarité est très courant mais réfléchir à son interprétation sur certains types de réseaux est peut-être utile.

J'insiste : mesurer tout ce qui est possible grâce à un logiciel donné puis chercher à comprendre les résultats est un très bon moyen pour 1. perdre son temps, 2. perdre son énergie et 3. commenter de manière impressionniste des résultats que l'on ne comprend pas très bien. Et, au pire, raconter ensuite n'importe quoi...

1. Voir les *Synthèses méthodologiques* de la collection fmr sur HAL.
2. Dans le manuel de Wasserman et Faust, le terme *global* renvoie au réseau dans son ensemble ; le terme *local* à un sous-graphe de ce réseau (1994, p. 507). Merci à Paul Gourdon d'avoir attiré mon attention sur ce point dans sa thèse (2021, p. 254).

5.1 Connexité, densité, diamètre

Dans les chapitres précédents, nous avons déjà rencontré un certain nombre d'indicateurs permettant de caractériser un réseau : l'ordre (nombre de sommets, *order*), la taille (nombre de liens, *size*), le nombre de composantes connexes (*components*), le nombre de sommets isolés (*isolates*), la présence éventuelle de points d'articulation (*cut-points*, *cut-vertices*, *articulation points*) et d'isthmes (*bridges*). Il est toujours utile de rappeler ces mesures de base dans la mesure où de nombreux indicateurs sont sensibles à ces caractéristiques.

Si le réseau étudié n'est pas connexe, mesurer la proportion de sommets et de liens présents dans les différentes composantes est utile : la grande majorité des réseaux non connexes issus de données empiriques comporte une composante principale comprenant la grande majorité des sommets et des liens, on parle souvent de composante géante (*giant component*). Si votre réseau présente une configuration autre, cela vaut la peine d'être mis en évidence et si possible expliqué.

Densité

La densité (*density*) d'un réseau désigne le ratio entre le nombre de liens présents dans un réseau et le nombre de liens possibles. Elle varie entre 0 (réseau vide, *i.e.* absence de liens) et 1 (réseau complet, *i.e.* tous les liens possibles sont présents) et, multipliée par 100, elle peut s'exprimer en pourcentage (dans tel réseau, x% des liens possibles sont présents). Ces valeurs extrêmes sont évidemment des bornes qu'on ne rencontre jamais en pratique - un réseau vide ou un réseau complet serait peu intéressant à étudier.

Soit un réseau G avec V sommets et E liens. La formule permettant de calculer la densité est la suivante :

Réseau non orienté	Réseau orienté	Réseau planaire
$\frac{2E}{V(V-1)}$	$\frac{E}{V(V-1)}$	$\frac{E}{3(V-2)}$

En géographie des transports, la densité est parfois appelée indice gamma (γ *index*, encadré 5.2).

Il arrive dans certains cas que le nombre maximal de liens possibles doive être calculé différemment. Ainsi, les enquêtes sociométriques qui imposent une limite au nombre de réponses nécessitent d'adapter la formule.

> **Exemple** : dans l'enquête de Coleman *et al.* portant sur l'adoption d'innovation dans le milieu médical, il est demandé aux médecins à qui elles s'adressent pour obtenir des conseils. Les V médecins interrogées doivent fournir 3 noms maximum [1] : si les personnes enquêtées respectent les consignes, ce qu'il convient de vérifier, le nombre maximal de liens possibles est $3V$.

L'indicateur peut être utile pour comparer des réseaux de taille similaire, même s'il ne permet pas de différencier finement leur structure (la ligne et l'étoile de la figure 3.6 ont la même densité). Par contre, il est sensible à la taille et ne peut donc servir à comparer des réseaux d'ordres différents. Dans la plupart des réseaux, qu'ils s'agissent de réseaux

1. « Each doctor interviewed was asked three sociometric questions : To whom did he most often turn for advice and information ? With whom did he most often discuss his cases in the course of an ordinary week ? Who were the friends, among his colleagues, whom he saw most often socially ? In response to each of these questions, the names of three doctors were requested. », Coleman *et al.*, 1957, p. 254.

complets ou de réseaux personnels, plus le nombre de sommets augmente et plus la densité tend à baisser. L'exemple des relations interpersonnelles permet de le comprendre intuitivement : dans un collectif de 20 personnes, tout le monde connaît le prénom de tout le monde ; dans un collectif de 300 personnes, c'est beaucoup plus rarement le cas.

L'indicateur étant sensible à la taille, il est rarement pertinent, lorsqu'on étudie un réseau non connexe, de calculer les densités des différentes composantes pour les comparer les unes aux autres.

Les termes de réseau dense ou de réseau clairsemé (*sparse network*) utilisés pour décrire les réseaux en fonction de leur densité dépendent davantage des données et des questions de recherche que de la densité *stricto sensu*.

L'intensité des liens n'est généralement pas prise en compte pour le calcul de la densité. L'une des options proposées par Wasserman et Faust consiste à calculer l'intensité moyenne des liens (1994, p. 143). Ceci est peu satisfaisant car 1. cela ne renseigne pas sur la densité et 2. utiliser la moyenne n'est pertinent que si la distribution des intensités est approximativement gaussienne. En pratique, la méthode la plus utilisée aujourd'hui encore est de choisir un seuil (en le justifiant), de dichotomiser les liens en fonction de ce seuil (0 si l'intensité du lien est inférieure, 1 sinon) puis d'utiliser les formules vues précédemment. Il est également possible d'adapter la formule en considérant que le nombre maximal de liens possibles correspond au cas où chaque lien est porteur de l'intensité maximale rencontrée dans le réseau.

Comme toutes les mesures portant sur le réseau dans son ensemble, la densité seule est délicate à interpréter. Si je l'utilise pour comparer un même réseau à différents moments ou différents réseaux d'ordres similaires, je peux être plus loquace pour commenter les résultats.

Diamètre et longueur moyenne

Le diamètre correspond à la longueur, mesurée en nombre de liens, du plus long des plus courts chemins dans un réseau connexe. Si le réseau est composé de plusieurs composantes, chacune est susceptible d'avoir un diamètre différent. Le diamètre donne une indication sur la compacité ou le caractère étiré du réseau étudié.

Que le réseau soit planaire ou non ne change rien au diamètre. Si le réseau est valué, le diamètre correspond au plus long des plus courts chemins dont l'intensité totale est minimale. Il est possible alors d'obtenir deux diamètres différents : un diamètre topologique (plus long des plus courts chemins mesuré en nombre de liens) et un diamètre pondéré (plus long des plus courts chemins minimisant l'intensité) (figure 5.1).

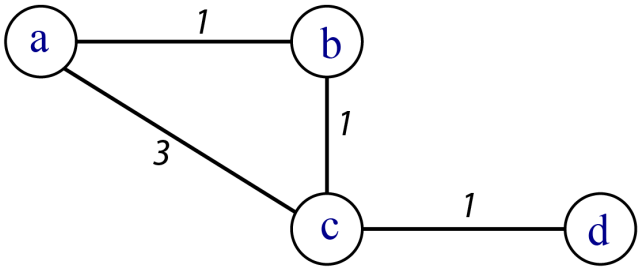
Une mesure différente, celle de la longueur moyenne des plus courts chemins, est devenue populaire en analyse de réseau, notamment grâce aux travaux de Milgram et à la notion de petit monde. L'expérience de Milgram dont il est question ici n'est pas la fameuse expérience concernant la soumission à l'autorité [1] mais celle sur les « six degrés de séparation [2] ». En deux mots, des personnes au centre des États-Unis sont censées envoyer un courrier à une

1. Sur la popularité délirante de certaines expériences sociologiques nord-américaines des années 1960-1970, expériences souvent fragiles, je renvoie à l'ouvrage de Thibault Le Texier, 2018. S'il étudie l'expérience bidonnée dite de Stanford, il évoque celle de Milgram sur la soumission à l'autorité pp. 133-135.

2. Cette expression, apparemment commune dans le monde anglophone, et totalement inconnue dans le monde francophone, correspond à l'idée selon laquelle six liens nous séparerait de n'importe quelle autre personne située n'importe où sur Terre.

personne qu'elles ne connaissent pas à Boston. Elles doivent l'adresser à une personne qu'elles connaissent dont elles supposent qu'elles pourraient connaître la personne cible. L'expérience n'a pas un succès fou (Kleinfeld, 2002), 80% des paquets s'égarent en chemin mais les 20% qui arrivent à destination ont une longueur médiane de 5 (étendue de 2 à 10).

FIGURE 5.1 – Diamètre topologique et pondéré



Si je ne prends pas en compte la valuation des liens, le diamètre du réseau est égal à 2 (liens {ac,cd}). Si je prends en compte la valuation, le diamètre est égal à 3 liens {ab, bc,cd} et intensité totale égale à 3).

La notion de plus court chemin doit être maniée avec précaution lorsque la distance est supérieure à 2, notamment dans les réseaux sociaux. Supposons que je cherche un emploi. Je peux demander à mes connaissances si elles ont des pistes. Ces connaissances, si elles en ont le temps, l'énergie et la volonté, peuvent éventuellement mobiliser leurs propres connaissances. Il est cependant illusoire d'imaginer qu'une amie d'une amie d'une amie me contacte pour me proposer du travail. En d'autres termes, pour nombre de relations, une distance de 3 ou de 25 revient exactement au même, elle signale une ressource inaccessible. On pourrait d'ailleurs généraliser cette limite : il est rare que je choisisse un voyage en train ou en avion qui m'impose plus de 2 changements ; si je navigue sur internet à la recherche d'une information, je commence par la liste de résultats de mon moteur de recherche, vais voir l'un de ces sites puis éventuellement l'un des sites indiqués sur le site en question mais il est rare que j'aille plus loin.

Les mesures qui suivent permettent de caractériser des éléments individuels du réseau (sommets essentiellement) ou d'étudier des micro-configurations (deux ou trois sommets et les liens éventuels entre ces sommets). Ils peuvent aussi être utilisés pour caractériser le réseau dans son ensemble.

5.2 Mesures de centralité

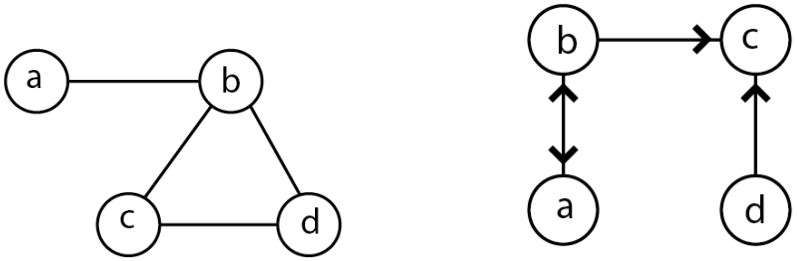
L'une des questions de recherche les plus courantes en analyse de réseau concerne la centralité des individus : est-ce que certains individus sont plus actifs, plus populaires, plus accessibles (tout dépend du type d'individus et de relations) que d'autres et comment l'expliquer ? Cette question étant souvent traitée, il existe bien sûr différentes façons d'y répondre.

Degré

La mesure de centralité la plus simple à calculer est le degré (*degree*) : il s'agit du nombre de liens adjacents à un sommet, que le réseau soit planaire ou non planaire. Si

le réseau étudié est orienté, on distingue les degrés entrants (*in-degree*), nombre de liens ayant pour destination le sommet étudié, et le degré sortant (*out-degree*), nombre de liens ayant pour origine le sommet étudié. L'examen de la matrice d'adjacence correspondant au réseau montre que le degré correspond aux sommes marginales de la matrice (figure 5.2). Par ailleurs, dans un réseau orienté, le nombre total de liens entrants est égal au nombre total de liens sortants.

FIGURE 5.2 – Degrés et sommes marginales



	a	b	c	d	∑
a	0	1	0	0	**1**
b	1	0	1	1	**3**
c	0	1	0	1	**2**
d	0	1	1	0	**2**
∑	**1**	**3**	**2**	**2**	8

	a	b	c	d	∑
a	0	1	0	0	**1**
b	1	0	1	0	**2**
c	0	0	0	0	**0**
d	0	0	1	0	**1**
∑	**1**	**1**	**2**	**0**	4

Dans les deux cas, le réseau est simple donc la diagonale de la matrice d'adjacence est vide. La matrice du réseau non orienté à gauche est symétrique, les sommes marginales sont donc égales. Le sommet b *est celui qui a le degré le plus élevé ; le sommet* a *celui qui a le degré le plus faible. Lorsque le réseau est orienté, il est nécessaire d'examiner degré entrant d'un côté et degré sortant de l'autre. Ici,* c *est le plus « attractif » (deux liens entrants),* b *le plus « actif » (deux liens sortants).*

Le degré dans un réseau unimodal non planaire simple (pas de boucle, pas de liens multiples) d'ordre V varie entre 0 (sommet isolé) et V - 1. Il est donc possible de normaliser le degré en le divisant par V - 1. Il arrive que certains sommets aient des degrés très élevés et on parle alors de *hubs*. Si tous les sommets d'un réseau ont le même degré k, on parle de réseau k-régulier ; ceci n'arrive évidemment jamais avec des données empiriques mais permet de construire un modèle de réseau (cf *infra* la description des réseaux petit-monde).

En fonction de la nature des données, l'interprétation du résultat varie mais le fait qu'un sommet soit voisin de nombreux autres sommets est signe qu'il occupe une place importance dans le réseau étudié. Lorsque le réseau est orienté, un fort degré sortant peut généralement être considéré comme le marqueur d'une forte *activité*, un fort degré entrant comme celui d'une forte *attractivité*. Si l'on examine les interactions sur un site de rencontres, envoyer de nombreux like - degré sortant élevé - n'a pas le même sens qu'en recevoir beaucoup - degré entrant élevé. Tester la relation entre ces deux indicateurs peut être utile [1].

1. En règle générale, il est utile de tester les relations entre les différents indicateurs calculés sur les sommets, notamment pour détecter d'éventuelles redondances.

Lorsque le réseau est valué, je peux choisir de mesurer le degré sans prendre en compte l'intensité des liens. Cela est pertinent thématiquement si je considère que la présence ou l'absence d'un lien, quelle que soit son intensité, est le critère le plus pertinent pour étudier le phénomène observé. Je peux également calculer le degré en sommant les intensités des liens adjacents à un sommet, je calcule alors un degré pondéré (*weighted degree*, on rencontre parfois le terme *strength*). Si le réseau est à la fois valué et orienté, je distinguerai des degrés pondérés entrants (somme des intensités reçues par un sommet) et des degrés pondérés sortants (somme des intensités émises par un sommet). La normalisation des degrés pondérés se fait généralement sur la somme des intensités présente dans le réseau étudié.

Si l'étendue des intensités est forte (commerce international entre États par exemple), je peux transformer mes données d'intensité avant les calculs pour obtenir des résultats plus faciles à interpréter. Il est également possible de transformer des variables continues en variables ordinales (volume faible, moyen ou fort par exemple).

Le degré est une mesure simple et très employée. Elle est calculée pour chaque sommet et il est donc possible de calculer les paramètres statistiques de centralité et de dispersion de cet indicateur (degré moyen, étendue, variance, etc.). L'analyse de la distribution des degrés est devenue une démarche classique en analyse de réseau, elle est évoquée dans la section consacrée aux modèles de réseau (section 5.4).

Degré et voisinage

La mesure du degré prend en compte les seuls voisins immédiats des sommets ; on parle des voisins d'ordre 1 (sommets situés à un lien de distance). Il peut être intéressant au niveau thématique de prendre en compte des voisinages plus larges. Si j'étudie des relations professionnelles, avoir de nombreux contacts est important mais avoir de nombreux contacts avec des personnes qui ont elles-mêmes de nombreux contacts est sans doute un avantage.

Prendre en compte le voisinage d'ordre 2 revient à pondérer les degrés des sommets par les degrés de leurs voisins d'ordre 1. Différentes formules ont été proposées des années 1950 à nos jours pour prendre en compte ces paramètres : centralité de Katz (1953), centralité de vecteur propre (Bonacich, 1987), PageRank, etc. [1].

Le calcul de ces différents indicateurs repose sur le calcul des vecteurs propres de la matrice d'adjacence (cf annexe A) et suppose de choisir la valeur des paramètres utilisés pour ce calcul [2]. Certes, la mesure est un petit peu plus complexe à prendre en main mais elle peut être utile, notamment pour caractériser de manière fine des sommets ayant de même degré (figure 5.4).

Cette mesure peut être adaptée sans difficulté particulière aux réseaux planaires et/ou valués.

Intermédiarité

Le degré prend en compte les voisins d'ordre 1, les indicateurs basés sur les vecteurs propres les voisinages d'ordre 2, l'intermédiarité (*betweenness*) prend en compte la structure

1. On trouvera dans le manuel de Newman un exposé clair et synthétique des avantages et inconvénients des différentes formules disponibles (2018, pp. 159-167).
2. Garder les valeurs par défaut du logiciel utilisé sans savoir ce qu'elles signifient est une option possible mais peu satisfaisante.

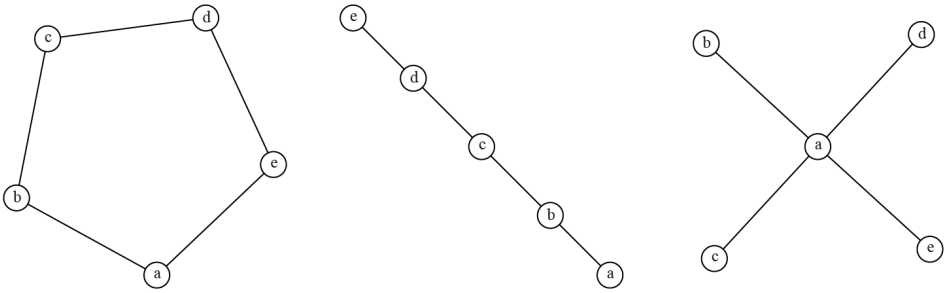
complète du réseau. L'intermédiarité d'un sommet est le rapport entre le nombre de plus courts chemins passant par un sommet donné et le nombre de plus courts chemins du réseau étudié. Les sommets les plus centraux ne seront pas nécessairement ceux qui ont le plus de liens mais ceux qui sont des points de passage obligés pour passer d'une zone du réseau à une autre.

Cet indicateur peut être calculé pour les sommets, il peut également être calculé pour les liens (*edge betweenness*) : un lien sera considéré comme plus central s'il est situé sur de nombreux plus courts chemins entre paires de sommets.

Trois remarques apparaissent nécessaires concernant cet indicateur très fréquemment utilisé pour tous types de données relationnelles :

- on suppose implicitement que seuls les plus courts chemins sont employés ;
- le calcul prend en compte la structure du réseau dans son ensemble, que celle-ci soit connue ou non des acteurs ;
- son interprétation a plus ou moins de sens selon les données étudiées.

FIGURE 5.3 – Réseaux idéaux-typiques et centralités



	a	.	e	a	b	c	d	e	a	b	c	d	e
Degré	2	.	2	1	2	2	2	1	4	1	1	1	1
Interm.	1	.	1	0	3	4	3	0	6	0	0	0	0
Proxim.	1/6	.	1/6	1/10	1/7	1/6	1/7	1/10	1/4	1/7	1/7	1/7	1/7

*Dans le réseau de type cercle, tous les sommets ont les mêmes centralités : un même degré 2 (le réseau est 2-régulier) ; une intermédiarité égale à 1 (*a *a pour voisins* b *et* c *et un seul plus court chemin le traverse) et une centralité de proximité égale à 1/6 (2 plus courts chemins de 1 plus 2 plus courts chemins de longueur 2).*
Dans le réseau ligne, les extrémités ont des scores faibles voire nuls (aucun plus court chemin ne les traverse) et ces scores augmentent à mesure qu'on se rapproche du sommet central. Dans le réseau étoilé, le sommet central a un degré égal à $V-1$ *et tous ses voisins ont les mêmes faibles scores de centralité.*

Mesures pour réseaux planaires

Différents indicateurs mis au point dans les années 1960 par des géographes quantitativistes nord-américains visent à caractériser des réseaux d'infrastructures pouvant être modélisés sous forme de réseaux planaires. Ces indicateurs sont peu connus et peu utilisés hors de la géographie des transports et la liste qui suit n'est pas exhaustive, elle est extraite de l'ouvrage Haggett et Chorley (1969, p. 32).

Soit un réseau planaire R, composé de C composantes, avec V sommets et E liens :
Nombre cyclomatique (μ) : $\mu = E - V + C$. Indicateur sur le nombre de circuits présents dans le réseau. Il varie de 0 (arbre) à 1 (réseau complet) ;
Indice α : rapport entre le nombre de circuits présents (μ) et le nombre de circuits possibles ($2E - 5$). Plus il est élevé, plus le réseau est redondant donc coûteux à entretenir mais peu vulnérable (si une route est coupée, d'autres routes existent) ;
Indice β : nombre de liens divisés par le nombre de sommets ;
Indice γ : nombre de liens présents divisés par le nombre de liens possibles.

Dans sa thèse (1963), Kansky propose également des indices d'intensité moyenne, notamment les indices η (longueur moyenne des liens), π (longueur totale du réseau divisée par son diamètre), θ (volume moyen par lien) et ι (rapport entre kilomètrage total et volume transporté). Le chapitre correspondant est accessible en ligne (Dancoisne et Kansky, 1989), le lien entre ces indicateurs et la théorie des graphes est évoqué une entrée de l'encyclopédie en ligne *Hypergeo* (Beauguitte, 2020) ; une version traduite et commentée de la thèse de Kansky est disponible dans la collection « textes » du groupe fmr.

Nombre associé

Une mesure directement issue de la théorie des graphes, celle de nombre associé (*associated number*), peut être utile pour déterminer des sommets centraux. Le nombre associé d'un sommet v_i est la longueur du plus long des plus courts chemins entre v_i et les autres sommets du graphe. Dans le réseau à gauche de la figure 5.2, les nombres associés des différents sommets sont a : 2, b : 1 ; c : 2 et d : 2. Le ou les sommets ayant le nombre associé le plus faible sont considérés comme des sommets centraux.

Proximité

Différents indicateurs existent pour mesurer la proximité moyenne entre un sommet et tous les autres sommets du réseau, ce dernier pouvant être planaire et/ou valué.

L'un des plus employés en analyse des réseaux sociaux est la centralité de proximité (*closeness*). Il s'agit de mesurer la distance moyenne des plus courts chemins entre un sommet v et l'ensemble des autres sommets du réseau. Afin que les sommets les plus centraux pour ce critère aient les scores les plus élevés, on calcule généralement l'inverse de cette distance moyenne.

Si l'on étudie des logiques de circulation d'information, le sommet ayant une forte centralité de proximité peut transmettre rapidement (*i.e.* avec le moins d'intermédiaires possibles) une information à l'ensemble des autres sommets.

Cette mesure est peu adaptée aux réseaux non connexes dans la mesure où la distance entre deux sommets appartenant à deux composantes différentes est, par convention, considérée comme infinie. Il est nécessaire dans ce cas de calculer cet indicateur composante par

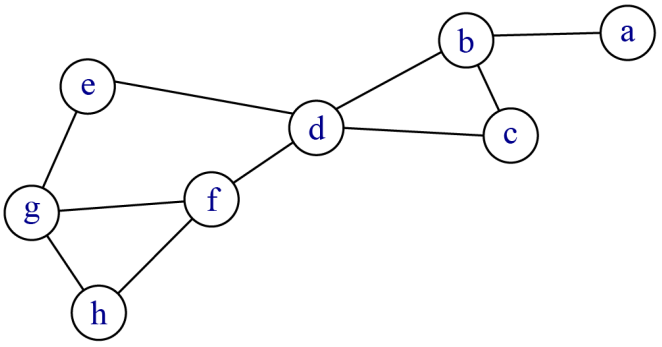
composante mais la comparaison n'a de sens qu'entre les sommets situés dans la même composante.

Cet indicateur est en partie lié à l'ordre du réseau (plus le nombre de sommets est faible, plus les distances moyennes tendent à être faibles) et au diamètre (plus il est faible, plus les distances moyennes le sont).

Si les liens portent une intensité (réseau valué), la longueur des plus courts chemins peut être mesurée en sommant ces intensités. Les deux indices de Shimbel [1], régulièrement utilisés pour l'analyse des réseaux de transport, prennent en compte une mesure de distance (généralement kilométriques) entre les sommets :

- l'indice d'accessibilité, calculé pour chaque sommet, est la somme de la longueur des plus courts chemins entre un sommet et les autres sommets du réseau connexe. Plus l'indice est élevé, plus le sommet est considéré comme éloigné des autres ;
- l'indice de dispersion porte sur le réseau connexe dans son ensemble et consiste à sommer tous les plus courts chemins présents dans le réseau. Plus l'indice est bas, plus le réseau est considéré comme compact.

FIGURE 5.4 – Centralités de vecteur propre



Dans ce réseau, les sommets b, g *et* f *ont un degré de 3. Le tableau compare les résultats obtenus avec la centralité de vecteur propre et l'algorithme PageRang. Dans les deux cas,* a *obtient le plus faible score (degré de 1 et voisin de degré 3) et* d *le score le plus élevé (degré de 4, 2 voisins de degré 2 et 2 de degré 3). On pourrait s'attendre à ce que* c *et* e *aient les mêmes scores : tous deux ont un degré de 2, un voisin de degré 4 (*d*) et un de degré 3 ; leurs scores sont identiques dans un cas et très proches dans l'autre.*

	a	b	c	d	e	f	g	h
Eigen.	0.24	0.67	0.61	1.00	0.63	0.84	0.75	0.58
PageRank	0.06	0.16	0.10	0.19	0.10	0.14	0.14	0.10

Pour obtenir les mêmes résultats avec deux logiciels différents, il est nécessaire de vérifier les paramètres utilisés par défaut par ces logiciels.

Toutes les mesures évoquées dans cette section concernent les sommets pris individuellement, plus les liens pour l'intermédiarité. L'étude de la distribution de ces indicateurs, et notamment celle des degrés, est une pratique devenue très courante en analyse de réseau. Dans la mesure où elle permet de distinguer différents modèles de réseaux, elle est évoquée dans la section 5.4.

1. Les deux articles de Shimbel (1951 et 1953) sont disponibles en version bilingue et commentée dans la collection « textes »du groupe fmr. Shimbel apparaît régulièrement dans les manuels de *Social Network Analysis*, non pas pour ces deux mesures, mais parce qu'à la fin de son article de 1953, il évoque une mesure de *stress* dont la description annonce la centralité d'intermédiarité.

Les distributions des autres indicateurs sont beaucoup plus rarement étudiées alors qu'elles peuvent s'avérer intéressantes. Si on souhaite utiliser ces indicateurs pour des analyses multivariées, tester leurs relations éventuelles est utile pour éviter d'inclure des variables redondantes : il n'est pas rare que les sommets les plus centraux le soient selon les différents indicateurs évoqués ici et cela est vrai également pour les sommets les moins centraux (un sommet isolé ou de degré 1 a une intermédiarité nulle). Pour faciliter la comparaison des résultats, normaliser les résultats ou les passer en rangs peut être utile.

Toutes les mesures évoquées précédemment portaient sur des éléments considérés individuellement, les mesures qui suivent s'intéressent aux micro-configurations formées par les dyades (deux sommets et les liens éventuels entre eux) et les triades (trois sommets et les liens éventuels entre eux).

5.3 Dyades, triades et motifs

Si le réseau étudié est non orienté, les liens dans les dyades sont soit présents soit absents et la densité est un indicateur suffisant. Par contre, si le réseau est orienté, il est intéressant de calculer la proportion de liens mutuels (v_1 envoie un lien vers v_2 et v_2 envoie un lien vers v_1) et asymétriques (présence du lien $\{v_1, v_2\}$ ou du lien $\{v_2, v_1\}$). On peut calculer un indicateur de réciprocité (*reciprocity*) en divisant le nombre de liens mutuels par le nombre de liens total.

Une mesure plus récente concernant les dyades est l'assortativité (*assortativity* ou *assortative mixing*) : il s'agit d'une mesure de l'homophilie des sommets voisins. Cette mesure peut porter sur une variable attributaire ou sur une variable structurale [1] comme le degré. La formalisation mathématique est un peu longue à expliquer mais le principe est simple : on compare le nombre de liens dont les deux sommets partagent la même caractéristique au nombre de liens de ce type qu'on obtiendrait dans un réseau où les liens seraient placés au hasard ; on parle dans ce dernier cas de graphe aléatoire (*random graph*).

Quand la variable attributaire testée est qualitative, une mesure possible de l'assortativité repose sur le calcul de la modularité (*modularity*), indicateur que nous retrouverons dans la section consacrée à la détection de communautés (6.4) ; quand la variable attributaire testée est quantitative, la mesure peut reposer sur un coefficient de corrélation de Pearson.

Lorsque la mesure est inférieure à 0, le réseau est disassortatif : les liens existent plus souvent qu'attendu entre individus ayant des attributs différents. Lorsqu'elle est supérieure à 0, le réseau est dit assortatif et les liens sont présents plus souvent qu'attendu entre individus partageant le même attribut. Il est possible de normaliser cette mesure pour la faire varier entre 0 et 1 ; sa version normalisée est appelée coefficient d'assortativité.

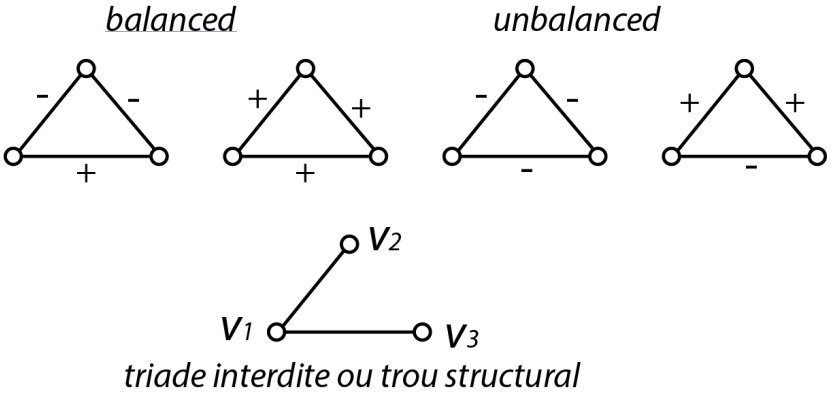
Triades et transitivité

Une triade désigne un ensemble de trois sommets et les liens présents entre ces trois sommets. Cette microconfiguration a fait l'objet de plusieurs théorisations en sociologie depuis le milieu des années 1960, qu'il s'agisse de la notion d'équilibre, de trou structural ou de triade interdite.

1. L'adjectif structural désigne une propriété liée à la structure du réseau. Mes sommets peuvent avoir x attributs exogènes (pour des personnes, âge, pays de naissance, genre, etc.) ; la centralité de degré, de vecteur propre, etc. sont des variables structurales ou endogènes.

Un même objet, trois formalisations théoriques

Dans leur manuel de 1965, Harary *et al.* proposent une conceptualisation sur le caractère équilibré ou non (*un-balanced*) des triades. Ils étudient des réseaux signés où les liens sont porteurs d'une intensité positive ou négative. Remplacer le signe par une expression du type « apprécie - n'apprécie pas » permet de saisir intuitivement la notion d'équilibre dans le réseau.
Granovetter propose une conceptualisation très proche avec la notion de triade interdite (*forbidden triad*). Si l'on prend une triade $\{v_1, v_2, v_3\}$ et que des liens intenses existent entre v_1 et v_2 et entre v_1 et v_3 alors il est très probable qu'un lien fort existe également entre v_2 et v_3. La raison est liée au coût (en temps, en énergie) que prend l'entretien d'un lien fort : si mes deux meilleures amies ne se supportent pas (cas de triade interdite), je vais avoir du mal à maintenir ces deux liens.
Dans un cadre différent, Burt utilise la même triade pour proposer le concept de trou structural (*structural hole*). L'absence de lien entre v_2 et v_3 procure un avantage stratégique à v_1 qui peut choisir de faire circuler ou de bloquer (d'où les termes anglophones de *broker* et *brokerage*) une ressource quelle qu'elle soit.

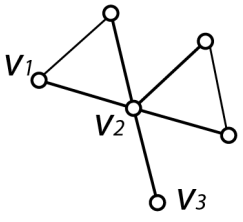


Il est important lorsqu'on étudie nos données de se rappeler qu'une même configuration peut se prêter à des interprétations très différentes les unes des autres, en fonction notamment du sens donné aux relations entre les entités.

L'une des mesures les plus courantes est la recherche de la transitivité (*transitivity*), parfois appelée *clustering coefficient*. Il s'agit de mesurer au niveau des sommets et/ou au niveau du réseau dans son ensemble la proportion entre le nombre de triades fermées et le nombre de triades ouvertes.

Si le réseau (non planaire) est complet, la transitivité est égale à 1 au niveau local et au niveau global. Inversement, tout réseau en forme d'arbre (absence de cycle) a une transitivité nulle, localement et globalement.

Dans la figure ci-contre, le sommet v_1 a une transitivité égale à 1 : il a deux voisins et ses deux voisins sont eux-même voisins ; le sommet participe à une seule triade et elle est fermée. Inversement, le sommet v_3 n'ayant qu'un seul voisin, il n'est inclus dans aucune triade et l'indicateur ne peut être calculé. Le sommet v_2 a 5 voisins mais seules deux triades sont fermées, il a donc une transitivité de 0.2.



Si dans le cadre d'un petit guide pratique, s'arrêter au niveau de la triade apparaît suffisant, il existe un certain nombre de recherches, notamment en informatique [1], portant sur la détection de motifs de quatre sommets ou plus dans les réseaux.

Chaque fois que l'on procède à une mesure sur l'ensemble des sommets, on crée de fait un nouvel attribut. Comme toute variable quantitative, il est possible d'examiner sa distribution (minimum et maximum, quartiles, moyenne, écart-type, coefficient de variation). Il est également possible d'utiliser ces variables pour mener des analyses multivariées. Il est préférable de contrôler que les différents indicateurs calculés ne soient pas trop corrélées les uns aux autres.

Une autre possibilité est de mesurer l'écart entre les données obtenues pour un indicateur sur le réseau empirique étudié et le score qu'on obtiendrait sur le réseau le plus hiérarchique qui soit. C'est le principe des indicateurs de centralisation (*centralization*) disponibles dans certains logiciels : pour le degré par exemple, ils comparent la distribution observée à celle d'un réseau étoilé ayant le même nombre de sommets.

5.4 Des mesures aux modèles

Lorsqu'on étudie un réseau, à quoi le comparer et comment le qualifier ? Depuis les années 1950, il était possible de le comparer au graphe aléatoire (*random graph*) proposé par les deux mathématiciens Erdös et Rényi dans une série d'articles.

La construction d'un tel graphe est relativement simple : soit un ensemble de sommets et une probabilité de présence de liens entre ces sommets créant un ensemble de liens. Les deux auteurs ont montré qu'il se produit une bifurcation à mesure que la probabilité augmente : passé un certain seuil, variable en fonction du nombre de sommets, une composante géante rassemblant la grande majorité des sommets et des liens apparaît et la distribution des degrés tend à suivre une loi normale. On observe peu de valeurs extrêmes (faibles ou fortes), un mode correspondant à la moyenne et une courbe en cloche.

Il s'agit d'un modèle mathématique et les auteurs n'ont jamais prétendu que ces graphes aléatoires puissent avoir un quelconque intérêt pour étudier des réseaux issus de données empiriques. Pourtant, ces graphes aléatoires servent encore aujourd'hui pour les modèles statistiques de réseaux (chapitre 11).

Le modèle du réseau petit-monde (*small-world network*) proposé par Watts et Strogatz en 1998 se situe entre le graphe k-régulier et le graphe aléatoire (figure 5.5). En supprimant de façon itérative un lien dans le graphe k-régulier de départ et en le replaçant de façon aléatoire, on obtient un graphe qui présente deux propriétés intéressantes : il existe une forte transitivité locale (nombre important de triades fermées) mais la distance moyenne entre paires de sommets diminue rapidement (il y a besoin de peu d'intermédiaires pour passer d'un sommet à l'autre).

Le réseau sans-échelle (*scale-free network*) proposé par Barabási et Albert en 1999 correspond à un réseau très hiérarchisé : une poignée de sommets est très connectée (degré très élevé), la très grande majorité a peu de liens. La distribution des degrés peut être représentée par une droite sur une échelle logarithmique. Aucune valeur centrale (mode, moyenne) ne permet de qualifier cette distribution des degrés, d'où le terme sans échelle. Le

1. Voir notamment les travaux de Christophe Prieur, MCF en informatique devenu professeur de sociologie à l'Université Gustave Eiffel. Attention, le profil Google Scholar est celui d'un homonyme mathématicien grenoblois.

FIGURE 5.5 – Deux figures devenues iconiques

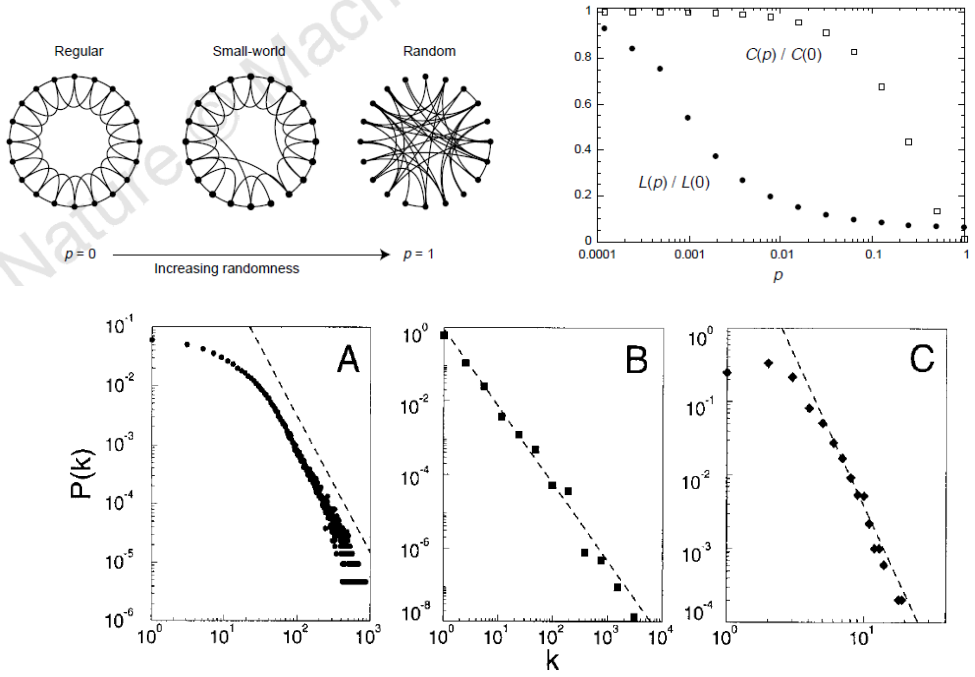


Fig. 1. The distribution function of connectivities for various large networks. (**A**) Actor collaboration graph with $N = 212{,}250$ vertices and average connectivity $\langle k \rangle = 28.78$. (**B**) WWW, $N = 325{,}729$, $\langle k \rangle = 5.46$ (6). (**C**) Power grid data, $N = 4941$, $\langle k \rangle = 2.67$. The dashed lines have slopes (A) $\gamma_{actor} = 2.3$, (B) $\gamma_{www} = 2.1$ and (C) $\gamma_{power} = 4$.

Figures extraites des articles de Watts et Strogatz (haut) et Barabási et Albert (bas). La première montre tout d'abord la construction des réseaux petit-monde, le graphique de droite croise distance moyenne entre paires de sommets L *et transitivité locale* C*, la première diminuant beaucoup plus vite que la seconde. La figure du bas montre la distribution des degrés de trois gros réseaux ; la comparaison des courbes est rendue difficile par la non harmonisation des abscisses mais dans les trois cas, la hiérarchie est très forte et calculer un degré moyen n'aurait pas grand sens.*

---

mécanisme expliquant l'apparition de cette structure est nommé par les autrices l'attachement préférentiel : un sommet nouveau dans un réseau tend à se connecter aux sommets les plus connectés. Ce mécanisme fonctionne pour un grand nombre de phénomènes sociaux comme les réseaux de transport, les réseaux de citations ou les liens entre sites internet.

Ces deux modèles ont eu un très fort impact, tant chez les physiciennes analystes de réseau que dans les autres disciplines. On trouve facilement en ligne des milliers d'articles visant à déterminer si tel ou tel réseau (aérien, lexical, social, numérique, etc.) peut être considéré comme sans-échelle et/ou petit-monde. Un grand nombre de phénomènes sociaux étant très hiérarchisés, il n'est pas rare de trouver des réseaux sans-échelle ; la très grande majorité des réseaux issus de données empiriques n'étant ni aléatoires ni k-réguliers, ils sont le plus souvent petit-monde. Ceci étant, si dans votre discipline, la vague n'est pas encore retombée, ne vous privez surtout pas de publier ce type de résultats.

Chapitre 6

Simplifier, partitionner

Deux démarches complémentaires sont présentées dans ce chapitre :
- les démarches de simplification d'un réseau ;
- les démarches de partitionnement des sommets.

Dans le premier cas, il s'agit de se débarrasser d'éléments (sommets et/ou liens) considérés comme peu utiles pour l'analyse ; dans le deuxième cas, de créer des classes de sommets basées sur les propriétés structurales du réseau étudié. Les méthodes de partitionnement sont nombreuses et seules celles couramment implémentés dans les logiciels les plus utilisés sont présentées ici.

Toutes les démarches présentées ne sont pas adaptées à tous les réseaux étudiés et, comme pour les mesures présentées dans le chapitre précédent, il est prudent de comprendre ce que l'on fait (propriétés générales de l'algorithme choisi, rôle des paramètres) avant de commenter les résultats obtenus, certaines méthodes ayant un aspect boîte noire quelque peu gênant (cf *infra* l'exemple de la méthode CONCOR).

6.1 Supprimer des éléments

Une fois les données relationnelles mises en forme, une fois les mesures de base effectuées, il peut être utile de simplifier le réseau obtenu. C'est notamment le cas lorsque le réseau est gros et/ou dense. Les méthodes de simplification visent généralement à préserver la structure d'ensemble du réseau tout en éliminant au maximum les phénomènes considérés comme peu significatifs. Trois démarches complémentaires peuvent être mises en œuvre pour faciliter l'analyse et/ou proposer des visualisations lisibles du ou des réseaux étudiés (ce dernier aspect sera abordé dans le chapitre 12) : supprimer des liens, agréger des sommets et enfin supprimer des sommets. Dans tous les cas, il est bien sûr nécessaire de décrire la méthode utilisée, de la justifier et de ne pas oublier les transformations effectuées lorsqu'on commente les résultats obtenus sur le réseau simplifié.

Arbre couvrant minimum et flux dominants

La recherche de l'arbre couvrant minimum (*minimum spanning tree*) peut permettre de mieux comprendre la structure générale du réseau, que celui-ci soit ou non valué. Comme le nom le suggère, il s'agit d'un arbre (suppression de tous les cycles présents dans le réseau étudié) couvrant (il passe par tous les sommets) et il contient le nombre de liens le plus faible possible. Cette manière de supprimer des liens peut donner des résultats plus ou

moins probants en fonction des données étudiées. Lorsque le réseau est valué, l'objectif est d'obtenir un arbre où la somme du poids des liens est minimale. Les réseaux d'infrastructures se prêtent généralement bien à ce type de simplification ; les réseaux sociaux très denses avec un diamètre faible gagnent peu à être simplifiés ainsi.

Une autre méthode visant à supprimer des liens valués a été proposé par les géographes étasuniens Nystuen et Dacey en 1961[1]. L'objectif est, en partir d'une matrice de flux orientés, de créer un arbre (ou une forêt) hiérarchisant les sommets.

Les flux sont sélectionnés en appliquant les deux règles suivantes :
- pour chaque sommet v_i, ne garder que le flux sortant le plus important f_i ;
- si le flux est émis vers un sommet v_j dont le degré entrant pondéré est inférieur à celui de v_i, supprimer ce flux.

On obtient quatre catégories de sommets :
- les sommets isolés qui ne dominent personne et ne sont pas dominés ;
- les sommets dominés avec un lien sortant ;
- les sommets intermédiaires avec un ou plusieurs liens entrants et un lien sortant ;
- les sommets dominants avec uniquement des liens entrants.

Cette méthode simple crée une partition et une hiérarchie entre sommets. À ma connaissance, cette méthode n'a jamais été testée en dehors de la géographie quantitative où elle donne souvent des résultats intéressants et des régionalisations convaincantes.

Les autres méthodes disponibles pour sélectionner des liens valués sont généralement basées sur des critères statistiques et elles supposent d'étudier la distribution des intensités. La grande majorité des matrices de flux suivent des distributions de Paretto (20 % des liens concentrent au moins 80 % du volume) : il est donc souvent possible de supprimer les liens les plus faibles, qui sont aussi les plus nombreux, tout en conservant la très grande majorité du volume du flux étudié.

Suppression et/ou agrégation de sommets

Une des méthodes les plus répandues pour mieux appréhender la structure globale d'un réseau consiste à supprimer les sommets les moins connectés, qu'ils soient isolés ou de degré faible. Il est recommandé de rappeler la proportion de sommets supprimés lorsqu'on commente les résultats obtenus sur le réseau ainsi modifié.

Inversement, dans certains cas, il peut être intéressant de supprimer les sommets les plus centraux. Si j'étudie les relations diplomatiques au niveau mondial, il est propable que tous les États ou presque entretiennent des relations avec les États-Unis, la Belgique (UE oblige) et la Chine. Supprimer ces liens attendus peut permettre de mettre en évidence des configurations régionales intéressantes.

Une troisième option consiste à agréger les sommets qui occupent la même position dans le réseau ; on parle dans ce cas d'isomorphisme. Les méthodes développées dans les sections suivantes permettent de créer des partitions, partielles ou exhaustives, des sommets afin d'obtenir des représentations simplifiées du réseau analysé.

1. Une version bilingue et commentée est disponible dans la collection « textes » du groupe fmr.

Si le réseau étudié est particulièrement dense, il peut être intéressant, plutôt que de supprimer des liens et/ou des sommets, de se pencher sur les *liens manquants* en étudiant le complément du graphe [1].

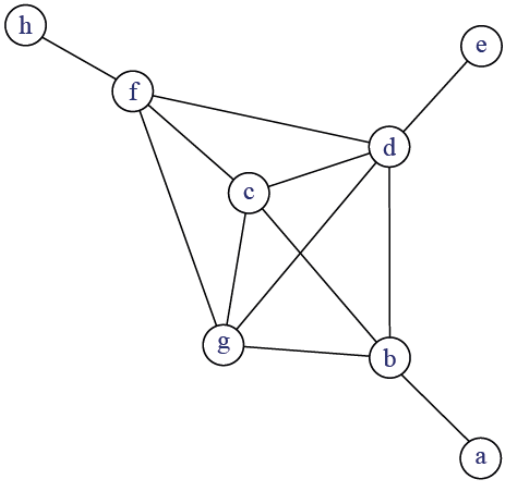
6.2 Rechercher les cliques

La recherche de sous-ensembles denses dans un réseau présente un intérêt, notamment que le réseau étudié a une structure de type centre - périphérie ou, au contraire, lorsqu'il est éclaté en différents sous-réseaux denses.

L'une des définitions les plus anciennes, et qui peut garder sa pertinence si le réseau est de taille réduite, est celle de clique (terme identique en anglais [2]). Une clique est un sous-graphe maximal complet de trois sommets ou plus. Si la définition brille par sa concision, elle mérite d'être expliquée terme à terme. Un sous-graphe consiste en un sous-ensemble du réseau. Il est maximal si et seulement si il n'est pas possible d'y ajouter le moindre sommet sans lui faire perdre sa caractéristique. Enfin, il est complet : tous les liens possibles entre les sommets de ce sous-graphe sont présents. La contrainte sur le nombre de sommets évite que chaque lien ne soit compté comme une clique.

Un même sommet peut être présent dans plusieurs cliques (figure 6.1) : le terme d'*overlap* (chevauchement) est utilisé dans la littérature anglophone. Si la clique a une définition mathématique claire, elle présente deux inconvénients majeurs. Elle peut être considérée comme trop exigeante dans la mesure où il suffit qu'un seul lien manque pour que des sommets soient dans des cliques différentes. Par ailleurs, si elle a une utilité pour les réseaux de petite dimension, la recherche des cliques présente peu d'intérêt lorsque le nombre de sommets dépasse les centaines (nombre élevé de cliques de petite taille, chevauchements importants, temps de calcul élevé).

FIGURE 6.1 – Cliques



Intuitivement, on voudrait obtenir un ensemble de sommets centraux ($bcdfg$) et des sommets périphériques (aeh). L'absence du lien entre b et f entraine la présence de deux cliques différentes : $cdfg$ et $bcdg$.

1. Le complément du graphe, appelé aussi graphe complémentaire ou graphe inversé, du graphe simple G est le graphe G' tel que deux sommets de G' sont adjacents si et seulement si ils ne sont pas adjacents dans G.
2. Le terme a souvent été, et reste encore, employé dans la sociologie anglophone dans un sens métaphorique de petit groupe soudé d'individus ; ce n'est pas le sens retenu ici.

Un certain nombre de définitions plus souples ont donc été proposées en *SNA* (Wasserman et Faust, 1994, pp. 257-270 [1]). Un *k-core* est un sous-graphe maximal dans lequel chaque sommet à un degré minimal de k. La figure 6.1 comprend un 4-core constitué par les sommets $bcdfg$. La valeur de k est fixée par la chercheuse et dépend de la densité du réseau étudié (plus la densité augmente, plus je peux augmenter k).

Qu'il s'agisse des cliques ou des *k-cores*, ces méthodes sont principalement employées pour des réseaux non planaires, non orientés, unimodaux et binaires. Des adaptations ont été proposées pour les réseaux bimodaux (chapitre 7) ; pour les réseaux valués, une stratégie possible serait de tester différents seuils pour conserver les liens (supprimer les liens inférieurs à 1, à 2, etc.) et de voir quelles cliques restent présentes avec des seuils élevés. Il s'agirait alors moins de définir des cliques *stricto sensu* que de déterminer des zones d'intensité plus ou moins forte.

Si la recherche de cliques peut mettre en évidence des sous-réseaux denses, elle ne permet pas une partition de l'ensemble des sommets. À l'inverse, les méthodes de *blockmodeling* et les méthodes de détection de communautés créent des partitions exhaustives, tout sommet du réseau appartient à un bloc/une communauté, et généralement exclusives, aucun sommet ne peut être membre de plus d'un bloc/une communauté.

6.3 *Blockmodel* et équivalences

L'objectif du *blockmodel* est de produire une image simplifiée du réseau étudié. Il suppose trois étapes :

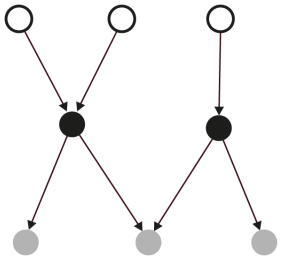
- ordonner la matrice d'adjacence afin de produire des ensembles présentant une forte densité interne ;
- agréger les ensembles détectés afin de créer une partition des sommets en des sous-ensembles discrets appelés positions ;
- créer la matrice d'adjacence entre ces positions (*image matrix*) ; le réseau entre ces positions est parfois qualifié de réseau réduit (*reduced graph*).

Plusieurs méthodes sont disponibles pour créer cette partition. Classiquement en *SNA*, deux formes d'équivalence sont régulièrement utilisées : l'équivalence dite structurale (*structural equivalence*) et l'équivalence régulière (*regular equivalence*). Dans le premier cas, deux sommets sont regroupés lorsqu'ils sont en relation avec les mêmes autres sommets ; dans le deuxième cas, deux sommets sont regroupés s'ils ont le même type de relation avec le même type de sommets. L'équivalence structurale est particulièrement exigeante : le réseau de 8 sommets à gauche de la figure 6.2 permet la création d'un seul ensemble de deux sommets seulement (les deux sommets blancs émettant un lien vers le même sommet noir).

> **Exemple** : Snyder et Kick proposent dans leur article de 1979 une analyse structurale des relations internationales basée sur les théories de l'échange inégal. Ils partent de quatre matrices binaires et orientées de relations entre États (relations économiques, diplomatiques, commerciales, militaires) et utilisent la méthode CONCOR pour produire un *blockmodel* avec trois positions correspondants au centre, à la périphérie intégrée et à la périphérie. Cette méthode, fondée sur une itération de corrélations de matrices, produit souvent des résultats interprétables même si, comme le notent Wasserman et Faust dans leur

1. La détection de cliques et de *k-cores* est présente dans plusieurs logiciels ; les autres méthodes sont plus rarement implémentées.

FIGURE 6.2 – Équivalence régulière et *image matrix*



	A	B	C
A	0	1	0
B	0	0	1
C	0	0	0

La teinte des sommets à gauche est fonction de l'équivalence régulière et donc les liens avec le même type de sommets. La matrice à droite résume le réseau de départ en montrant les liens entre les trois positions.

---

manuel, les propriétés mathématiques de cette méthode sont mal comprises voire obscures [1].

Ces méthodes ont été développées dans la *SNA* au début des années 1970 pour des réseaux uni-modaux, orientés et non valuées. Différentes adaptations ont été proposées depuis pour des réseaux valués, non orientés, bimodaux, etc. Elles sont semble-t-il moins utilisées aujourd'hui pour partitionner les sommets que les méthodes de détection de communautés présentées dans la section suivante.

6.4 Détecter des communautés

En analyse de réseau, notamment en physique et en informatique, on parle de communautés (*communities)* pour désigner des sous-graphes où la densité de liens entre les sommets de ce sous-graphe est plus importante qu'avec les sommets extérieurs. Le terme *cluster* est parfois utilisé dans le même sens. Là où les méthodes de *blockmodeling* créent des partitions de sommets non nécessairement voisins les uns des autres, les méthodes de détections de communautés créent des sous-graphes fortement connexes.

Une détection de communautés crée une partition exhaustive (tous les sommets sont assignés à une communauté) et généralement exclusive (les sommets sont assignés à une seule communauté). Une détection de communautés se fait en deux temps : choisir l'algorithme (il en existe beaucoup) puis mesurer la qualité de la partition obtenue (là aussi, plusieurs indicateurs existent pour mesurer cette qualité).

Deux petites mises en garde me paraissent utiles. Quand vous présentez vos résultats, s'il vous plaît, n'affirmez pas « dans mon réseau, on a les communautés suivantes » tout en projetant un « joli » réseau plein de couleurs. Ce que vous présentez est la partition obtenue avec un algorithme donné et des paramètres donnés : d'autres algorithmes créeraient des partitions différentes et peut-être tout aussi pertinentes. Le fait que certains logiciels ne proposent qu'une seule méthode de détection de communautés favorise sans doute cette façon maladroite de présenter ses résultats. N'hésitez pas, comme le fait par exemple Paul Gourdon dans sa thèse (2021, pp. 278-315), à tester différents algorithmes pour repérer

1. « the formal properties of the procedure are not well understood [...] the exact mathematical properties of CONCOR remain obscure (it is not clear what, if anything, it is optimizing) » (Wasserman et Faust, 1994, pp. 380-381).

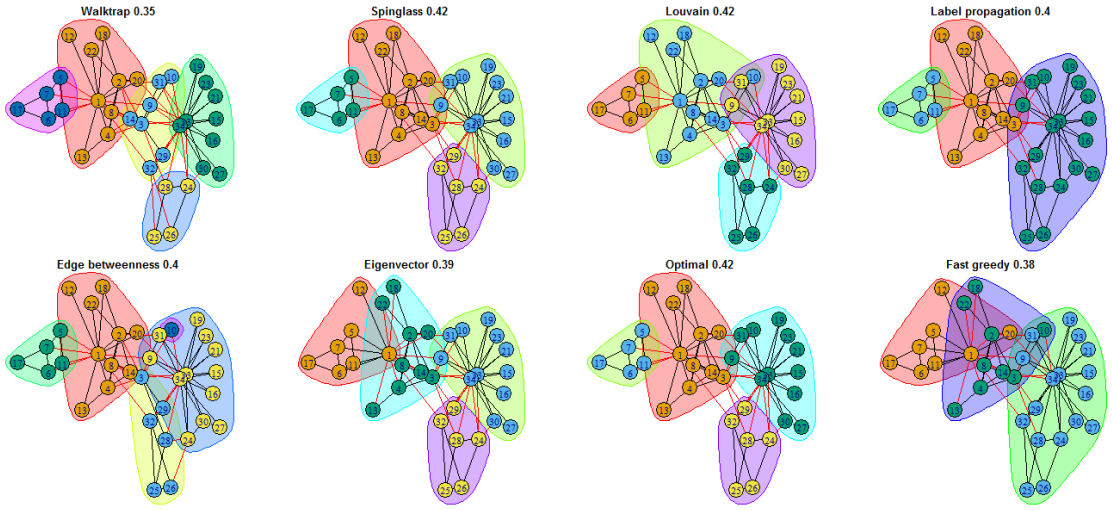
quels sommets sont systématiquement mis ensemble et, à l'inverse, quels sommets ont des appartenances changeantes. Il est tout à fait possible de créer ainsi des gradients d'appartenance aux communautés qui évitent le côté parfois arbitraire des assignations obtenues avec un seul algorithme.

Plus important peut-être : il existe des structures de réseaux où la détection de communautés n'a strictement aucun intérêt. Chercher des communautés dans un arbre est inutile : chercher des communautés dans un réseau de type centre très densément connecté et périphérie très peu connectée est tout aussi inutile ; préférez les *blockmodels* dans ce dernier cas.

Si l'on en croit Rémy Cazabet (2013, p. 16), il y aurait plus de 250 algorithmes de détection de communautés ; sa thèse ayant dix ans, le nombre a dû sensiblement augmenter. En pratique, une poignée seulement d'algorithmes est utilisée : certains sont dits déterministes , ils produiront toujours la même partition ; d'autres non déterministes, la partition créée sur un même réseau peut varier légèrement. Par ailleurs, certaines méthodes sont descendantes (la population de sommets est divisée par itérations successives) ; d'autres sont ascendantes (les sommets sont groupés deux à deux par itérations successives). Dans les deux cas, cela produit un dendogramme qui sera coupé pour obtenir des classes, la coupe s'effectuant à un niveau maximisant une mesure de qualité de la partition. L'une des mesures les plus couramment utilisée est celle de la modularité : plus elle est proche de 1, meilleure est la partition.

La figure 6.3 montre les différentes partitions obtenues sur le jeu de données du *Zachary karate club* [1]. Le nom indiqué est celui de l'algorithme ayant produit la partition, l'indice celui de la modularité.

FIGURE 6.3 – Un réseau, x méthodes de détection de communautés



La détection de communautés est devenue une pratique très commune en analyse de réseau, la recherche sur le sujet est très dynamique et des adaptations sont régulièrement proposées pour tous les types possibles et imaginables de réseaux (valués, multiplexe, bimodaux, dynamiques, etc.). L'implémentation dans des logiciels utilisables en sciences sociales reste rare.

1. Figure tirée d'un billet de 2019 où je détaille les algorithmes disponibles avec *igraph*.

À l'intention des formatrices

Il n'est ni utile ni nécessaire de rentrer dans les subtilités mathématiques des différents algorithmes couramment utilisés en analyse de réseau ; par contre, avoir une toute petite idée du processus et des limites est intéressant (je renvoie au billet évoqué à la page précédente).
Le principal objectif quand on aborde ce sujet est d'expliquer voire de marteler :

- qu'il existe différents méthodes ;
- qu'elles donnent différents résultats ;
- que la qualité des partitions peut s'évaluer de différentes manières.

Pour le formuler autrement : aucun logiciel ne sait « détecter les communautés dans un réseau ». Par contre, plusieurs logiciels proposent d'utiliser la méthode de Louvain en maximisant la modularité - parce que c'est rapide et relativement efficace sur les gros réseaux.
N'hésitez pas à encourager les pratiques comparatives et à prôner les partitions floues qui peuvent être plus intéressantes à commenter d'un point de vue thématique.

Dans les quatre chapitres suivants, vous ne trouverez pas d'encadré *À l'intention des formatrices*. Les réseaux bimodaux, multiplexes et dynamiques sont souvent transformés pour être étudiés comme des réseaux simples, éventuellement valués. Garder la structure originale du réseau étudié suppose me semble-t-il un minimum de familiarité avec les méthodes les plus courantes. L'absence d'encadré dans le chapitre consacré à l'analyse des réseaux personnels est lié à ma faible pratique de ce type d'analyse.

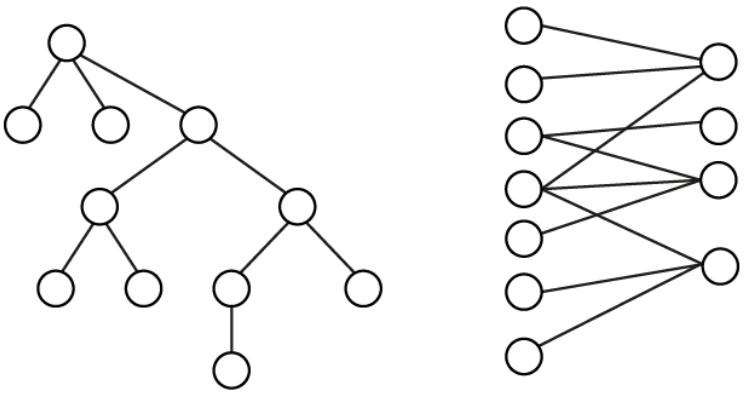
Chapitre 7

Analyser des réseaux bimodaux

7.1 Précisions terminologiques

Un petit point de vocabulaire pour commencer : les termes de réseaux bipartis (*bipartite*) et bimodaux (*2-mode*), ceux de réseaux d'affiliation (*affilation network*) et d'*interlock* sont régulièrement utilisés comme synonymes. Au sens strict du terme, un réseau R est biparti s'il est possible de diviser l'ensemble des sommets V en deux sous-ensembles disjoints V_1 et V_2 tels que tout lien de R a une extrémité dans V_1 et l'autre dans V_2. Tout arbre par exemple est un graphe biparti (figure 7.1).

FIGURE 7.1 – Tout arbre est un graphe biparti



En théorie des graphes, tout arbre (graphe connexe acyclique) et toute forêt (graphe non connexe acyclique) est un graphe biparti : il est en effet toujours possible de créer une partition des sommets en deux sous-ensembles V_1 et V_2 telle que tout lien ait une extrémité dans V_1 et l'autre dans V_2.

Quand on utilise ce terme en analyse de réseau, on suggère le plus souvent que le réseau étudié est bimodal, c'est-à-dire qu'on étudie les relations entre deux ensembles différents de sommets (des actrices (V_1) jouant dans des films (V_2), des autrices V_1 publiant dans des revues V_2, des relations entre enseignantes V_1 et élèves V_2 etc. etc.) ; les liens éventuels entre les sommets de V_1 et entre les sommets de V_2 ne sont pas pris en compte.

Le réseau d'affiliation lie des individus à des organisations. L'*interlock* est un cas particulier de réseau d'affiliation où les organisations en question sont des conseils d'administration. On peut également trouver le terme d'*actor-event network* signalant la participation de personnes à des événements. Ces trois types de réseau sont des formes particulières de réseau bimodal.

Dans la majorité des cas, la direction n'est pas prise en compte et le lien n'est pas valué. Mais on pourrait tout à fait imaginer des liens orientés et valués.

> **Exemple** : soit un réseau bimodal de participantes (V_1) à des conférences (V_2). On pourrait construire un réseau orienté en distinguant les personnes invitées par les organisatrices de la conférence (lien de V_2 vers V_1) et les personnes qui demandant à participer à la conférence (lien de V_1 vers V_2). Le lien pourrait être valué en indiquant le coût (en temps de voyage, en frais d'inscription, etc.) pour les participantes.

7.2 Du réseau bimodal aux réseaux unimodaux

La méthode d'analyse la plus couramment utilisée pour les réseaux bimodaux est de réaliser une projection afin les transformer en deux réseaux unimodaux non orientés et valués[1]. Ces liens sont ensuite généralement dichotomisés (1 si l'intensité de la relation dépasse un seuil donné, 0 sinon) et le réseau projeté est analysé comme un réseau unimodal simple.

> **Exemple** : soit un ensemble V_1 d'autrices et un ensemble V_2 d'articles (figure 3.7 page 29). Le moyen le plus fréquent d'analyser ces données est de créer deux réseaux unimodaux non orientés, l'un pour les autrices, l'autre pour les articles. Un lien entre les autrices indiquent qu'elles ont publié ensemble 1 à V_2 articles ; un lien entre deux articles indique qu'ils ont en commun 2 à V_1 autrices.

Il existe des méthodes plus subtiles pour passer du réseau bimodal au réseau unimodal valué, méthodes issues notamment de la bibliométrie. Si un article a deux autrices, on peut supposer qu'elles se connaissent et qu'elles ont réellement travaillé ensemble. Si, comme c'est le cas dans certaines disciplines, un article a plusieurs dizaines d'autrices, il est probable que l'interconnaissance soit plus faible. On peut donc choisir la valuation du lien de cosignaure qui sera d'autant plus faible que le nombre d'autrices est élevé.

Les méthodes permettant d'analyser les réseaux unimodaux valués ont déjà été abordées précédemment. L'objectif de ce chapitre est de montrer comment analyser les données tout en gardant la structure bimodale du réseau. Je rappelle par ailleurs que ces données pourraient être modélisées sous forme d'hypergraphe (un article = 1 lien, une autrice = 1 sommet) ; les analyses d'hypergraphes restant très rares en sciences sociales, elles ne seront pas abordées.

7.3 Quelques mesures

Les mesures vues dans les chapitres précédents doivent généralement être adaptées pour prendre en compte la nature bimodale du réseau ; les définitions restent les mêmes et ne sont rappelées que pour mémoire.

1. L'opération matricielle est décrite dans l'annexe A.

Ordres, taille & diamètre

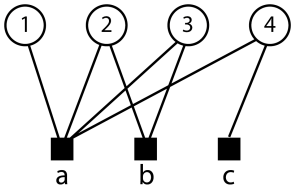
L'ordre est le nombre de sommets d'un réseau, on précisera dans le cas d'un réseau bimodal l'ordre des deux ensembles de sommets V_1 et V_2. La taille est le nombre de liens du réseau : les réseaux bimodaux ne sont par construction jamais vides. Le nombre maximum de liens dans un réseau bimodal est égal à $V_1 \times V_2$.

La définition de la connexité et des composantes ne change pas. Le calcul du diamètre (plus long des plus courts chemins) d'un réseau bimodal connexe ne pose aucun problème particulier ; on peut cependant distinguer un diamètre pour chaque ensemble de sommets.

Centralités

La définition du degré ne change pas et reste le nombre de liens adjacents à un sommet. La seule différence concerne la normalisation de la mesure : le degré maximal d'un sommet de l'ensemble V_1 correspond au nombre de sommets présents dans l'ensemble V_2 et, par construction, il est rare d'avoir des sommets isolés donc le minimum est généralement 1 (figure 7.2). Lorsqu'on commente les degrés d'un réseau biparti et leur distribution, on commente les résultats pour l'un des ensembles de sommets puis ceux obtenus pour l'autre ensemble.

FIGURE 7.2 – Degré et degré normalisé dans un réseau bimodal



	1	2	3	4	degré	d. norm.
a	1	1	1	1	4	1
b	0	1	1	0	2	0.5
c	0	0	0	1	1	0.25

Pour calculer le degré, on fait la somme marginale de la matrice rectangulaire correspondant au réseau étudié. Pour normaliser cette mesure, on divise par le nombre de sommets présents dans l'autre ensemble de sommets. L'opération pour les sommets du haut n'est pas indiquée car ils ont tous le même degré (2) et donc le même degré pondéré (2/3), à l'exception du sommet 1 (degré 1 et degré normalisé 1/3).

Les centralités de vecteur propres sont à ma connaissance rarement mobilisées pour étudier des réseaux bimodaux. Pourtant Bonacich a proposé dès 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.

Le calcul des centralités d'intermédiarité et de proximité ne pose pas de problème particulier d'un point de vue mathématique mais il peut-être plus délicat d'analyser les résultats d'un point de vue thématique.

De la triade au cycle de longueur 4

Il ne peut exister par construction de triade fermée au sein d'un réseau bimodal. La mesure de la transitivité étant devenue très populaire suite à l'article de Watts et Strogatz sur les réseaux petit-monde, il existe des dizaines de propositions permettant d'adapter cette mesure aux réseaux bimodaux [1] et de manière générale à tous les types de réseaux.

1. Voir en première approche les références listées par l'informaticien Tore Opsahl sur son site à la page *Two-mode network clustering*. J'utilise dans la suite de cette section les formules qu'il a proposées ; je ne

L'indicateur global est obtenu en divisant le nombre de nombre de cycles de longueur 4 par le nombre de chemins de longueur 4. Il varie entre 0 (absence de chemin) et 1 (réseau complet). L'indicateur est calculé pour chaque sommet v_i en divisant le nombre de cycles de longueur 4 où se trouve v_i par le nombre de chemins de longueur 4 où se trouve v_i.

> **Exemple** : dans le réseau de la figure 7.2, le seul cycle de longueur 4 présent est le cycle $\{a2, 2b, b3, 3a\}$ (le point de départ n'a aucune importance). Un exemple de chemin de longueur 4 est $\{c4, 4a, a3, 3b\}$.

Cliques et communautés

Il est possible d'adapter la définition de la clique en prenant en compte la nature bimodale du réseau étudié : il s'agit du sous-graphe maximal complet composé d'au moins 4 sommets, 2 sommets minimum étant inclus dans V_1 et deux sommets minimum étant inclus dans V_2. Dans le réseau de la figure 7.2, l'ensemble de sommets $\{ab23\}$ forme la seule bi-clique du réseau.

La détection de communautés étant tout aussi populaire que la transitivité, de très nombreuses méthodes ont été et sont proposées pour prendre en compte la nature bimodale du réseau étudié [1].

N'ayant pas les compétences mathématiques pour hiérarchiser les méthodes proposées d'une part, considérant que ces méthodes ne sont pas implémentées dans les logiciels courants en analyse de réseau d'autre part [2], il ne me parait pas tout à fait utile de poursuivre sur ce sujet.

Boucle et copinage (mais pas que)

Il est souvent intéressant de garder la structure bimodale des données et il est possible de produire des résultats thématiquement intéressants, même avec des mesures basiques. Les deux exemples qui suivent ne prétendent absolument pas être représentatifs ou exemplaires et sont très (trop) clairement ancrés dans ma discipline mais ils me paraissent cependant dignes d'être évoqués.

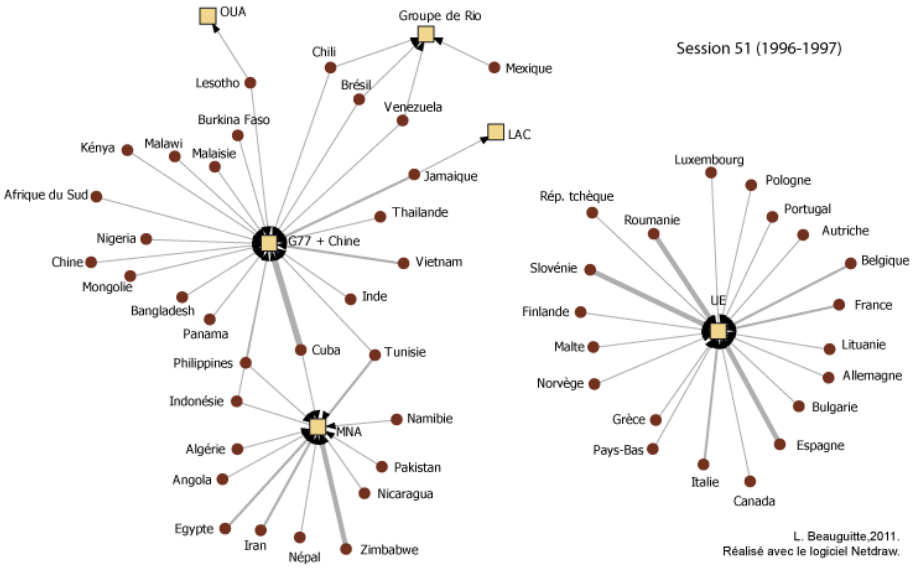
> **Exemple 1** : dans ma thèse (2011), j'ai notamment étudié les déclarations faites à l'Assemblée générale de l'ONU par les groupes régionaux puis j'ai relevé les déclarations faites par des représentant.e.s des États membres qui déclaraient soutenir ou d'associer aux déclarations faites au nom des groupes régionaux. Ce relevé a été fait à quatre dates différentes et sur les réseaux bipartis obtenus, j'ai mesuré l'ordre (V_1 nombre d'États et V_2 nombre de groupes régionaux), la densité, la distance moyenne à l'intérieur de la plus grande composante connexe, le diamètre (idem) et le nombre de composantes connexes.

> **Exemple 2** : dans sa thèse soutenue en 2021, Paul Gourdon s'intéresse à la participation de villes à des associations européennes ainsi qu'à leur participation à des projets de recherche. Le chapitre 4 est consacré à l'étude des réseaux

sais pas si ce sont les plus pertinentes mais elles sont implémentées dans R (*package tnet*) et donc utilisables en sciences sociales.

1. Test réalisé à l'instant sur Google scholar, chercher « community AND "bipartite network" », sans inclure brevets et citations, obtenir 9 140 résultats (requête faite le 28 février 2023).
2. Je ne suis même pas certain qu'il y ait des *packages* R permettant ce type de traitements.

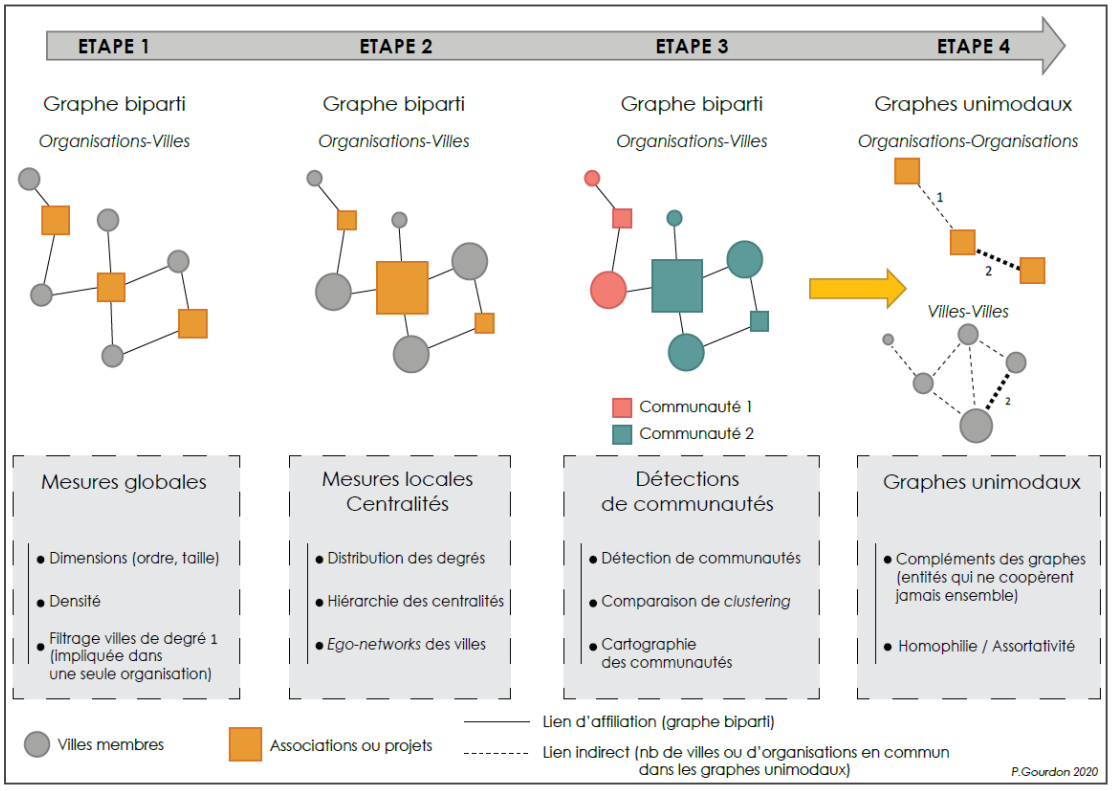
FIGURE 7.3 – Exemple de réseau bimodal : les déclarations des groupes régionaux à l'AG de l'ONU



Les liens sont valués, un État pouvant soutenir à plusieurs reprises des déclarations faites par un groupe régional ; la valuation des liens n'a pas été utilisée pour le calcul des indicateurs. Le simple examen visuel montre un clivage net UE et pays occidentaux vs *le reste du monde et l'absence des États-Unis qui se méfient beaucoup de l'AG de l'ONU.*

bimodaux. Il mobilise différents indicateurs (ordre, taille, densité, distribution des degrés, centralités, réseaux personnels des villes) avant de tester différentes méthodes de détection de communautés. La transformation en réseau unimodal termine l'analyse et cherche à mettre en évidence les liens absents ainsi que l'homophilie des villes et des organisations.

FIGURE 7.4 – Une chaîne de traitements qui prend en compte la structure bimodale du réseau



Extrait de la thèse de Paul Gourdon, 2021, p. 262. La lecture de l'ensemble du chapitre 4 est conseillée ; l'auteur a mis à disposition le code R écrit pour réaliser ses analyses et la thèse est évidemment disponible en ligne.

Chapitre 8

Analyser des réseaux multiplexes

Un réseau multiplexe ou multi-couches (*multilayers*) est un réseau comprenant un ensemble de sommets V et un ensemble de relations différentes entre ces sommets (E_1, E_2, ..., E_n). Certaines relations peuvent être orientées et d'autres non, certaines peuvent être valuées et d'autres non.

8.1 Agréger ou comparer

La méthode d'analyse la plus couramment utilisée est la transformation de ce réseau multiplexe en réseau simple, éventuellement valué. La tactique suppose tout d'abord de transformer les couches une à une pour que chacune soit du même type en ce qui concerne la valuation des liens et l'orientation. L'étape suivante consiste à agréger les couches, en les pondérant ou non en fonction des données étudiées, afin de produire un réseau synthétique. On obtient un réseau valué qu'il est possible d'analyser avec les méthodes vues dans les précédents chapitres.

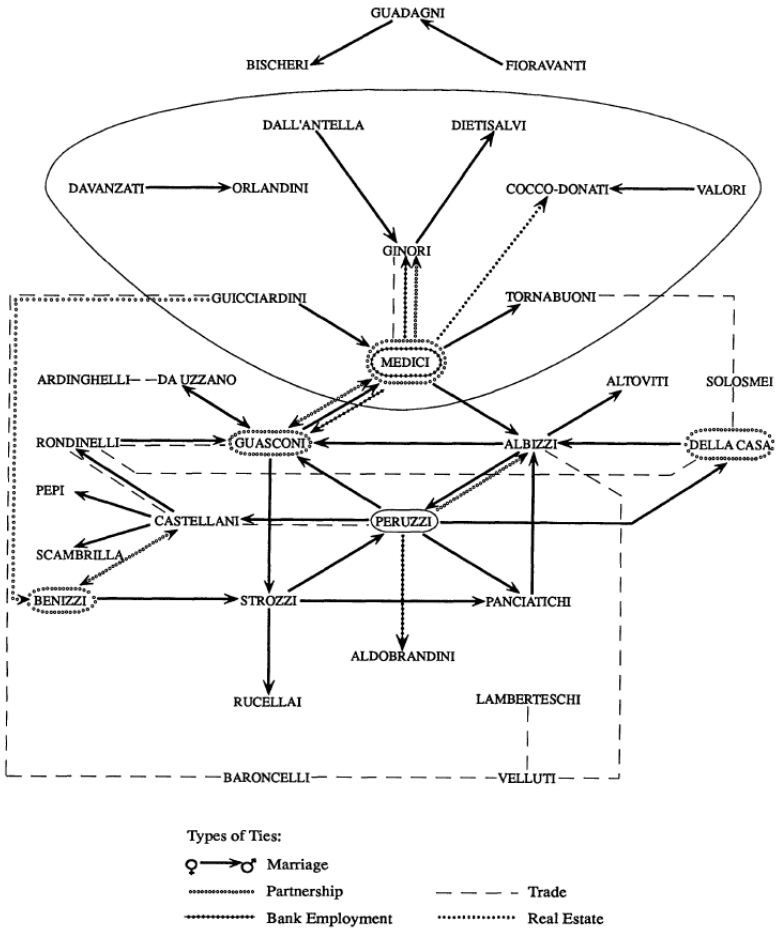
> **Exemple** : dans son article de 1977, Zachary observe les interactions entre des membres d'un club de karaté pendant trois ans. Il étudie notamment les relations en dehors du club et définit huit contextes différents donnant lieu à huit types de relations. Ces différentes couches sont ensuite empilées pour donner une matrice valuée synthétisant ces relations.

Une autre option fréquemment utilisée est d'analyser chaque couche indépendamment les unes des autres puis d'interpréter de manière synthétique les différents résultats.

> **Exemple** : Padgett et Ansell étudient dans leur article de 1993 neuf types de relations (familiales, économiques, politiques, amicales) entre 92 familles florentines au début du XV^{e} siècle. Les liens matrimoniaux et économiques sont considérés comme des liens forts ; les matrices correspondantes sont soumises à des méthodes de *blockmodeling* ; idem pour les liens politiques et amicaux considérés comme des liens faibles. Les scores de centralisation des différentes couches sont comparés sur les réseaux symétrisés et rendus binaires.

Il est bien entendu possible de combiner ces deux approches en analysant chaque couche de manière indépendante puis un réseau de synthèse.

FIGURE 8.1 – Un réseau multiplexe célèbre : les familles florentines du XV^{e} siècle



Cette figure est extraite de l'article de Padgett et Ansell (1993, p. 1276). Comme le précisent les auteurs, les noms indiqués ne sont pas des noms de familles mais ceux des familles dominantes au sein des blocs trouvés par l'analyse et un lien entre blocs signifie qu'au moins deux liens du type indiqué existent entre deux blocs familiaux. Les jeux de données disponibles en ligne ne proposent que cette version agrégée des données construites par les auteurs.

8.2 Conserver la multiplexité des liens

Les méthodes d'analyse des réseaux multiplexes ne sont pas stabilisées et de nombreuses propositions sont faites chaque année. Ce qui suit est donc est sélection très partielle des méthodes possibles.

Les méthodes proposées par Battiston *et al.* dans leur article de 2014 me paraissent à la fois intéressantes et relativement simples à mettre en œuvre. Ils proposent de mesurer notamment les distributions des degrés, le chevauchement des liens, la transitivité des sommets ainsi que les centralités d'intermédiarité et de proximité. Je détaille ici seulement ce qui concerne le degré mais le principe est valable pour les autres mesures de centralité.

Soit un réseau multiplexe de V sommets avec n types de liens entre ces sommets. Une première étape consiste à mesurer les propriétés des sommets (degré par exemple)

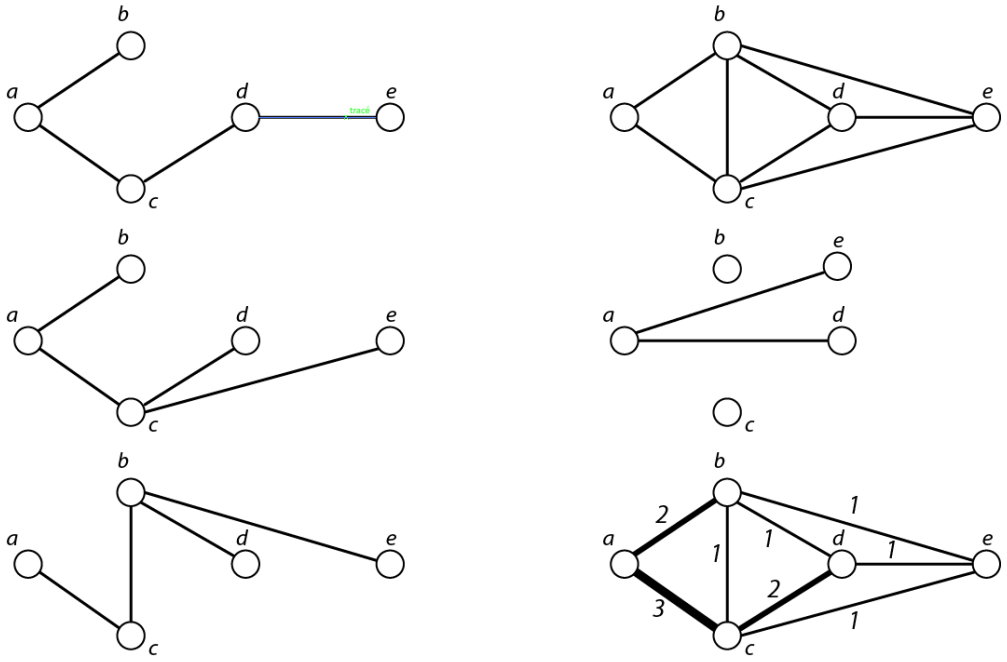
dans chacune des couches, à transformer le résultat en rang : on obtient alors n vecteurs caractérisant chaque sommet et il est possible de faire une matrice de corrélation des rangs pour étudier, en fonction de ce critère, quelles sont les couches les plus similaires et les plus dissemblables.

Cette méthode suppose que l'ensemble des couches est soit orienté soit non orienté. Elle peut évidemment être utilisé pour d'autres indicateurs (transitivité locale, proximité, etc.). Toujours à propos du degré, ils proposent une mesure d'entropie variant de 0 (tous les liens adjacents au sommet v_i se trouvent dans une même couche) et 1 (les liens adjacents au sommet v_i sont répartis de façon uniforme dans les différentes couches).

Il est possible de combiner les n couches pour créer différents réseaux de synthèse pouvant être analysées (figure 8.2) :

- un réseau topologique de synthèse : un lien est créé entre deux sommets si ce lien existe dans au moins une des couches du réseau ;
- le complément du réseau précédent permet de mettre en évidence les liens toujours absents, ce qui peut être intéressant, notamment si les relations sont denses ;
- un réseau agrégé de synthèse où les liens entre sommets où une intensité variant de 0 (lien jamais présent) à n (lien présent dans toutes les n couches). Ce dernier réseau peut être filtré afin de ne conserver par exemple que les liens *toujours* présents.

FIGURE 8.2 – Réseau multiplexe et réseaux synthétiques



À gauche, trois types de relations au sein d'un réseau multiplexe ; à droite de haut en bas, le réseau topologique de synthèse ; le complémentaire du précédent (le sommet e *a été déplacé pour augmenter la lisibilité du réseau) ; le réseau agrégé de synthèse.*

Les distances topologiques peuvent prendre en compte le caractère multiplexe des couches. Il est en effet possible de déterminer les plus courts chemins entre deux sommets non voisins au sein d'une même couche de relations, l'enjeu étant de minimiser tant le nombre de liens que le nombre de couches à parcourir.

Bien d'autres mesures ont été proposées et certaines sont à la fois simples à calculer et à interpréter. C'est le cas par exemple de l'activité d'un sommet, mesurée par le nombre de couches où ce sommet est non isolé ou de l'activité d'une couche, mesurée par le nombre de sommets non isolés dans cette couche. Un taux de multiplexité des liens peut être calculé en faisant le rapport entre le nombre de liens présents dans plus d'une couche et le nombre de liens total. Il est également possible de mettre en évidence des composantes connexes multiplexes, à savoir des ensembles de sommets entre lesquels existe au moins un chemin dans l'ensemble des couches du réseau étudié.

Pour aller plus loin

Ce chapitre est exagérément court, je m'en excuse, et j'indique donc une poignée de ressources permettant d'approfondir cet aspect.

La documentation du *package* R *multinet* donne des pistes de traitement intéressantes [1], notamment des méthodes de détection de communautés prenant en compte la multiplexité des liens. Le livre d'Artime *et al.* intitulé *Multilayer Network Science* (2022), disponible sur lib-gen, présente différentes approches possibles pour l'analyse de ces réseaux. L'acheter n'est pas indispensable : la qualité de l'illustration (souvent empruntée à la référence qui suit) est variable et la formalisation mathématique parfois confuse. L'ouvrage de De Domenico, 2022, *Multilayer Networks : Analysis and Visualization* est en partie le mode d'emploi payant, et hors de prix, d'un *package* R en développement (MuxViz). Ne l'achetez pas, trouvez-le en ligne, il est mieux illustré et plus complet que le précédent.

1. Voir également la présentation de ce *package* écrite avec Paul Gourdon et disponible en ligne.

Chapitre 9

Analyser la dynamique des réseaux

La démarche la plus fréquente consiste à analyser un réseau donné à une date donnée : le nombre de sommets est stable, les relations sont présentes ou absentes mais également stables. Le temps souvent long nécessaire pour recueillir des données relationnelles à l'aide d'entretiens, de questionnaires ou d'observations (ou un mélange des trois) peut expliquer l'aspect « arrêt sur image » de nombres d'articles. Il existe certes des exemples de données sociologiques temporelles (cf les articles de Bidart sur le panel de Caen, l'étude du club de karaté par Zachary ou les relations entre moines étudiées par Sampson [1]) mais si l'on donne sans cesse les mêmes exemples, ce n'est pas par hasard.

La situation a changé suite au développement des outils numériques en ligne : il est désormais théoriquement possible de générer des réseaux dynamiques [2] au sein d'une population donnée. Qu'il s'agisse de Facebook (pour les moins jeunes), de Twitter ou d'Instagram, il est en théorie possible de choisir une méthode (réseau complet ou réseau personnel), une population et d'étudier l'évolution de ce ou ces réseaux (création et suppression de liens et de sommets). D'autres dispositifs (puces RFID notamment) ont également été utilisés pour enregistrer les interactions au sein d'un groupe sur une période donnée.

> **Exemple** : le projet *SocioPatterns* a permis la construction de plusieurs jeux de données relationnelles dynamiques. La collecte s'est faite à l'aide de puces RFID portées par les enquêtées durant un temps plus ou moins long. Toutes les 20 secondes, une trace numérique permet d'identifier les personnes situées face à face à 1 mètre maximum de distance. La vidéo disponible en ligne concernant les interactions au sein d'une école primaire montre le potentiel d'un tel dispositif.

Dans un premier temps, j'évoque des pistes d'analyse possibles quand le nombre de sommets est stable ; dans un deuxième temps, des pistes possibles quand ce nombre est variable.

1. Le jeu de données Sampson est un classique de l'analyse de réseaux sociaux, il montre l'évolution des relations au sein d'un groupe de moines. La thèse intitulée *A novitiate in a period of change : An experimental and case study of social relationships*, soutenue en 1968, a été éditée en 2002 ; elle est introuvable en ligne et présente dans seulement deux bibliothèques en Europe d'après Google (recherche effectuée le 29 juillet 2022) ; ne l'ayant pas lue, je m'abstiens de la citer. **Update** : on peut la trouver sur l'indispensable liste https ://github.com/briatte/awesome-network-analysis de François Briatte. Mais je ne l'ai toujours pas lue...

2. Je considère dans ce chapitre que les expressions »réseau dynamique » et « réseau temporel » sont synonymes. Comme souvent en analyse de réseau, les termes et les significations varient en fonction des autrices.

9.1 Analyser un réseau dynamique d'ordre V

Lorsque l'étude porte sur une même population de sommets dont on étudie les liens à différents temps t (t_0, t_1, ..., t_n), la démarche est relativement proche de celle concernant l'étude des réseaux multiplexes. On peut en effet considérer que chaque temps t correspond à une couche, que le réseau temporel correspond à un réseau multiplexe chronologiquement ordonné et les réseaux de synthèse évoqués précédemment peuvent être créés notamment pour mettre en évidence les liens toujours présents (présence- absence et valuation par le nombre de pas de temps concernés) et toujours absents.

Au moins trois types d'analyse peuvent être menées à bien :
- des analyses à chaque pas de temps ;
- des analyses prenant en compte l'évolution du réseau ;
- des analyses sur le réseau agrégé (topologique ou valué).

Mesurer l'évolution du nombre de composantes, de la densité, du diamètre ou de la distribution des degrés ne pose pas de problème particulier. Pour suivre l'évolution des mesures de centralité, une mesure à chaque temps t, une transformation en rang puis une corrélation des rangs permettent de mettre en évidence la stabilité ou l'instabilité temporelle des hiérarchies au sein du réseau étudié. Il est également possible de proposer des mesures pondérées en fonction de la persistance temporelle. À un niveau plus global, il est possible de corréler les matrices d'adjacence du réseau entre chaque pas de temps t et $t+1$ pour repérer les éventuelles bifurcations [1].

La mesure de la distance topologique couramment utilisée peut être enrichie en prenant en compte une distance temporelle entre sommets : il s'agit de mesurer le nombre de liens et le nombre de pas de temps nécessaires pour joindre deux sommets non voisins au moment t. Un diamètre temporel peut être calculé sur le même principe. Le temps peut également être pris en compte pour mesurer la persistance des liens et des triades fermées.

La détection des communautés dans un réseau dynamique devrait permettre d'identifier les processus à l'œuvre au sein du réseau étudié. La typologie de Cazabet et Rossetti (2019) distingue par exemple les communautés qui apparaissent, grandissent, diminuent, disparaissent, éclatent, fusionnent, se maintiennent et réapparaissent (figure 9.1).

9.2 Analyser un réseau dynamique d'ordre variable

Lorsque l'ordre du réseau dynamique est variable, les méthodes d'analyse sont moins normalisées encore. L'avantage est donc qu'il est possible de laisser libre cours à sa créativité ; l'inconvénient est que l'on n'est pas toujours certain de la pertinence de l'approche que l'on propose.

Une des premières étapes pourrait être de mesurer l'instabilité du réseau en étudiant la fréquence et le rythme d'apparition et de disparition des sommets et des liens.

Si le nombre de sommets et de liens ne cesse de croître, il est possible de vérifier si la création des liens suit une logique d'attachement préférentiel à savoir la probabilité des sommets nouveaux à $t+1$ à créer des liens avec les sommets centraux au moment t.

1. Différentes méthodes permettent de mesurer la similarité entre deux matrices de même taille, l'une des plus simples à mettre en œuvre étant la distance de Jaccard (voir la page anglophone de Wikipedia Jaccard index).

FIGURE 9.1 – Une typologie des communautés dans un réseau dynamique

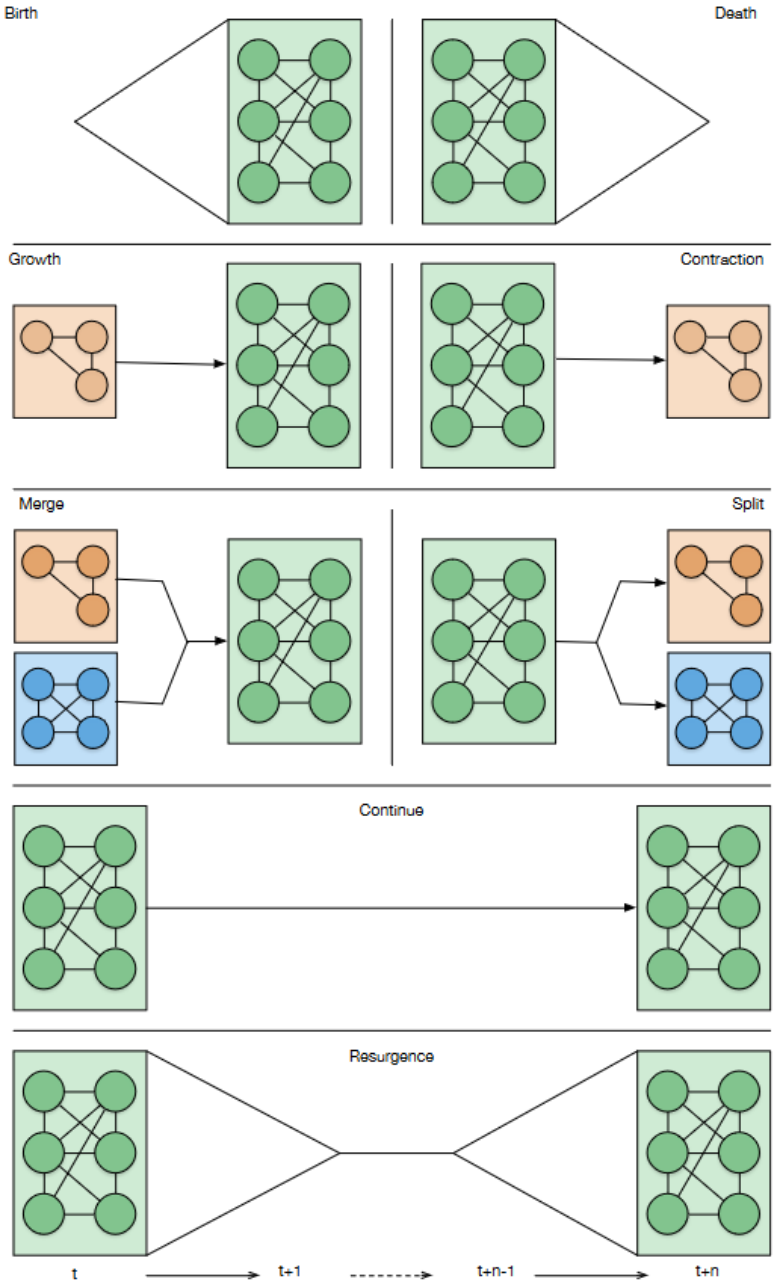


Figure 1. Different types of community events

Figure extraite de l'article de Cazabet et Rossetti (2019). Le pas de temps doit être suffisamment fin pour mettre en évidence ce type de dynamiques : une structure identique à t et à $t+n$ peut en effet masquer ces différents processus.

Exemple : le panel de Caen est une enquête longitudinale en réseau personnel menée par la sociologue Claire Bidart depuis 1995. Les réseaux personnels qu'elle étudie varie à chaque vague d'enquête : certains se contractent, d'autres se diversifient, d'autres encore se renouvellent fortement. L'un des enjeux de

l'analyse est alors de chercher et si possible d'expliquer les déterminants sociaux de ces évolutions très contrastées.

Si le nombre de sommets et de liens peut, selon les cas, varier fortement dans un sens ou dans l'autre, il n'existe pas de modèle statistique simple auquel vous raccrochez, ce qui n'est *a priori* pas un problème majeur si vous êtes chercheuse en sciences sociales. L'un des enjeux de la recherche sera de mesurer ces variations puis de chercher à les expliquer.

Pour aller plus loin

Le plus compliqué avec l'étude de réseau dynamique est d'obtenir des données de qualité. Une fois ce petit obstacle franchi, la liste des possibilités est immense et si vous étudiez ce type de réseau, n'hésitez pas à faire preuve d'imagination pour répondre aux questions que vous vous posez. Il est tout à fait possible de produire des résultats intéressants avec des méthodes simples. L'un de mes exemples favoris est l'ouvrage de BidartBidart *et al.*, *La vie en réseaux* (2011) : les analyses proposées reposent sur une poignée d'indicateurs basiques mais les données étant riches, les questions de recherche pertinentes, les résultats le sont aussi.

La littérature sur le sujet croît chaque année, avec des niveaux de complexité très variable. Deux ouvrages récents sont utiles pour avoir un aperçu des derniers développements : l'ouvrage de Masuda et Lambiotte, *A Guide to Temoral Networks* (2e édition, 2020), très utile mais conseillé aux fans d'équations, et l'ouvrage *Temporal Network Theory* dirigé par Holme et Saramäki, plus hétérogène mais plus accessible que le précédent. Les deux peuvent être trouvés en ligne.

Chapitre 10

Analyser des réseaux personnels

Je ne suis pas un praticien de l'analyse de réseau personnel au sens strict, j'ai parfois extrait des réseaux personnels de réseaux complets pour caractériser les sommets. Ce chapitre s'inspire de manuels récents (Crossley et al. *, 2015; Perry* et al.*, 2018; McCarty* et al. *, 2019). Les aspects liés à la visualisation et aux logiciels sont traités dans les chapitres correspondants.*

Un réseau personnel (*ego-network*, *personal network*) désigne un réseau constitué d'ego, ses voisins d'ordre 1 (*alters*) et les liens entre ces voisins [1]. Une analyse de réseaux personnels suppose le recueil d'une collection de réseaux personnels dans un échantillon de population donnée ou la transformation d'un réseau complet en V réseaux personnels, V étant l'ordre du réseau. Plusieurs stratégies d'analyse complémentaires sont possibles et ce chapitre ne prétend nullement à l'exhaustivité. Je m'intéresse d'abord aux analyses possibles sur un réseau personnel avant de décrire celles comparant différents réseaux personnels.

10.1 Analyser un réseau personnel

Certaines mesures présentent peu d'intérêt pour analyser un réseau personnel. Par construction, le diamètre varie entre 1 (tous les alters sont connectés entre eux) et 2 (cas le plus fréquent) et le réseau est obligatoirement connexe. Ego est dans l'immense majorité des cas le sommet le plus central et ce quel que soit l'indicateur de centralité choisi [2]. La mesure du degré a peu d'intérêt pour ego : dans un réseau personnel de V sommets, il est nécessairement égal à $V-1$. La distribution des degrés des alters peut présenter un intérêt, le degré étant susceptible de varier entre 1 et $V-1$.

Il est par contre intéressant de supprimer ego puis de mesurer le réseau ainsi modifié à l'aide des indicateurs usuels vus au chapitre 5 (nombre d'isolés, de composantes, densité, diamètre, transitivité, etc.). La figure 10.1 montre trois réseaux personnels de même ordre (7 sommets) et de même taille (9 liens). Par définition, ego est connecté à tous ses alters : garder ces liens n'apporte non seulement aucune information utile mais masque des différences structurelles notables. Une fois ego supprimé, ces dernières apparaissent clairement.

1. On pourrait tout à fait imaginer des protocoles de recueil de données visant à récolter les voisins d'ordre 2 ou 3 (situés à des plus courts chemins de longueur 2 ou 3). Un tel recueil serait évidemment très coûteux avec des dispositifs type enquête + entretiens mais le serait peu avec des données nativement numériques.

2. Là encore, ce n'est pas vrai pour un réseau où tous les alters sont connectés entre eux mais il s'agit d'un cas rare.

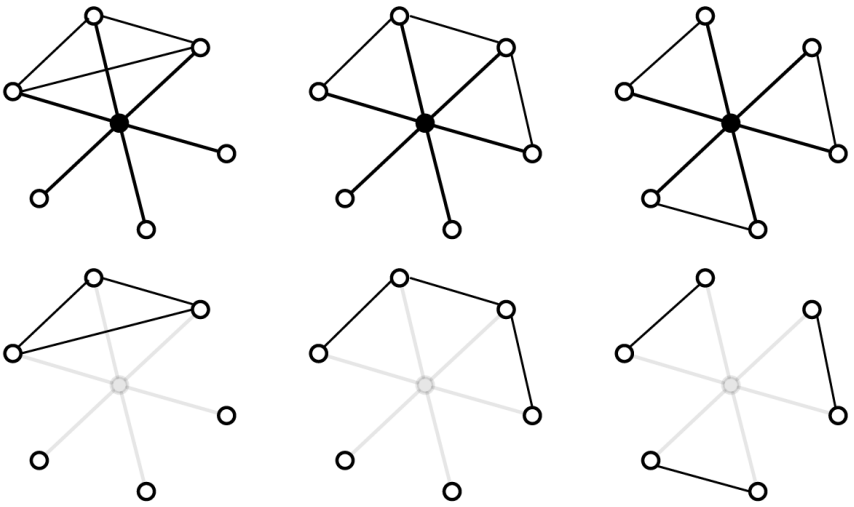
Certaines valeurs restent les mêmes (densité, degré moyen) mais le nombre de triades fermés et d'isolé(s) varient. Les mesures étant différentes d'un égo réseau à l'autre, cela peut donner des pistes pour l'interprétation thématique.

Plusieurs indicateurs ont été proposés spécifiquement pour l'étude des réseaux personnels.Il est possible par exemple d'évaluer la fragmentation du réseau personnel une fois supprimé ego en calculant l'indicateur suivant :

$$\frac{(C-1)}{(N-1)}$$

où C est le nombre de composantes et N le nombre d'alters. L'indice varie entre 0 (les alters forment une composante connexe) et 1 (tous les alters sont isolés). L'indicateur ne permet pas une différenciation fine : un réseau avec un isolé d'un côté et une composante connexe de l'autre obtient le même score qu'un réseau formé de deux composantes de taille égale (cf figure 10.1).

FIGURE 10.1 – Supprimer les egos



Un ego-réseau est nécessairement connexe, son diamètre ne dépasse jamais 2 et ego est très souvent le sommet le plus central, quelle que soit la mesure de centralité choisie. Le calcul d'indicateurs avant et après la suppression d'ego est susceptible de révéler des configurations plus intéressantes. Certains indicateurs ne permettent pas d'identifier des structures différentes : la fragmentation du réseau est la même pour les deux réseaux en bas à droite (1/3).

Burt a proposé plusieurs indicateurs pour mettre en évidence les propriétés des réseaux personnels. Dans une logique utilitariste des liens sociaux, le fait de connaître deux personnes qui se connaissent a peu d'intérêt pour ego car elles risquent de lui apporter la même information. Par ailleurs, plus les alters se connaissent, plus iels sont susceptibles d'exercer une forte pression sociale sur ego.

La taille effective (*effective size)* est égale au degré d'ego $(V-1)$ [1] moins la moyenne des degrés des alters, liens avec ego non pris en compte. Si le réseau est en étoile (aucun

1. Les formules indiquées ici sont celles modifiées par Borgatti pour les adapter à des réseaux personnels non valués. La formule de la taille effective est parfois indiquée comme étant le nombre de sommets moins la moyenne des degrés des alters, liens avec ego non pris en compte. Dans ce cas, le minimum est égal à 1.

lien entre alters), la taille effective est égale au degré d'égo ($V-1$). Si le réseau personnel est complet, la taille effective est égale à 0 : le degré d'ego est par définition $V-1$, le degré de chaque alter aussi. Plus la taille effective s'approche de 0, plus le réseau personnel d'ego est dense.

L'efficacité (*efficiency*) est égale à la taille effective divisée par le nombre d'alters, ce qui permet de normaliser la mesure précédente. Si l'efficacité de mon réseau est égale à 0 (réseau complet), cela signifie que tous les alters apportent la même information ; si elle est égale à 1 (réseau en étoile), chaque alter est susceptible d'apporter une information différente. Burt a également proposé une mesure de redondance égale à $2T/N$ où T est le nombre de liens non adjacents à ego et N le nombre d'alters. Elle varie entre 0 (réseau en étoile) et $N-1$ (réseau complet) ; plus elle est élevée, plus le réseau personnel contient des liens « redondants ».

Burt a proposé d'autres indicateurs qui sont évoqués dans les manuels cités plus hauts. Peu compliqués à calculer, ils supposent pour être interprétés certains types de liens entre certains types de personnes. Pour se limiter à des relations en partie instrumentales, un niveau minimum de redondance est utile dans les relations professionnelles, par exemple pour s'assurer de la fiabilité d'une partenaire potentielle. Si on travaille sur des liens personnels autres (liens familiaux par exemple), d'autres indicateurs sont peut-être plus utiles. Enfin, quand les sommets ne sont pas des personnes, il peut être difficile de supposer une signification stratégique aux liens étudiés.

Si les indicateurs de Burt sont spécifiques à certains types de relations, il explique des indicateurs plus généralistes permettant de mettre en évidence le niveau d'homogénéité ou d'hétérogénéité du réseau et le niveau d'homophilie des liens [1].

Si la variable attributaire considérée est qualitative, plusieurs dizaines d'indicateurs sont disponibles pour la mesurer [2]. Un indicateur fréquemment utilisé est l'indice de diversité d'Agresti (voir l'annexe B pour l'équation). Il varie entre 0 (tous les alters appartiennent à la même catégorie) et 1 (les alters sont distribués de manière homogène entre les différentes catégories). Un indice proche est l'indice de Blau (également appelé indice d'Herfindahl, d'Hirschman ou D de Simpson) qui s'interprète de la même manière.

Si la variable attributaire des alters est ordinale (classes d'âge par exemple), la médiane peut être utilisés. Si elle est continue (âge par exemple), calculer les paramètres statistiques de centralité et de dispersion peut fournir des informations utiles : plus l'écart-type est élevé, plus l'hétérogénéité est forte.

10.2 Comparer les réseaux personnels

Le petit échantillon de mesures listées à l'instant portait sur *un* réseau personnel. Les analyses sont évidemment faites sur les réseaux personnels de l'ensemble des individus enquêtés. Étudier la variation des indicateurs en fonction des individus est une démarche nécessaire ; utiliser ces indicateurs dans le cadre d'une analyse multivariée pour créer des catégories d'individus est une démarche fréquente.

1. Certaines autrices distinguent l'homophilie (liens ego-alters) de l'homogénéité (liens entre alters). Une femme qui n'aurait que des amis hommes aurait un réseau personnel hétérophile et homogène ; une femme qui n'aurait que des amies femmes aurait un réseau personnel homophile et homogène.

2. Non, je n'exagère pas : voir la page https ://en.wikipedia.org/wiki/Qualitative_variation sur Wikipedia.

Exemple 1 : dans l'ouvrage déjà évoqué de Bidart *et al.* (2011), l'utilisation de 7 indicateurs (taille, densité, centralité de proximité, d'intermédiarité, nombre de triades fermées, nombre de composantes et nombre d'isolés) sur l'ensemble des réseaux personnels du panel de Caen et d'une enquête toulousaine permet de construire une typologie en quatre classes : les réseaux denses, centrés (sur ego), dissociés (souvent entre composantes familiale, amicale et professionnelle) et composites (structures complexes combinant des éléments hétérogènes).

Exemple 2 : l'étude des réseaux personnels d'usagers de soins psychiatriques menée par Wyngaerden *et al.* (2020) a permis de collecter des données concernant 390 personnes. Une fois les indicateurs calculés (figure 10.2), une analyse de la variance est menée à l'aide de quatre variables susceptibles d'impacter la structure du réseau (statut résidentiel, niveau d'éducation, sévérité du diagnostic et durée de l'histoire psychiatrique). Ces quatre variables ont ensuite été utilisées comme variables explicatives dans des régressions linéaires visant à expliquer la densité, le nombre de composantes et la centralisation de degré.

FIGURE 10.2 – Variables structurales de réseaux personnels

Table 1. Definition of the main measures used to describe the social networks collected, in terms of size, composition, and cohesion, Morpheus Study, Belgium 2014–2015.

	Definition
Size and composition measures	
Network size	Number of alters (network members) supporting ego (the focal service user)
Diversity of professional services	Number of different professional services (community mental health teams, psychiatric wards, sheltered housing, etc.) to which the alters supporting ego are attached
Diversity of professional functions	Number of different professional functions (psychiatrist, nurse, social worker, etc.) among the alters supporting ego
Diversity of types of alters	Number of types of alters (mental health services, health services, social services, justice, generic services, and non-professionals) among the alters supporting ego
Cohesion measures	
Density	Proportion of effective ties among all possible ties between alters
Fragmentation	Proportion of pairs of alters that are not connected to each other directly or indirectly
Component	Subset of alters that is disconnected from other subsets in ego's network
Clique	Subset of alters that are all connected to each other (maximum density subnetwork)
Isolates	Alter which is not in contact with any other alter in ego's network
Degree centrality	Number of other alters with which an alter is in contact within ego's network
Degree centralization	Sum of the differences in degree centrality between the most central alter and the others, divided by the largest sum of differences that can exist in a network of the same size
Betweenness centrality	Number of times an alter is a crossing point along the shortest path between two other alters
Betweenness centralisation	Sum of the differences in betweenness centrality between the most central alter and the others, divided by the largest sum of differences that can exist in a network of the same size
Average degree	Average number of other alters with which all ego's alters are in contact
Homophily	Number of links between different alters (according to a specific characteristic) minus the number of links between similar alters, divided by the total number of alters (range from −1 to 1)

Extrait de l'article de Wyngaerden et al. *(2020).*

Chapitre 11

Modèles graphiques, modèles statistiques

11.1 Modèles graphiques

Les termes de modèle graphique ou de modélisation graphique sont peu utilisés me semble-t-il en sciences sociales en général et en analyse de réseau en particulier. Mon attachement à cette forme de modélisation, et donc la présence de cette section, est sans aucun doute due à l'influence de mon directeur de thèse, Christian Grataloup, qui dans son ouvrage (hélas épuisé et introuvable en ligne), *Lieux d'histoire. Essai de géohistoire systématique* proposait une série de modèles graphiques tout à fait stimulants.

Cet intérêt rejoint par ailleurs les réflexions de l'analyste de réseaux sociaux Alden Klovdahl qui, dans un article de 1981, écrivait notamment : « Les représentations visuelles fondées sur une théorie correcte sont généralement beaucoup plus faciles à comprendre que les données empiriques qui ont pu les inspirer et permettent donc souvent de transmettre plus clairement les principaux points théoriques. Il convient donc de garder à l'esprit que les mêmes techniques utilisées pour produire des images visuelles de données empiriques peuvent être utilisées pour produire des représentations théoriques » [1].

L'objectif de la modélisation graphique est donc la création d'un schéma illustrant un processus ou un concept susceptible d'expliquer un ou plusieurs des aspects structuraux du réseau étudié (triade interdite ou trou structural par exemple).

Un exercice d'imagination utile quand on réfléchit à ses données et à ses questions de recherche est de se poser la question suivante : que se passerait-il, et donc quelle forme aurait mon réseau, si tel processus expliquait complètement sa structure ?

> **Exemple 1** : je construis un réseau bimodal avec les postes de maîtresses de conférence ouverts au concours sur Galaxie (ensemble de sommets V_1) et les membres des comités de sélection (ensemble de sommets V_2) ; un lien existe quand une personne de V_2 siège dans le comité de sélection du poste V_1. Je peux imaginer par exemple qu'en section 23 (géographie humaine et physique),

1. « Visual representations based on good theory are usually much easier to comprehend than the empirical data that may have inspired them, and hence often help to convey key theoretical points more clearly. Thus, it is worth keeping in mind that the same techniques used to produce visual images of empirical data can be used to produce theoretical representations. »

je vais obtenir deux composantes connexes, l'une créée par les postes de géographie humaine, l'autre par les postes de géographie physique et construire un petit modèle graphique. Je peux ajouter des éléments liés à ma connaissance thématique du sujet pour enrichir ce modèle théorique [1].

Exemple 2 : je m'intéresse à la place des groupes régionaux à l'Assemblée générale de l'ONU (cf *supra*, p. 70). Je peux créer un modèle où l'on passerait d'une situation totalement stato-centrée (seuls les États membres prennent la parole) à une situation centrée sur les seuls groupes puis comparer mes réseaux empiriques à ce modèle.

Le modèle graphique construit sert à imaginer un réseau « idéal » qu'on ne rencontre évidemment jamais. Comparer les données empiriques aux modèles de réseau peut permettre de trouver de nouvelles questions et d'enrichir l'analyse.

FIGURE 11.1 – Un modèle graphique : les collaborations entre villes européennes à l'épreuve du covid

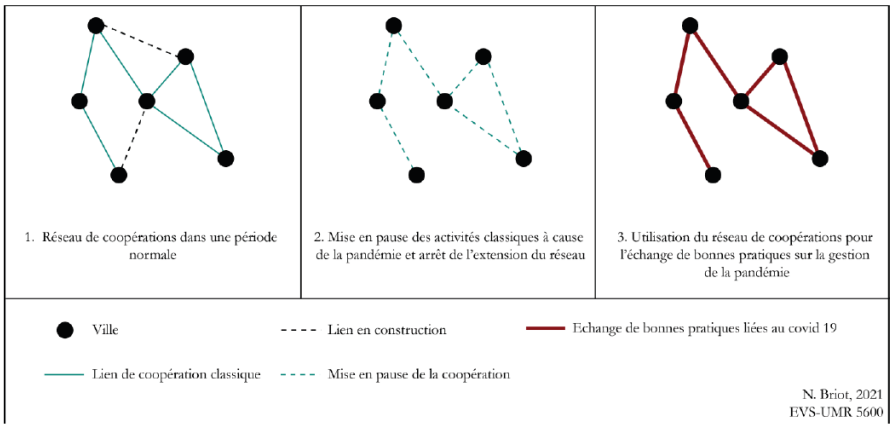


Figure 7.8 : Le réseau de coopérations entre villes pendant la pandémie

Dans sa thèse (2021), sans jamais utiliser le terme, Ninon Briot construit à plusieurs reprises des modèles graphiques permettant d'illustrer un processus qu'elle souhaite étudier. Le fait de produire ces représentations stylisées constitue un bon moyen pour ne pas s'égarer ensuite dans le caractère touffu des données analysées.

1. Je pourrais par exemple supposer, étant données la géographie et la démographie universitaires, que les personnes faisant office de point d'articulation sont plutôt des femmes de rang A travaillant en Île-de-France.

À l'intention des formatrices

À titre personnel, je suis convaincu que l'exercice de la modélisation graphique - auquel on peut donner le nom de schématisation si on ne souhaite pas effrayer les participantes - est utile et pertinent. Construire des hypothèses à partir de ses données et traduire ces hypothèses en images de réseau ne peut que faciliter la construction d'une problématique solide et une analyse de réseau pertinente. En effet, les formes de réseau créées peuvent suggérer les mesures et méthodes pertinentes (si mon modèle est un arbre, inutile de mesurer la transitivité par exemple).
Il est possible cependant à lire les différents manuels disponibles sur le marché que je sois l'un des seuls dans ce cas. Et il n'est sans doute pas indispensable d'aborder le sujet dans une initiation à l'analyse de réseau...

11.2 Modèles statistiques

Les modélisations statistiques de réseau existent depuis plusieurs décennies et semblent occuper une place croissante dans les analyses de réseaux sociaux. Les quelques paragraphes qui suivent présentent très (trop) brièvement un de ces modèles, à savoir le modèle ERGM (*Exponential Random Graph Model*) ; les références bibliographiques devraient permettre aux personnes intéressées par ces méthodes de les approfondir pour les mettre en œuvre.

Le principe général d'un modèle ERGM est de comparer un réseau empirique à x réseaux aléatoires afin d'expliquer (statistiquement) la création des liens au sein du réseau étudié. Les réseaux aléatoires générés ont *a minima* le même nombre de sommets et de liens que le réseau étudié mais d'autres contraintes peuvent être incluses dans le modèle (graphes aléatoires respectant la distribution des degrés du réseau étudié par exemple).

D'un point de vue statistique, le modèle ERGM présente de fortes similitudes avec une régression logistique [1], la différence majeure étant que les données ne sont pas supposées indépendantes. La variable à expliquer est l'existence de liens entre deux sommets et les variables explicatives mêlent généralement propriétés structurales (densité, degré entrant et sortant, liens mutuels etc.) et variables attributaires (figure 11.2). Des tests statistiques permettent d'évaluer la significativité de chacune des variables du modèle.

Comme tout modèle de régression multiple, le choix des variables explicatives est lié à des hypothèses aussi explicites que possibles sachant que plus le nombre de variables est élevé, plus le modèle est délicat à interpréter.

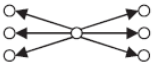
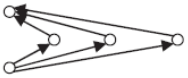
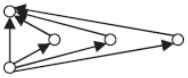
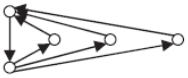
Si les modèles ERGM tendent à s'imposer dans certaines disciplines, ils présentent un certain nombre de limites. La principale limite réside dans le nombre de sommets qui ne peut pas dépasser quelques centaines. Le temps de calcul peut être long et l'interprétation délicate. Comme toutes les régressions multiples, le choix des variables explicatives pèse lourdement sur les résultats : il est tout à fait possible de construire des modèles statistiquement performants (les variables sélectionnées sont toutes très significatives) et peu pertinents au niveau thématique.

Des adaptations des modèles ERGM ont été proposées pour les réseaux valués, bimodaux, temporels, personnels, etc.

D'autres formes de modélisation statistique sont également couramment employés, notamment pour la détection de communautés (*Stochastic block model*). Les compétences

1. Forme de régression linéaire multiple où la variable à expliquer est binaire (0 absence, 1 présence).

FIGURE 11.2 – Variables explicatives possibles d'un modèle ERGM

Network effect		Estimate (SE)
Purely structural effects (endogenous)		
Arc		−1.96 (0.73)*
Reciprocity		2.88 (0.46)*
Popularity (in-degree)		−0.27 (0.32)
Activity (out-degree)		−0.34 (0.34)
Simple 2-path[3]		−0.06 (0.08)
Multiple 2-paths		−0.06 (0.09)
Transitivity (transitive path closure of multiple 2-paths)		1.22 (0.19)*
Cyclic closure (cyclic closure of multiple 2-paths)		−0.37 (0.17)*
Actor relation effects (exogenous) (black nodes indicates actor with attribute)		
Sender (seniority)		−0.56 (0.29)
Sender (projects)		0.01 (0.02)
Receiver (seniority)		0.08 (0.23)
Receiver (projects)		−0.02 (0.02)
[4] Homophily (seniority)		0.64 (0.26)*
Heterophily (projects)		−0.08 (0.02)*
Homophily (office)		−0.01 (0.17)
Covariate network (exogenous)		
Advice entrainment (covariate arc)		1.76 (0.30)*

* = parameter estimate is greater than two times the standard error in absolute value, indicating the effect is significant (see Section 12.5.1 for details).

Extrait du manuel de Lusher et al. *(2013, p. 43). La colonne de gauche liste les variables explicatives endogènes (structurales) et exogènes (attributs des sommets) utilisées, la colonne de droite leurs significativités statistiques.*

mathématiques nécessaires pour les comprendre dépassent tant mes compétences actuelles que le cadre de ce guide pratique.

Pour aller plus loin

La page *workshops* du projet statnet (ensemble de *packages* R dédiés à l'analyse de réseau) offre de nombreuses ressources régulièrement actualisées sur les différents modèles ERGM. Il est possible de les utiliser pour se former même si l'on n'utilise pas R. Le manuel de Lusher *et al.* date un peu (2013), il reste néanmoins très utile pour comprendre ces modèles ; on peut le trouver en ligne sans trop de problème. Les personnes curieuses des *Stochastic block model* peuvent consulter en première approche l'article de Lee et Wilkinson (2019).

Chapitre 12

Visualiser les données relationnelles

La visualisation de données (*dataviz*), qu'elles soient ou non relationnelles, est un champ de recherche à part entière et ce chapitre ne fait qu'effleurer quelques aspects liés à ce sujet [1].

La visualisation de données est indispensable au départ d'une recherche : elle permet d'explorer ses données, de les comprendre, de repérer d'éventuelles anomalies et, avec un petit peu de pratique, de faire émerger de nouvelles questions de recherche. Elle est également utile à l'arrivée, lorsqu'il s'agit de présenter ses résultats.

Si la visualisation sous forme dite « liens - nœuds » est sans aucun doute la plus courante, y compris dans ce petit guide pratique, elle n'est pas la seule forme possible et elle n'est pas non plus nécessairement la plus pertinente.

12.1 Visualiser pour explorer

Lorsque j'explore mes données, je n'ai pas besoin de produire de belles images et de construire une légende impeccable. J'ai aussi le droit de m'affranchir de toutes les règles de la sémiologie graphique. L'exploration de données relationnelles commence presque toujours par une déception : l'image produite est moche, on n'y voit rien, il y a des liens partout et c'est tout à fait normal ! Produire une image lisible prend du temps et suppose généralement une sélection drastique - mais raisonnée - de ce que l'on souhaite montrer [2].

Avant de commencer à filtrer sommets et liens pour tenter d'y voir quelque chose, ayez le réflexe de faire quelques mesures basiques (densité, diamètre, composantes, degré, etc.) et visualisez leur distribution. S'il y a plusieurs composantes, mesurez-les et visualisez-les une à une. Pensez à examiner vos isolés, peut-être partagent-ils des caractéristiques intéressantes. Il y a sans doute dans votre réseau des individus qui vous intéressent plus que d'autres : visualiser leur réseau personnel d'ordre 1, 2 ou plus.

Pensez à varier les algorithmes de visualisation : on ne voit pas la même chose et on ne se pose pas les mêmes questions en fonction de l'image que l'on a sous les yeux. Vous avez trois grandes familles d'algorithmes de visualisation à votre disposition :

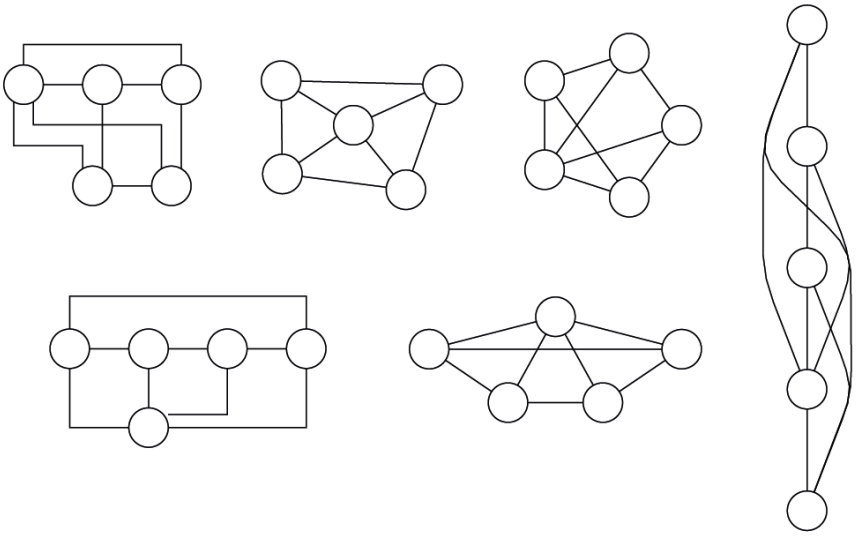
1. Ce chapitre synthétise et actualise des éléments qui étaient déjà présents dans un document produit au sein du groupe fmr avec Françoise Bahoken et Serge Lhomme (2013).
2. Certaines revues acceptent de publier des grosses boules de liens, avec parfois des tâches de couleur, mais c'est leur problème.

- les algorithmes plaçant les sommets en fonction de la géométrie (grille / *grid*, circulaire / *circular*) ;
- les algorithmes plaçant les sommets en fonction de propriétés statistiques de la matrice d'adjacence (MDS) ;
- les algorithmes dits *force-based* qui procèdent par itérations afin de minimiser un certain nombre de critères (chevauchement de sommets, croisement de liens notamment). Les algorithmes Force Atlas (*Gephi*), Kamada-Kawai, Fruchterman-Reingold, nicely (*igraph*) appartiennent à cette catégorie.

Cette dernière famille d'algorithmes est sans doute la plus utilisée car elle est produit des projections lisibles, sauf quand le réseau est très dense. Gardez à l'esprit que la plupart de ces algorithmes sont non déterministes (vous obtiendrez une image légèrement différente chaque fois que vous les relancerez) et qu'ils donnent également des résultats différents d'un logiciel à l'autre. L'algorithme *random* (aléatoire) est disponible dans tous les logiciels. Comme le nom l'indique, les sommets sont placés de façon aléatoire et c'est évidemment le meilleur moyen de produire un réseau illisible.

La figure 12.1 montre le même réseau projeté de différentes façons. Toutes ces images sont strictement équivalentes mais certaines pourraient sans doute donner lieu à des commentaires différents.

FIGURE 12.1 – Un réseau, six algorithmes de visualisation



Minuscule réseau (5 sommets, 8 liens) projetés avec 6 algorithmes différents parmi la vingtaine proposée par le logiciel yEd. Il est tout à fait possible de produire des réseaux esthétiquement agréables et illisibles (celui de droite notamment).

Les stratégies les plus communes pour y voir un peu plus clair sont :
- supprimer les isolés (qu'on a parfois tendance à oublier dans les analyses) ;
- supprimer les sommets de degré inférieur à x, x variant selon la distribution des degrés ;
- en cas de liens valués, supprimer les liens d'intensité inférieure à x (même remarque que précédemment) puis suppression des sommets isolés ;
- si elle existe, ne conserver que la composante géante.

Il est possible en fonction de la structure de votre réseau d'adopter des solutions autres, par exemple de supprimer les liens les plus intenses ou les sommets les plus centraux.

Les conseils précédents s'appliquent à tous les types de réseaux. Un mot sur les réseaux autres vus dans les chapitres précédents.

Il existe un algorithme courant généralement appelé *bipartite* pour les réseaux bimodaux. Il place les sommets V_1 sur une ligne en haut et les sommets V_2 sur une ligne en bas. Cet algorithme n'a aucun intérêt et n'aide pas à explorer ses données. Choisissez deux couleurs différentes (ou deux formes si vous n'êtes pas sensible aux couleurs) pour vos deux populations de sommets et utilisez de préférence des algorithmes *force-based*, vous y verrez plus clair.

Lorsque vous étudiez des réseaux multiplexes, inutile de chercher à visualiser toutes vos relations sur la même image, vous ne verrez rien. Explorez vos relations une à une et explorer les réseaux synthétiques que vous avez produits. Testez d'une part en gardant la même position pour les sommets, ce qui facilite la comparaison, et d'autre part avec des algorithmes *force-based* adaptés à chaque couche, ce qui produira une image lisible pour chacune d'entre elles.

En ce qui concerne la visualisation de réseaux dynamiques, plusieurs logiciels permettent la réalisation d'animation. En phase exploratoire, il est probable que le résultat soit décevant en terme de lisibilité. Le conseil donné à l'instant pour les réseaux multiplexes (1. figer les sommets et étudier les relations à chaque temps t puis 2. utiliser un algorithme *force-based* à chaque temps t) reste valable.

Dans le cas des réseaux personnels, vous gagnerez du temps en supprimant ego avant de visualiser.

À l'intention des formatrices

Animer un atelier sur la visualisation de données suppose que les participantes 1. connaissent un minimum l'analyse de réseau et 2. aient commencé à prendre en main un logiciel.
Pour montrer en quoi consiste l'exploration de données (statistique et visuel), le plus efficace me semble être de fournir un jeu de données aux participantes et de donner une consigne du type « vous devez présenter ces données en 5 minutes lors d'une réunion, vous pouvez présenter des tableaux, des figures, etc. À vous de jouer ».
Choisissez de préférence des données suffisamment généralistes pour pouvoir être manipulées sans être spécialiste de la question. Les jeux de données du *Correlates of War project* (relations commerciales et relations diplomatiques entre États par exemple) m'ont souvent servi.
Le même jeu de données peut bien entendu servir pour le volet communication visuelle (voir partie suivante) mais il est nécessaire dans ce cas que les participantes aient des connaissances thématiques afin de pouvoir tester des hypothèses ou que vous leur prépariez des demandes précises.

Quel que soit le type de réseau analysé, il est souvent plus facile de l'explorer visuellement avec un logiciel clicodrome (avec interface graphique et menus déroulants) : déplacer un sommet ou plusieurs, les sélectionner, les supprimer est plus rapide. Ce qui n'empêche pas en parallèle de l'explorer statistiquement, si besoin avec un logiciel autre.

12.2 Visualiser pour communiquer

Que ce soit un histogramme, un nuage de points, une carte ou un réseau, à partir du moment où vous produisez une image destinée à un public, c'est parce que vous souhaitez faire passer un message clair et que vous souhaitez que ce message soit compris. Pour que ce message soit compris, il est recommandé de respecter quelques règles de sémiologie graphique et de fournir à vos lectrices un titre (problématisé si possible) et une légende (explicite et exhaustive).

La sémiologie graphique étant enseignée dans une poignée de disciplines seulement, rappeler quelques règles s'impose :

- la taille (des sommets) et l'épaisseur (des liens) sont des variables adaptées pour représenter des variables quantitatives (degré, intensité du lien) ;
- la couleur et la forme sont adaptées pour représenter des variables catégorielles (attribut des sommet, type de relation) ;
- la valeur (dégradé de gris, gamme de couleurs ordonnées) est adaptée pour représenter une variable ordinale.

Une remarque supplémentaire sur les variables visuelles : ne surestimez la capacité de lecture du cerveau humain. Nous avons des ordinateurs puissants capables de générer des figures avec des dizaines de nuances de couleur et des milliers d'objets. Certes, mais notre cerveau renâcle à distinguer plus de 7 couleurs[1] et plus de 5 formes différentes dans une figure. Rappelez-vous également qu'au moins 10 % de votre lectorat ne perçoit pas correctement les couleurs : des palettes adaptées existent[2], il est possible aussi de produire des réseaux lisibles en noir et blanc.

Un certain nombre de figures présentes dans les chapitres précédents montre une utilisation adaptée de ces variables visuelles. Dans la figure 2.2 à gauche, la forme des sommets varie en fonction du genre et la forme des liens varie en fonction de leur nature. Idem pour la figure 8.1 où, cependant, la légende ne permet pas de comprendre pourquoi les noms de certaines familles sont entourées et d'autres non ni à quoi correspondent les patates dessinées avec des figurés variables.

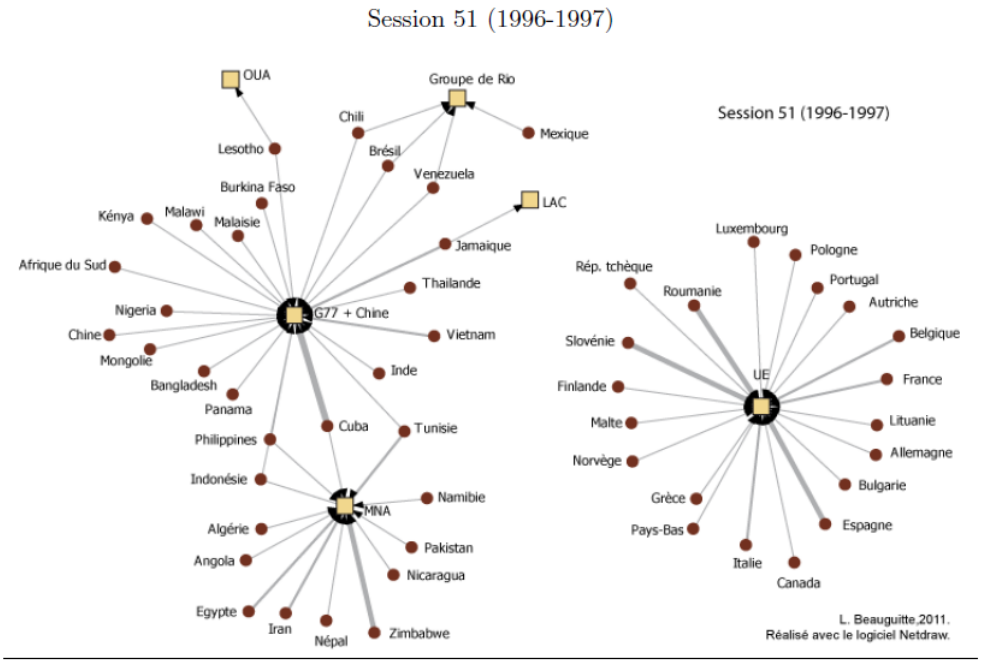
Ce qui m'amène au point suivant : une légende est nécessaire[3] pour que vos lectrices comprennent ce que vous souhaitez montrer. Si dans votre image, il y a des tailles, des couleurs et des formes différentes, il faut que la légende permette de comprendre ce qu'elles signifient. Il est important également d'expliciter ce que signifient les liens entre vos sommets.

Si par exemple vous estimez indispensable de faire de la détection de communautés, indiquez que les couleurs de sommets correspondent aux communautés détectées par l'algorithme x. Vous pouvez même, et personne ne vous en voudra, indiquer la modularité entre parenthèses et préciser quel logiciel vous avez utilisé pour produire votre figure[4].

Petite précision de géographe formé à la cartographie : vous n'avez pas besoin d'indiquer légende au dessus de la légende.

1. Non, rien à voir avec le roman de Brasillach ou avec la maison d'édition fondée par son beau-frère.
2. Une ressource indispensable : https://colorbrewer2.org. Ne vous laissez pas intimider par la carte, ce n'est absolument pas réservé aux géographes.
3. Ce que n'ont toujours pas compris les personnes développant des logiciels d'analyse de réseau. *Cytoscape* a une fonction légende qui donne un résultat très peu satisfaisant ; il est possible de bricoler une légende correcte avec les *packages* de R.
4. Avec un peu d'entrainement, on reconnaît les logiciels utilisés mais parfois on voit des réseaux surprenants et on aimerait savoir avec quoi ils ont été faits.

FIGURE 12.2 – Légender son réseau : peut mieux faire



La place des différentes composantes connexes les unes par rapport aux autres n'a aucune importance. L'épaisseur des liens est proportionnelle au nombre de déclarations de soutien faites. Les graphes ont été réalisés avec le logiciel Netdraw et l'algorithme de visualisation prend en compte le nombre et l'intensité des liens. Reste que les algorithmes les plus courants donnent des résultats peu fiables pour les graphes bipartis. Ces derniers commentaires s'appliquent également aux graphes des sessions 59 et 63.

Soit un extrait de ma thèse (2011, p. 233). J'ai reproduit ici le bas de la figure 3.22 intitulée « États soutenant des déclarations faites par des groupes (sessions 45 et 51) ». Le titre n'est pas bon : trop descriptif. Il n'y a pas de légende : on devine que couleur et forme des sommets différencient États (membres de l'Assemblée générale de l'ONU) et groupes régionaux mais ça reste implicite. Un carton avec les types de sommet et la définition de l'épaisseur des liens serait bienvenu. Indiquer en commentaire (et/ou sur l'image elle-même) le logiciel utilisé et le sens à donner au placement des composantes est, me semble-t-il, une bonne pratique. Par contre, critiquer l'algorithme de visualisation utilisé sans préciser duquel il s'agit n'est pas très sérieux. . .

À l'intention des formatrices

Rien de plus simple (malheureusement) que d'animer une séquence pédagogique sur le sujet : vous exposez les règles de base de sémiologie, vous expliquez ce que doit contenir la légende et vous prenez le réseau de votre choix pour le soumettre à la critique. La deuxième étape consiste évidemment à construire une légende correcte.
Par contre, demander de trouver un titre problématisé est plus compliqué si les participantes n'ont pas de connaissances thématiques sur le réseau utilisé comme exemple. De manière plus générale, encouragez l'esprit critique. Ce n'est pas parce que la spécialiste mondialement reconnue a publié une figure dans la revue de référence que cette figure n'est pas perfectible. C'est en critiquant x dizaines de cartes qu'on apprend peu à peu à en produire des correctes et c'est exactement le même apprentissage avec les réseaux.

Deux derniers conseils pour finir :

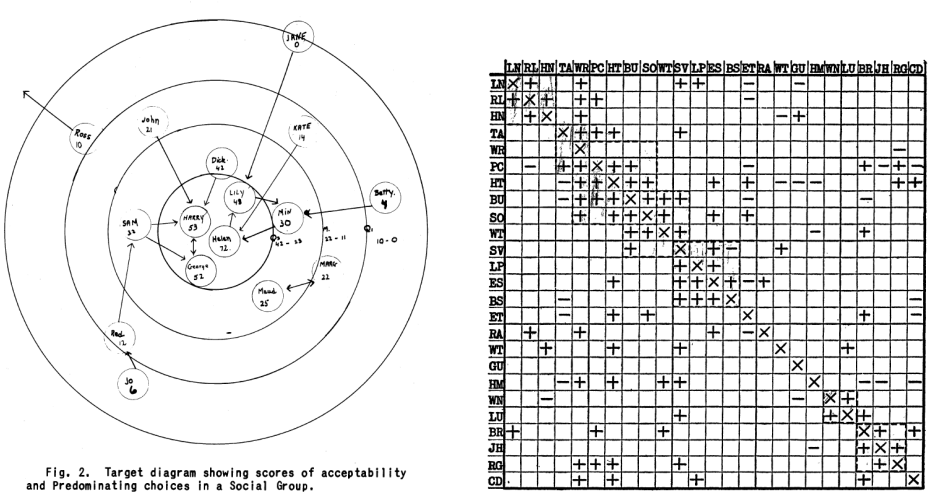
- si vous avez beaucoup d'informations à transmettre, faites plusieurs figures lisibles plutôt que d'essayer de tout mettre dans une seule ;
- adaptez vos figures à votre lectorat. Si vous vous adressez à des spécialistes de l'analyse de réseau, inutile de représenter ego dans un réseau personnel. Si vous vous adressez à des publics autres, garder ego rendra la figure plus compréhensible.

12.3 Au-delà du lien-nœud

Cela fait plus de 70 ans que des chercheuses indiquent à juste titre que, dans certains cas, la forme lien-nœud est peu lisible (réseau dense)[1]. De nombreuses alternatives sont pourtant disponibles.

La matrice ordonnée peut être beaucoup plus lisible que le lien-nœud pour représenter les communautés dans un réseau dense, ce que suggéraient Forsyth et Katz dès 1946 (figure 12.3) et a été confirmé à plusieurs reprises depuis. Plus récemment, des formes hybrides mêlant matrices (pour les parties denses) et lien-nœud ont été proposées et permettent si l'on en croit les travaux de Di Giacomo *et al.* (2021) d'améliorer la lisibilité des figures.

FIGURE 12.3 – Le diagramme cible de Northway (1940) et la matrice ordonnée de Forsyth et Katz (1946)



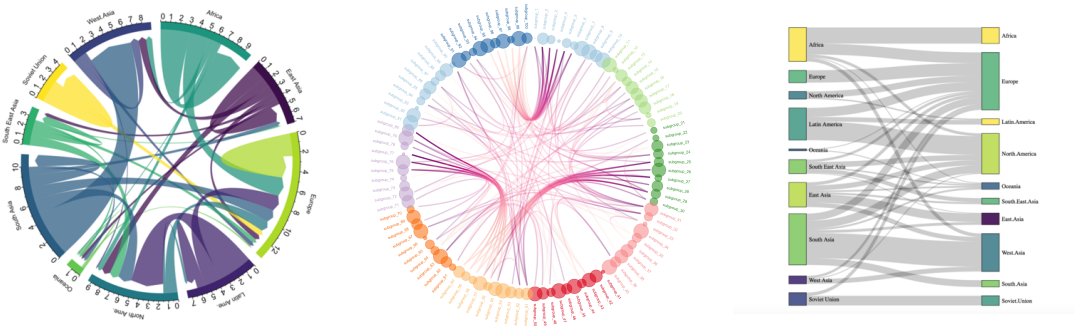
Mary Northway étudie les relations entre élèves. Les cercles correspondent aux quartiles des degrés entrants, ce qui met au centre les élèves les plus populaires. Les sociogrammes sont évidemment dessinés à la main à l'époque et cette proposition vise à justifier statistiquement le placement des sommets. Elaine Forsyth et Leo Katz plaident pour l'utilisation de la matrice ordonnée qui permet de mettre en évidence les sous-groupes fortement connexes autour de la diagonale. Par contre, la matrice est moins efficace pour identifier des chemins entre sommets, ce que confirmeront des études ultérieures.

Les diagrammes en cordes (*chords*) peuvent parfois donner des résultats intéressants si la partition des sommets en ensembles disjoints a un sens thématique fort. L'exemple des migrations internationales où pays émetteurs et récepteurs sont groupés par grandes régions

1. Voir l'article de Forsyth et Katz (1946). Une version bilingue et commentée par Françoise Bahoken, Julie Fen-Chong et moi-même est disponible en ligne.

fait partie de ces cas. La plupart du temps, ce type de visualisation est intéressante si elle autorise l'interactivité (sélection d'un sommet ou d'un groupe de sommets notamment).

FIGURE 12.4 – Diagramme en cordes, *edge-bundling* & diagramme de Sankey



Les trois figures sont tirées du site r-graph-gallery.com.

Parfois combinée avec le diagramme en cordes, le regroupement des liens (*edge-bundling*) consiste à agréger des liens ayant un tracé à peu près similaire. Ce type de visualisation peut donner d'excellents résultats quand il existe une contrainte de localisation sur les sommets (cartographie de flux).

L'utilisation de diagramme de Sankey peut donner des résultats esthétiquement agréables ; je ne suis pas tout à fait certain de la lisibilité des figures et de la clarté du message délivré (figure 12.4). Par contre, je suis de plus en plus convaincu par l'utilisation de l'hypergraphe et de la visualisation dite *BioFabric* pour la visualisation de réseaux bimodaux (Valdivia *et al.*, 2018 ; figure 12.5).

Enfin, si le réseau est de grande taille, peut-être que vouloir le représenter *in extenso* sous forme graphique n'est pas utile. Une distribution des degrés sera sans doute plus lisible et très largement suffisante si vous voulez signalez le caractère hiérarchisé de votre réseau. Plutôt que de colorer une grosse boule dense si vous souhaitez montrer les communautés détectées avec une méthode précise, agrégez les sommets (cf la section *blockmodeling*). N'oubliez pas : votre figure vise à transmettre un message clair.

Pour aller plus loin

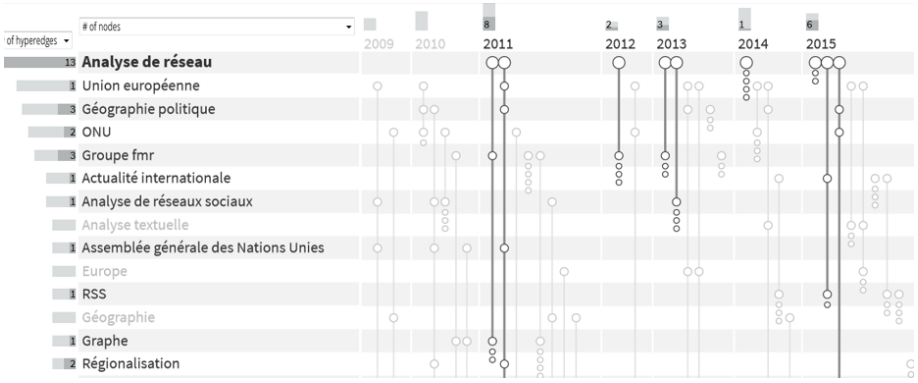
L'un de mes ouvrages préférés de sémiologie graphique est le superbe livre de Tufte (2001). Certes, il assène un certain nombre de règles sans les appuyer sur la moindre analyse empirique mais l'ouvrage est superbe et la démonstration brillante. Il traite de la visualisation de données quantitatives, qu'elles soient ou non relationnelles. Plus récent (2019), l'ouvrage librement accessible en ligne de Wilke, *Fundamentals of Data Vizualization* est également chaudement recommandé : les conseils sont précis et argumentés.

Le travail mené depuis plusieurs décennies par Jean-Daniel Fekete à l'INRIA (équipe Aviz) sur les données relationnelles mérite qu'on s'y attarde, notamment parce qu'il propose des alternatives souvent stimulantes à la visualisation lien-noeud. Mettre au format attendu un mini-jeu de données pour tester les outils disponibles en ligne est susceptible d'enrichir vos visualisations.

J'anticipe sur le chapitre suivant mais si vous voulez avoir une idée à peu près exhaustive de ce qu'on peut faire aujourd'hui en visualisation de données (relationnelles ou

non), n'hésitez pas à visiter la *R Graph Gallery*. Comme le nom l'indique, l'ensemble des représentations peut être réalisé avec le logiciel R.

FIGURE 12.5 – Visualisation sous forme d'hypergraphe



Chaque ligne verticale est un texte disponible sur hal, textes rangés par ordre chronologique ; les termes dans la colonne de gauche sont les mots clés. Les traits mis en évidence concernent des textes liés à l'analyse de réseau. Figure produite à l'aide du logiciel Paohvis.

Chapitre 13

Choisir un ou plusieurs logiciels

L'analyse de réseau suppose 1. du calcul matriciel et 2. la possibilité d'explorer visuellement ses données ; il est donc nécessaire d'apprendre à utiliser un logiciel. Le nombre de logiciels disponibles dépasse la vingtaine et il ne saurait être question de les présenter tous. Je distinguerai deux types de logiciels : ceux qui obligent à coder (R, Python) et ceux qui peuvent s'utiliser avec une interface graphique (Cytoscape, Gephi, Tulip [1]). J'évoquerai enfin quelques outils de visualisation et parfois d'analyse disponibles en ligne.

13.1 Identifier vos besoins

Commençons par un constat un brin pessimiste : le logiciel facile à prendre en main et permettant de réaliser toutes les analyses dont vous avez besoin n'existe pas. L'un des moyens de choisir un outil est de répondre aux questions suivantes :

- quel type de réseau j'analyse ?
- de quelles mesures, de quelles méthodes ai-je besoin pour répondre à mes questions de recherche ?
- est-ce que c'est l'analyse ou la visualisation qui importe dans mes travaux ?
- ai-je besoin de réaliser d'autres traitements de données (statistiques, cartographie, analyse lexicale, etc.) ?

Si le ou les réseaux que vous étudiez sont simples, unimodaux et non valués, à peu près tous les logiciels peuvent être utilisés. Si vos réseaux sont bipartis, multiplexes ou valués et que vous souhaitez les étudier en tant que tels, le choix est plus restreint. Pour être plus explicite, à peu près tous les logiciels savent analyser, au moins sommairement, un réseau simple. Si le réseau est autre, R ou Python risquent rapidement de devenir indispensables.

Les mesures globales et locales basiques (densité, diamètre, degré...) sont disponibles dans tous les logiciels. La plupart des logiciels proposent des méthodes de détection des communautés (souvent l'algorithme de Louvain, parfois le choix est plus large). Réaliser des modèles statistiques type ERGM limite davantage le choix (R, Python ou les logiciels non libres et tournant sur Windows seulement développés à Melbourne par le Melnet).

1. Ces logiciels peuvent également être utilisés avec des scripts. Je n'évoque plus depuis des années les logiciels historiques comme Unicet ou Pajek. Ils ont été très utiles et très utilisés il y a dix ou vingt ans, je ne suis pas tout à fait certain qu'ils le soient encore aujourd'hui, excepté pour les personnes s'étant formées avec et ne souhaitant pas changer d'outils. Aucun des deux n'est libre et les deux ne tournent que sous Windows.

Si vous avez uniquement besoin de produire quelques jolies images de réseaux [1], préférez un logiciel avec interface graphique de type Yed [2], Gephi ou Cytoscape. L'important dans ce cas est que l'outil propose des algorithmes de visualisation efficaces, des exportations au format vectoriel (oui, il vous faudra toujours ajouter une légende) et que le format des données en entrée soit simple (une liste de liens et éventuellement une liste de sommets avec des attributs).

Enfin, un critère important concerne l'importance du traitement de données dans vos travaux. Si vous avez besoin ou envie de faire des cartes, de l'analyse lexicale, de la statistique multivariée et que sais-je encore, formez-vous à R ou à Python [3]. Les tutoriels de qualité pullulent, les formations sont nombreuses, les communautés d'utilisatrices très actives et surtout, une fois passée la phase ardue d'apprentissage, vous gagnerez un temps fou. Une fois vos données mises en forme et votre script écrit, vous n'aurez plus qu'à lancer le programme pour récupérer l'ensemble de vos résultats et de vos figures

À l'intention des formatrices

Lorsqu'on organise des formations d'initiation, on a toutes les chances d'avoir un public très hétérogène. Rappeler des choses très basiques (qu'est-ce qu'un chemin ? comment et pourquoi il est nécessaire de décompresser une archive ?) au départ permet de gagner du temps par la suite.
Le choix du logiciel à utiliser est largement fonction de la durée de la formation proposée. Former à l'analyse de réseau *et* à l'utilisation de R en deux jours ou moins est un pari que je n'ai jamais osé tenter. Si la connexion wifi est de bonne qualité, choisir une application en ligne évite les problèmes d'installation.
Si vous envisagez l'utilisation d'un logiciel, prenez le temps en amont de la formation d'envoyer des consignes très complètes concernant l'installation du logiciel choisi et essayez d'anticiper les problèmes (est-ce que Java doit être installé par exemple). Passer une demi-heure ou plus à tenter de comprendre pourquoi tel programme ne tourne pas sur Mac (exemple choisi de manière non arbitraire) ajoute un stress inutile en début de formation, pour vous comme pour les participantes.

13.2 Mobiliser l'entourage

Le moyen le plus rapide de se former à un logiciel est de connaître des personnes qui l'utilisent déjà. S'il y a des personnes autour de vous qui utilisent Cytoscape, demandez-leur de vous montrer comment ça fonctionne. Si plusieurs personnes utilisent R - y compris pour autre chose que de l'analyse de réseau - et que vous pensez en avoir besoin, lancez-vous. Seul inconvénient : il est possible que votre entourage utilise un logiciel bof [4].

La bonne nouvelle tout de même, c'est que la plupart des logiciels fonctionnent à peu près de la même façon : ils demandent une liste de liens et une liste de sommets en entrée et les noms des fonctions se ressemblent d'un outil à l'autre. Conséquence logique : une fois que l'on sait se servir à peu près d'un logiciel d'analyse de réseau, la prise en main

1. Ce n'est pas du tout péjoratif de ma part : on peut parfaitement souhaiter illustrer des travaux avec une (jolie) image de réseau et ne pas du tout avoir besoin ou envie de faire de l'analyse de réseau.
2. Je ne parle pas de Yed par la suite car il permet uniquement de faire des images de réseau : l'outil est gratuit, multiplateforme et très simple d'utilisation (https ://www.yworks.com/products/yed).
3. Si on ne vous oblige pas à travailler avec, oubliez les équivalents payants : ils ont perdu la bataille dans le monde de l'enseignement et de la recherche et la perdent aussi dans le secteur privé.
4. Le logiciel bof est un logiciel qui propose peu de choix (mesures, méthodes) et/ou impose beaucoup de contraintes (format de données) pour des résultats qu'on obtient à moindre coût avec un logiciel équivalent.

d'un autre logiciel est beaucoup plus rapide. Seule exception à ce cercle vertueux, certains *packages* R [1] sont si mal fichus qu'une pratique assidue du logiciel n'est pas suffisante pour réussir à les utiliser [2]. Vous les repérerez facilement : aucun tutoriel ne les utilise ni ne les cite.

13.3 Un outil à votre service

Petit rappel : le logiciel est un outil qui est là pour faire ce dont vous avez besoin. Ce n'est pas parce qu'un logiciel propose de la détection de communautés que vous êtes obligée d'en faire si ce n'est pas utile pour vos questions de recherche. Tester l'ensemble des indicateurs proposés « des fois que ça donne un résultat intéressant » sans trop comprendre les nuances entre les différentes mesures est un bon moyen de perdre du temps. Et surtout, surtout, si le logiciel que vous utilisez ne propose pas ce dont vous avez besoin, utilisez-en un autre.

Un logiciel digne de ce nom en 2023 devrait vous permettre de créer vos propres mesures et d'adapter les mesures proposées (chapitre 5). Il doit être capable de tourner sur les différents systèmes d'exploitation, ne serait-ce que pour faciliter les collaborations. Enfin, il est préférable que le logiciel soit libre : les erreurs éventuelles passent inaperçues moins longtemps quand le code source est accessible d'une part ; il est beaucoup plus facile de développer des modules permettant de nouvelles approches d'autre part.

13.4 Tester les outils en ligne

Les outils gratuits en libre permettant de visualiser des données relationnelles sont très nombreux et j'en ignore la plupart. J'en présente ici une poignée qui peuvent permettre une analyse rudimentaire de vos données (une poignée de mesures basiques, rarement plus) mais surtout une exploration visuelle facile de vos données (possibilité de zoomer et dézoomer, de filtrer les liens, de sélectionner les sommets, etc.). Ces outils peuvent être intéressants mais le rapport nombre de clics - pertinence des résultats ou des nouveaux questionnements obtenus peut s'avérer décevant à l'usage.

The Vistorian est un outil en ligne permettant de représenter vos données relationnelles sous forme de réseau, sous forme de matrice et, si vos sommets ont des coordonnées, sous forme de carte. L'équipe Aviz, partenaire du projet, a également développé d'autres outils de visualisation de données relationnelles. La documentation (en anglais) est généralement claire, les images produites sont rarement exportables au format vectoriel et ce sont avant tout des outils d'exploration visuelle qui ne proposent pas ou peu de mesures.

VOSViewer est un outil libre et gratuit, il peut être installé en local ou utilisé en ligne. Cet outil s'affiche comme « a software tool for constructing and visualizing bibliometric networks » : s'il peut afficher des réseaux bibliométriques, il peut afficher n'importe quel type de réseau valué. Il est régulièrement actualisé depuis sa création en 2007. On peut trouver les couleurs laides et l'utilisation de la 3D peu efficace. La méthode de clustering utilisée semble spécifique à l'outil et ce dernier ne semble pas proposer de mesures [3].

1. R, comme Python, est un logiciel modulaire : en fonction de ses besoins, on fait appel à des *packages* qui regroupent un ensemble de fonctions adaptées à des besoins spécifiques.
2. Je n'utilise pas assez Python pour savoir si on rencontre le même problème.
3. J'ai survolé le manuel disponible en ligne, j'ai pu rater un passage.

Outil de visualisation de flux géographiques, Arabesque permet de choisir ses variables visuelles (taille, forme et couleur des liens et des sommets), la projection, de filtrer les données, etc. Il fournit également un histogramme des flux étudiés ainsi que des données utiles concernant la sélection de flux représentés.

Autre type de flux, les collaborations scientifiques à l'échelle mondiale[1] : pour avoir une idée de ce qui est possible (articles parus lors de la période 1999-2008), voir le site GéoScimo (représentation en matrices, cordes, cartes et graphes) ; l'outil Netscity permet de visualiser vos données bibliographiques (Maisonobe *et al.*, 2019).

egoSlider est un outil dédié à la visualisation de réseaux personnels et temporels assez intriguant (visualisation de qualité et originale) mais 1. la documentation est indigente (1 article de 10 pages) et 2. c'est juste une démo en ligne, impossible de charger ses données, dommage.

Il existe enfin x outils en outil nécessitant de créer un compte comme Nodegoat ou histograph. Certains sont peut-être de bonne qualité.

13.5 Logiciels à interface graphique

Je distingue d'un côté les logiciels que j'appelle faute de mieux généralistes, utiles pour l'analyse de réseaux unimodaux, et les logiciels « spécialisés » adaptés à un type précis de réseau. Certains de ces logiciels peuvent être utilisés aussi en ligne de commande et/ou avec R et Python. Par défaut, les logiciels indiqués sont libres, gratuits et multiplateformes (Windows, Mac, Linux).

Logiciels « généralistes »

L'un des plus utilisés est sans doute Gephi : plutôt simple d'accès, il permet de réaliser des visualisations efficaces et de mener toutes les analyses standards. Grosse communauté d'utilisatrices et documentation abondante. Un peu léger à mon goût en ce qui concerne la détection de communautés mais il existe des modules permettant d'élargir la palette de choix disponibles.

Très similaire tant en terme d'interface que de rendu visuel, Cytoscape me paraît à la fois plus stable et plus complet. La communauté étant plus réduite, la documentation disponible l'est aussi. Un de mes logiciels préférés pour l'initiation et la visualisation malgré tout.

Tulip est un logiciel de visualisation et d'analyse adapté aux grands réseaux et plutôt complet. Il est très fortement conseillé d'avoir un ordinateur performant et une carte graphique tout aussi performante.

Les trois logiciels qui viennent d'être cités peuvent être utilisés en ligne de commande et sont mis à jour très régulièrement. Des valeurs sûres donc.

SocNetV (Social Network Vizualizer) est un logiciel plus récent, assez simple à prendre en main, avec un choix correct d'indicateurs. Il est par contre étonnamment lent, y compris avec de petits réseaux, que ce soit pour le calcul de certains indicateurs ou les algorithmes

1. Une co-publication issue de chercheuses de deux institutions situées dans des métropoles différentes crée un lien entre ces deux métropoles. Pour comprendre l'intérêt de la méthode et avoir un bel aperçu du potentiel de l'analyse de réseau, voir la thèse de Marion Maisonobe sur le sujet (2015).

de visualisation *force-based*. Je ne suis pas certain qu'il y ait un public pour quelque chose d'aussi générique et peu performant.

Logiciels « spécialisés »

Comme je l'ai déjà écrit, je ne travaille pas sur des réseaux personnels et je connais mal le paysage logiciel. Un certain nombre d'outils commencent à dater : Egonet n'a pas été mis à jour depuis 2017 ; la dernière version E-Net date de 2012 (logiciel gratuit, non libre et seulement pour Windows). Ceci étant, n'importe quel logiciel généraliste peut les visualiser et les analyser ; les indicateurs spécifiques liés à l'homophilie et à l'homogénéité peuvent se calculer avec un tableur ou un logiciel de statistique.

Pour la visualisation de réseau dynamique, DyNetVis semble un candidat intéressant (dernière version datée de septembre 2021) mais ça manque terriblement de documentation ; j'ai du mal quand la seule documentation fournie est une vidéo youtube de 2'42...

Ceci étant, si vous souhaitez travailler sur des réseaux multiplexes, bimodaux, valués, temporels etc., peut-être que le plus efficace est de vous former à R ou à Python.

13.6 igraph, R & Python

Le coût d'entrée est plus élevé, les options sont beaucoup plus nombreuses. Si on utilise ces logiciels régulièrement, on finit par gagner du temps. Dans le cas contraire, on passe des heures à se battre avec des messages d'erreur pénibles... Et, contrairement au vélo, ça s'oublie très vite.

igraph

igraph est un logiciel d'analyse de réseau pouvant s'utiliser soit de façon autonome, soit avec R, soit avec Python. Il permet de faire presque tout avec des réseaux unimodaux (binaires ou valués), presque car il manque les modèles ERGM. Si vous étudiez des réseaux bimodaux et que vous souhaitez les étudier sans les transformer, il y a par contre peu d'options disponibles. La bibliothèque d'algorithmes de détection de communautés est tout à fait correcte (8 choix possibles) et il est également possible de faire des visualisations de qualité. Logiciel stable et actualisé fréquemment, documentation complète, recommandé donc ; ces commentaires sont valables que vous l'utilisiez seul, avec R ou Python.

R

Le couteau suisse du traitement de données avec ses avantages - quel que soit le type de traitement que vous envisagez et le type de réseau que vous étudiez, vous trouverez un ou plusieurs *packages* adaptés. Et ses inconvénients : documentation parfois médiocre, *packages* devenant obsolètes, rétro-compatibilité souvent catastrophique des scripts (*i.e.* le fait de pouvoir utiliser dans 1, 2 ou 5 ans un script qui fonctionne aujourd'hui).

Pour faire simple, avec igraph, vous faites à peu près ce que vous voulez sur de l'unimodal (sauf ERGM et blokmodeling). Si vous voulez des modèles ERGM, vous passez à statnet. Pour le blockmodeling, plusieur packages sont disponibles (blockmodeling, blockmodels, etc.). Et pour les autres types de réseau, vous fouillez et vous trouvez par exemple les packages bipartite (réseaux bimodaux), egor (réseaux personnels) etc. etc.

Pour savoir si ça vaut le coup de lire la documentation du *package*, regardez la version, la date et qui développe. Si le *package* est 0. quelque chose, que la date a plus de deux ans et qu'une seule personne développe, laissez tomber. Et si vous avez encore un doute, cherchez de la documentation en ligne : si vous ne trouvez rien, c'est peut-être parce que personne ne l'utilise et il y a sans doute de bonnes raisons pour ça.

Python

Je connais mal Python donc je vais être beaucoup plus bref. Les principaux modules d'analyse de réseau semblent être igraph (cf ci-dessus), NetworkX et graph-tools. Je ne désespère pas de prendre le temps de me former dans les mois qui viennent pour développer un peu cette partie...

Le foisonnement de modules semble un peu mieux contrôlé qu'avec R et vous pouvez trouver des modules spécialisés type Reticula pour l'analyse des réseaux temporels.

À l'intention des formatrices

Le choix du logiciel que vous utilisez lors des formations dépend 1. de vos capacités 2. de la durée de la formation et 3. du public - il s'agit dans mon esprit d'un ordre décroissant d'importance.
C'est vous qui animez, il est indispensable que vous vous sentiez à l'aise avec l'outil. Un bon moyen pour se former à de nouveaux outils et développer ses capacités est d'écrire des tutoriels. Même courts, n'hésitez pas, écrivez et partagez vos ressources, ça servira.
Pour une formation de deux jours ou moins, je recommande les logiciels à interface graphique, sans hésiter. Si vous avez plus de temps, je recommande R et ce quel que soit le public.
Pourquoi R ? Parce que RStudio le rend moins compliqué à utiliser, parce qu'on trouve plein de documentation de qualité et qu'on peut à peu près tout faire en analyse de données en général et en analyse de réseau en particulier.
« Oui mais c'est du code et en SHS on code pas ».
Je sais. Mais si votre public est peu sensible aux joies de l'informatique, *tout nouveau logiciel est compliqué**. Quitte à faire compliqué, autant former à un outil digne de ce nom. Surtout qu'avec un peu de pédagogie et d'habitude, R s'enseigne sans problème à des personnes pas quanti pour un sou.

*Lorsque j'ai repris mes études, je savais utiliser traitement de texte et tableur, rien d'autre ; apprendre Philcarto (logiciel de cartographie à interface graphique) a été un enfer, ce sont les seuls TD où je pouvais me retrouver au bord des larmes tellement je n'y arrivais pas...

Chapitre 14

Se (re)mettre à jour

Cela a été signalé, l'analyse de réseau est un domaine actif dans de nombreuses disciplines. La bonne nouvelle, c'est que de nouvelles méthodes, de nouveaux indicateurs et de nouveaux outils apparaissent régulièrement. La mauvaise nouvelle, c'est que se maintenir à niveau (notamment quand on donne des cours et qu'on anime des formations) est chronophage. Les pages qui suivent présentent une liste raisonnée - et nécessairement incomplète - de sites et de revues qu'il est utile de consulter régulièrement.

14.1 Logiciels

Les logiciels, et notamment les logiciels libres modulaires comme R, évoluent régulièrement et il est souvent nécessaire de se remettre à niveau pour mener à bien ses analyses et ses visualisations. Faire de temps en temps une recherche des termes *network analysis*, *complex networks* ou *bipartite* (à affiner en fonction de vos travaux) peut s'avérer utile sur la page du CRAN consacrée aux *packages* R [1]. Tester tous les *packages* est chronophage et peu utile : tous les *packages* ne se valent pas (documentation plus ou moins lisible, *package* non tenu à jour, format de données exotique nécessitant x manipulations, etc.). Un moyen plus rapide est de s'intéresser aux portails présentant des ressources pédagogiques comme le portail R-zine ou l'agrégateur de sites R-bloggers.

Je connais moins Python mais si vous l'utilisez pour mener à bien des analyses de réseau, j'imagine que vous savez où aller chercher.

Les logiciels Cytoscape et Gephi ont tous deux un wiki régulièrement actualisés ; le site de Tulip permet d'accéder à des manuels complets (en anglais). Mais on trouve aisément en ligne des ressources pédagogiques en français pour ces trois logiciels.

Enfin, et sans doute est-ce une absurdité pédagogique que de garder le meilleur pour la fin, n'hésitez pas à consulter, et à contribuer si vous le souhaitez, la très riche liste de ressources maintenue par François Briatte.

1. Une recherche rapide faite fin mai 2022 permet de repérer 27 *packages* contenant *network analysis* dans leur description, 8 contenant *complex networks*, 6 *bipartite graph* et 2 le terme *multigraph*.

14.2 Un aperçu partial du paysage éditorial

Le paysage éditorial lié à l'analyse de réseau s'est considérablement étoffé ces deux dernières décennies. Il reste malheureusement dominé par une poignée d'éditeurs privés qui demandent une fortune soit aux lectrices pour accéder aux articles soit aux autrices pour que l'article soit publié en « *open access* ». Le choix a donc été fait ici de présenter d'abord les revues librement accessibles avant d'évoquer plus rapidement les incontournables accessibles grâce à l'indispensable portail sci-hub[1] créé par Alexandra Elbakyan.

Revues spécialisées

La revue hispanophone en accès libre *REDES* existe depuis 2002 et publie essentiellement en espagnol et plus rarement en portugais. Deux numéros thématiques et pluridisciplinaires paraissent chaque année.

L'INSNA (*International Network for Social Network Analysis*) publie la revue anglophone gratuite *Connections* depuis 1977 et tous les numéros sont accessibles en ligne.

Issue du réseau *Historical Network Research*, le *Journal of Historical Network Research* est une revue pluridisciplinaire en accès libre. Un numéro par an depuis 2017. Le même collectif anime une très utile bibliographie multilingue (anglais, allemand, français, espagnol, etc.) centrée sur l'analyse de réseau en histoire.

La toute jeune revue ARCS - Analyse de réseaux pour les sciences sociales cherche à promouvoir une approche pluridisciplinaire et ouverte de l'analyse de réseau (articles, données et scripts en accès libre). Le succès de cette revue n'est absolument pas assuré à ce jour.

Social Networks reste la revue de référence anglophone de l'analyse de réseaux sociaux. Propriété d'Elsevier, elle demande un tarif délirant pour accéder aux (excellents) articles publiés depuis 1978 (25 dollars l'article). Pour que l'article soit en accès libre, il suffit à l'autrice de débourser 3 580 dollars.

Plus récentes, les revues anglophones *Network Science* (depuis 2013), *Journal of Complex Networks* (idem), *Social Network Analysis and Mining* (2011) ou *Applied Network Science* (2016) (liste non exhaustive) fonctionnent peu ou prou sur le même principe : l'autrice paye pour que son article soit en prétendu accès libre. Mais ces revues publient plein d'articles intéressants.

Revues généralistes utiles

La revue de géographie française *Netcom* existe depuis 1987 et est désormais hébergée sur OpenEdition Journals ; les numéros de la période 1987-2006 sont hébergés sur le portail Persée. La revue étant consacrée aux liens entre territoires et réseaux de communication, l'analyse de réseau est régulièrement présente dans ses pages.

Le *Bulletin de Méthodologie Sociologique* publie régulièrement des articles en anglais ou en français relatifs à l'analyse de réseau. Les numéros parus depuis 2008 sont librement accessibles, les numéros précédents continuent à être détenus par SAGE - 29 livres l'article, sci-hub est toujours votre amie.

1. Les liens changeant régulièrement, je ne l'indique pas ici.

De manière plus générale, dans à peu près toutes les disciplines, des numéros thématiques de revues ou des articles isolés mobilisent les concepts et les méthodes de l'analyse de réseau. La recherche par mots clés (*network analysis*, *complex network*) ou par autrices clairement identifiées à un aspect de l'analyse de réseau (Strogatz pour les réseaux petits mondes par exemple) permet de repérer les articles en question.

14.3 Rencontres scientifiques

La grande messe annuelle de l'analyse de réseaux sociaux est la *Sunbelt* qui se tient une année aux États-Unis, une année dans les restes du monde. Beaucoup de monde, x sessions parallèles, tarifs délirants. Si vous aimez apercevoir les stars du domaine et que votre labo a des sous [1].

Équivalent européen de la *Sunbelt* et issue elle aussi de l'INSNA (*International Network for Social Network Analysis*), l'*EUSN* [2] (*European Conference on Social Networks*) a lieu tous les ans depuis 2014. L'analyse de réseaux sociaux domine mais l'ouverture disciplinaire est réelle. Format très classique avec remise de prix divers (posters, jeunes chercheuses).

Les *rencontres Réseaux et Histoire* se tiennent de manière irrégulière en France depuis 2013. Si l'histoire domine, d'autres disciplines sont régulièrement présentes à cette conférence entièrement gratuite ; les actes de toutes les éditions sont disponibles en ligne. Le vivier d'historiennes pratiquant l'analyse de réseau étant limité, il n'est pas certain qu'un nouveau Res-Hist voit le jour dans un futur proche.

Connected Past est un équivalent de Réseaux et Histoire à l'échelle européenne, ce qui assure donc un vivier plus important de participantes. Débutant par une session spéciale en 2011 dans un colloque d'archéologie, la rencontre est devenue autonome l'année suivante, signe de la popularité croissante de l'analyse de réseau en archéologie et en histoire dans le monde anglophone. Les rencontres organisés par l'*Historical Network Research* déjà cité attirent sensiblement le même public me semble-t-il.

Le réseau thématique 26 de l'Association Française de Sociologie, outre la gestion de l'indispensable liste reseaux-sociaux, organise chaque année une session lors du colloque de l'association susnommée. Comme le nom l'indique, la composante sociologique domine.

Les géographes pratiquant l'analyse de réseau organisent parfois des sessions spéciales au colloque européen de géographie théorique et quantitative qui se tient tous les deux ans (*ECTQG*). Dans les colloques plus généralistes, le terme réseau est le plus souvent à prendre dans un sens métaphorique. Il est probable que le groupe fmr recommence à proposer des animations régulières dans un futur proche (information à vérifier).

Les événements scientifiques francophones sont généralement annoncés sur la très utile liste de diffusion reseaux-sociaux animée par le réseau thématique 26 de l'AFS déjà cité. La liste est peu bavarde, s'y abonner est fortement recommandé. La liste DH (*Digital Humanities*) relaye également nombre d'événements mobilisant l'analyse de réseau.

1. Je suis ici d'une totale mauvaise foi, n'ayant jamais mis les pieds dans cette conférence.
2. Chaque événement crée son site autonome.

Conclusion

J'espère que vous avez appris quelques bricoles utiles. J'espère aussi que le caractère inachevé de certains chapitres ne vous a pas trop déçu. J'espère enfin que ce texte ne contient pas trop d'erreurs ou de coquilles.

Je l'ai écrit dans l'avant-propos et je l'écris à nouveau ici : n'hésitez surtout pas à me faire part de vos retours sur ce petit guide pratique[1]. Que vous soyez débutante ou praticienne, que vous ayez suivi ou animé des formations d'analyse de réseau, vos avis m'intéressent.

1. Les personnes scandalisées que je crache sur les éditeurs privés et que je fasse de la pub pour sci-hub et lib-gen peuvent s'abstenir. Merci. Bisous.

Annexe A

Notations mathématiques et calcul matriciel

Les personnes qui savent lire une équation peuvent sauter la section suivante ; les personnes qui les survolent sans trop chercher à les comprendre[1] *trouveront peut-être un petit intérêt aux lignes qui suivent.*

Déchiffrer une équation

Il existe deux moyens principaux de décrire un indicateur : écrire ce qu'il permet de mesurer ; donner la formule permettant son calcul. Savoir lire une formule est utile pour plusieurs raisons. Certaines disciplines les utilisent couramment et il est difficile voire inutile d'essayer de lire un article de physique si on ne sait pas lire une équation. Lire les équations permet de démystifier certains travaux : un minimum d'habitude permet de réaliser que 90% des « nouveaux » indicateurs proposés dans certaines communautés scientifiques sont des indicateurs classiques plus un paramètre donné et/ou une constante. Avec un peu de pratique, il est également plus facile de prévoir les variations d'un indicateur et ses limites en étudiant sa formule mathématique plutôt que sa description verbale.

Il existe une poignée de symboles utilisés de manière à peu près homogène quelle que soit la discipline. Je liste ici uniquement les plus fréquents :

- $\subset$: l'ensemble situé à gauche est inclus dans l'ensemble situé à droite. Soit un graphe $G = \{V, E\}$ et un sous-graphe $G' = \{V', E'\}$ de G alors $V' \subset V$ et $E' \subset E$;
- $\in$: l'élément à gauche est un élément de l'ensemble situé à droite. Par exemple, dans un graphe $G = \{V, E\}$, tout sommet $v \in V$;
- $\cup$: union des deux ensembles situés de part et d'autre ;
- $\cap$: intersection des deux ensembles situés de part et d'autre ;
- $\bar{A}$: moyenne de A ;
- $|A|$: valeur absolue de A ;
- Σ désigne la somme des éléments situés à droite ;
- Π désigne le produit des éléments situés à droite.

1. Il n'y a pas de honte à avoir : j'ai commencé à traduire certains articles, d'optimisation linéaire notamment, quand j'ai réalisé que je ne les comprenais pas parce que je ne prenais pas le temps de *lire* les équations.

Les deux derniers symboles (sigma et pi) ont souvent une notation du type $i = 1$ en bas et n en haut, comme ceci : $\sum_{i=1}^{n}$. Les lettres utilisées importent peu : cela signifie que la somme concerne les n individus du premier ($i = 1$) au dernier (n) (voir les formules des indicateurs de Shimbel par exemple).

Classiquement, les formules en analyse de réseau sont des égalités où le terme de gauche est le nom de l'indicateur (parfois une lettre grecque, souvent une lettre en italique mais c'est une simple convention et il est rare que deux autrices adoptent les mêmes) et à droite la formule permettant le calcul de l'indicateur. La formule est précédée ou suivie d'une courte phrase définissant les termes employés dans l'équation.

Pour comprendre une formule un peu longue (modularité par exemple), il est conseillé de commencer par décomposer : comprendre ce qui est calculé dans les parenthèses, au numérateur puis au dénominateur.

Soit un des indicateurs les plus fréquemment calculés, la densité.

La densité δ d'un graphe simple non orienté et non planaire d'ordre V (nombre de sommets) et de taille E (nombre de liens) est :

$$\delta = \frac{2E}{V(V-1)}$$

Une même équation peut s'écrire de différentes façons. Les formules qui suivent sont totalement équivalentes :

$$\delta = \frac{2E}{V(V-1)} = 2\frac{E}{V(V-1)} = \frac{E}{(V \times (V-1))/2}$$

Les notations $V(V-1)$, $V \times (V-1)$ et $V \cdot (V-1)$ sont strictement équivalentes et désignent la multiplication du premier terme par le second.

Pour mémoire, la densité est le rapport entre le nombre de liens présents (noté ici E) et le nombre de liens possibles. Il s'agit d'un rapport donc la formule a la forme d'une fraction. Le graphe est simple, non planaire, il n'y a pas de boucle : chaque sommet peut être en relation avec tous les autres sommets. S'il y a V sommets, le nombre de liens possibles est donc $V(V-1)$. Si les boucles étaient autorisées, on remplacerait le dénominateur par V^2. Mais pourquoi multiplier le nombre de liens par 2 (ou diviser le nombre de liens possibles par 2) ? Le graphe est non orienté donc les liens entre les sommets v_i et v_j et v_j et v_i sont considérés comme un seul lien. Si le réseau était orienté, on supprimerait le terme 2 de l'équation.

Avec ces éléments à l'esprit, on peut déterminer les variations de l'indicateur. Si le graphe est vide (pas de lien), le numérateur (E ou $2E$) est égal à 0 et δ est égal à 0. Si le graphe est complet, le nombre de liens présents est égal au nombre de liens possibles et δ est égal à 1.

En résumé

L'objectif n'est pas de se transformer en mathématicienne mais il est utile de prendre le temps de réfléchir aux formules, aux indices, de se familiariser avec les notations les plus courantes. Ça vous permettra notamment de repérer quand un logiciel vous sort des résultats aberrants, ce qui arrive de temps à autre. Ça vous permettra aussi de repérer quand deux indicateurs sont tellement proches l'un de l'autre qu'en mesurer un seul suffit.

Ça vous donnera accès à des disciplines habituées à ces formalisations et actives en analyse de réseau. Et, avec un peu d'habitude, ça vous permettra surtout de mettre aux points les indicateurs dont vous aurez besoin pour répondre à vos questions de recherche.

Attention, certaines autrices sont spécialisées dans l'équation inutilement compliquée allant souvent de pair avec une description incomplète des termes de la formule donc n'hésitez pas à croiser les sources, par exemple en contrôlant la formule indiquée le Wikipedia anglophone [1].

Dernier conseil quand vous explorez un indicateur : dessinez de minuscules réseaux (une poignée de sommets, une poignée de liens) avec des formes spécifiques (ligne, cercle, nœud papillon), réfléchissez aux résultats que vous devriez obtenir puis calculez manuellement l'indicateur en question. Vérifiez ce que donne le logiciel que vous utilisez et si le résultat est normalisé par défaut, prenez le temps de comprendre comment il est normalisé. Vous serez beaucoup plus à votre aise ensuite lorsque vous commenterez et interpréterez vos résultats.

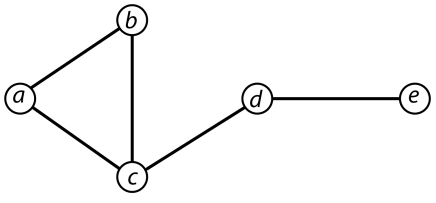
Matrices et calcul matriciel

Un certain nombre d'indicateurs en analyse de réseau reposent sur du calcul matriciel, calcul très vite chronophage s'il est réalisé à la main. Si l'analyse de réseau se développe à partir des années 1950-1960 dans différentes disciplines (chapitre 2), c'est en partie grâce à l'arrivée des premiers ordinateurs dans les universités. Si les réseaux étudiés depuis une vingtaine d'années sont de plus en plus gros, c'est en grande partie lié à la puissance de calcul informatique disponible aujourd'hui.

Ce qu'est une matrice

Une matrice est un tableau rectangulaire de nombres composé de x lignes et y colonnes permettant de représenter les propriétés d'un objet. Les matrices d'adjacence d'un graphe simple d'ordre V sont des tableaux carrés, remplis de 1 (présence d'un lien de v_i vers v_j) et de 0 (diagonale, absence de lien entre v_i et v_k). Si le graphe est non orienté, tout lien de v_i vers v_j suppose un lien de v_j vers v_i et la matrice est dite symétrique : la disposition des 1 et des 0 est identique de chaque côté de la diagonale.

L'exemple ci-dessous montre un réseau simple, non orienté, et la matrice d'adjacence correspondante.



	a	b	c	d	e
a	0	1	1	0	0
b	1	0	1	0	0
c	1	1	0	1	0
d	0	0	1	0	1
e	0	0	0	1	0

Par convention, la lecture se fait des lignes (origine) vers les colonnes (destination). Un calcul simple est la somme marginale (somme en ligne ou somme en colonne) permettant d'obtenir les degrés. Avec la matrice d'adjacence d'un graphe orienté, les sommes en ligne permettent d'obtenir le degré sortant, les sommes en colonne le degré entrant.

1. Ce que j'ai fait pour les formules indiquées *infra*.

Plus courts chemins et diamètre d'un réseau

Le calcul des plus courts chemins dans un graphe repose sur du calcul matriciel. Lorsque la matrice d'adjacence est élevée au carré, les nombres dans les cases correspondent au nombre de chemins de longueur deux entre chaque paire de sommets. Lorsque la matrice est élevée au cube, les nombres correspondent au nombre de chemins de longueur trois, etc. etc.

Si on additionne la matrice de départ et la matrice élevée au carré, les cases avec un 0 correspondent aux paires de sommets entre lesquels n'existe pas de chemin de longueur 1 ou 2. Si on additionne la matrice de départ, la matrice élevée au carré et la matrice élevée au cube, les cases avec un 0 correspondent aux paires de sommets entre lesquels il n'existe pas de chemin de longueur 1, 2 ou 3. Tant qu'il reste des 0 dans cette somme, l'opération continue jusqu'à A^{V-1} (le diamètre maximal d'un graphe de V sommets est $V-1$).

Le tableau ci-dessous montre la matrice d'adjacence A, la matrice au carré et la matrice au cube ainsi que la somme de ces trois matrices.

A

	a	b	c	d	e
a	0	1	1	0	0
b	1	0	1	0	0
c	1	1	0	1	0
d	0	0	1	0	1
e	0	0	0	1	0

A^2

	a	b	c	d	e
a	2	1	1	1	0
b	1	2	1	1	0
c	1	1	3	0	1
d	1	1	0	2	0
e	0	0	1	0	1

A^3

	a	b	c	d	e
a	2	3	4	1	1
b	3	2	4	1	1
c	4	4	2	4	0
d	1	1	4	0	2
e	1	1	0	2	0

$A + A^2 + A^3$

	a	b	c	d	e
a	4	5	6	2	1
b	5	4	6	2	1
c	6	6	5	5	1
d	2	2	5	2	3
e	1	1	1	3	1

La présence de nombres dans la diagonale peut surprendre. Si on examine A^2, la case (a, a) contient un 2 : il existe donc deux chemins de longueur 2 permettant d'aller de a à a ($\{ab,ba\}$ et $\{ac,ca\}$). La lecture est la même pour A^3 : il existe par exemple 3 chemins de longueur 3 entre a et b ($\{ac,ca,ab\}$, $\{ab,ba,ab\}$, $\{ab,bc,ca\}$).

Lorsqu'on additionne les différentes matrices ($A + A^2 + \ldots + A^n$), le diamètre est égal au plus petit exposant n pour lequel plus aucune case n'est égale à 0 (ici 3, ce qui se vérifie facilement sur le graphique). S'il reste des 0 pour $n = V - 1$, cela indique un réseau non connexe.

Matrice et transposée

La transposée d'une matrice d'adjacence A, souvent notée A' ou $t(A)$, s'obtient en permutant lignes et colonnes. Si je pars d'une matrice actrices - événements de taille $V_1 \times V_2$, la transposée sera de taille $V_2 \times V_1$. Multiplier une matrice rectangulaire par sa transposée ou une transposée par la matrice d'origine permet de passer d'un réseau bimodal à deux réseaux unimodaux non orientés et valués.

Je reprends l'exemple utilisé dans le chapitre 7 concernant la participation de 4 chercheuses à 3 colloques. Soit la matrice d'adjacence du réseau de départ A où la case a_i, C_j vaut 1 si a_i a assisté à la conférence C_j et 0 dans le cas contraire. On obtient la matrice suivante :

	C_1	C_2	C_3
a_1	1	0	0
a_2	1	1	1
a_3	0	1	0
a_4	0	1	1

La somme en ligne permet de connaître le nombre de conférences auxquelles ont assisté les personnes ; la somme en colonnes le nombre de personnes présentes à chaque conférence. Pour obtenir les réseaux unimodaux dérivés, on multiplie la matrice A par sa transposée A' obtenue en inversant lignes et colonnes. La multiplication $A \times A'$ permet d'obtenir une matrice individus - individus de co-présence aux évènements ; la multiplication $A' \times A$ une matrice événements - événements partageant un nombre donné de participantes.

A

	C_1	C_2	C_3
a_1	1	0	0
a_2	1	1	1
a_3	0	1	0
a_4	0	1	1

A'

	a_1	a_2	az_3	a_4
C_1	1	1	0	0
C_2	0	1	1	0
C_3	0	1	0	1

$A \times A'$

	a_1	a_2	a_3	a_4
a_1	1	1	0	0
a_2	1	3	1	2
a_3	0	1	1	0
a_4	0	1	1	2

$A' \times A$

	C_1	C_2	C_3
C_1	2	1	1
C_2	1	3	2
C_3	1	2	2

Les matrices obtenues sont valuées, symétriques (*i.e.* réseau non orienté) et la diagonale permet de retrouver le degré présent dans le réseau bimodal de départ (ex. : la case C_2, C_2 de la matrice $A \times A'$ vaut 3 : 3 personnes ont assisté à cette conférence).

Vecteurs et valeurs propres d'une matrice

Dans la section consacrée aux degrés pondérés par le degré des voisins, différents indices ont été évoqués (Katz, Eigenvector, PageRank). Tous ces indices reposent en partie sur le calcul des vecteurs propres - d'où le nom d'*eigenvector centrality* souvent utilisé dans les logiciels pour calculer l'un de ces indices.

Un vecteur propre x d'une matrice carrée est un vecteur solution de l'équation $Ax = \lambda x$ avec $x \neq 0$. Une matrice a en général plusieurs valeurs propres, la plus élevée est utilisée pour calculer les centralités. λ est appelé valeur propre.

L'ensemble des valeurs propres de A est appelé le spectre de A. Certaines méthodes d'analyse de réseau, non abordées dans ce guide, mobilisent ces valeurs propres (*spectral analysis*).

Annexe B

Quelques indicateurs fréquemment utilisés

Les formules sont classées par ordre alphabétique.

Liste des symboles utilisés et définitions

G : graphe $\{V, E\}$.

V : ensemble des sommets ou ordre du graphe, noté V_1, V_2 en cas de réseau bimodal. v_i désigne n'importe quel sommet, v_i, v_j n'importe quelle paire de sommets.

E : ensemble des liens et taille du graphe. Noté E_1, E_2, ...,E_n en cas de réseau multiplexe. e_{ij} désigne tout lien entre les deux sommets v_i et v_j.

I : ensemble des intensités des liens en cas de réseau valué.

G' : sous-graphe $\{V', E'\}$ de G où $V' \subset V$ et $E' \subset E$ ($\subset$ se lit « est inclus dans »).

A : matrice d'adjacence du réseau étudié.

Afin d'alléger certaines notations, un sommet v_i est parfois noté i.

Centralité d'intermédiarité (*betweenness*)

Indicateur très souvent utilisé, pouvant être calculé pour les sommets et pour les liens. La direction des liens n'est généralement pas prise en compte.

Soit $g_{v_j v_k}$ l'ensemble des plus courts chemins entre les sommets v_j et v_k et $g_{v_j v_k}(v_i)$ l'ensemble des plus courts chemins entre les sommets v_j et v_k passant par le sommet v_i avec $v_i \neq v_j \neq v_k$.

L'intermédiarité du sommet v_i est égale à :

$$B(v_i) = \sum_{v_j \neq v_i \neq v_k} \frac{g_{v_j v_k}(v_i)}{g_{v_j v_k}}$$

Il est possible de normaliser cette mesure par le nombre total de plus courts chemins n'incluant pas le sommet v_i :

$$B(v_i)' = \frac{2B(v_i)}{(V-1)(V-2)}$$

La formule est équivalente pour l'intermédiarité des liens.

Si le réseau étudié n'est pas connexe, les intermédiarités ne peuvent être calculées qu'au sein des mêmes composantes.

Si le réseau est valué, l'intensité des liens est considérée comme une distance entre deux sommets et le calcul peut se faire en prenant en compte les plus courtes distances et non les plus courts chemins.

L'intermédiarité peut être calculée sur un réseau bimodal : la formule est la même, le dénominateur devient $(V_1-1)(V_2-2)$ pour la forme standardisée mais il n'est pas totalement certain qu'elle puisse être interprétée aisément d'un point de vue thématique.

Centralité de proximité (*closeness centrality*)

Indicateur mesurant l'éloignement moyen d'un sommet par rapport à l'ensemble des autres sommets du réseau (voir également Shimbel (indices de)).

La centralité de proximité C d'un sommet v_i est égale à :

$$C(v_i) = \frac{V-1}{\sum_{v_j} d(v_i, v_j)}$$

où $d(v_i, v_j)$ désigne la longueur du plus court chemin entre deux sommets v_i et v_j.

Dans le cas d'un réseau valué, si l'intensité est interprétée comme une distance, il est possible de calculer l'inverse de la distance moyenne entre un sommet et tous les autres sommets. Si l'intensité traduit la force de la relation (*i.e.* échanges commerciaux par exemple), il apparaît plus logique de considérer l'intensité moyenne de la relation.

La centralité de proximité peut être calculée pour un réseau bimodal.

Centralité de vecteur propre (*eigenvector centrality*)

Je regroupe dans cette entrée les indicateurs proposés par Katz (1953), Bonacich (1987) et une déclinaison plus récente (PageRank). Cette courte liste n'est pas exhaustive. Dans les trois cas, le principe consiste à pondérer le degré d'un sommet par le degré de ses voisins.

Bonacich propose une mesure de centralité c_i basée sur deux paramètres, α et β. β permet de moduler l'importance du degré des voisins sur la centralité d'un sommet : si la centralité de degré de mes voisins n'a aucune importance pour un type donné de relations, $\beta = 0$ et le calcul est celui du degré ; si au contraire, cela compte beaucoup, $\beta = 1$. S'il est avantageux pour moi d'avoir des voisins avec un degré faible, alors $\beta < 0$. Le paramètre α sert uniquement à contrôler la longueur du vecteur propre obtenu.

$$c_i(\alpha, \beta) = \sum_j (\alpha + \beta c_j) A_{ij}$$

où A_{ij} sont les lignes et colonnes de la matrice d'adjacence correspondant aux voisins du sommet v_i. L'indicateur peut être calculé pour les réseaux orientés et non orientés mais est plus pertinent pour ces derniers [1].

En cas de réseau orienté, l'indice de centralité de Katz (1953) semble plus performant :

$$k_i(\alpha\beta) = \alpha \sum_j A_{ij} d_j + \beta$$

où A_{ij} sont les lignes et colonnes de la matrice d'adjacence correspondant aux voisins du sommet v_i, d_j le degré des voisins j, α et β des constantes positives.

L'indice *PageRank* a été développé par Google pour classer les sites web et est adapté pour les réseaux orientés. Il consiste à pondérer la centralité de degré d'un sommet par la centralité de degré des sommets voisins divisée par le degré sortant. Ceci est pertinent pour les liens entre sites web : le fait que le blog du groupe fmr reçoive un lien depuis google.fr ne suffit pas à en faire un blog central car google.fr envoie x millions de liens vers des sites souvent obscurs.

$$p_i(\alpha\beta) = \alpha \sum_j A_{ij} \frac{d_j}{d_j^{out}} + \beta$$

où d_j est le degré total des voisins j et d_j^{out} le degré sortant de ces mêmes voisins.

Degré (*degree*)

Mesure de centralité d'un sommet (d). Très utilisée, elle correspond au nombre de liens adjacents à un sommet. Soit la matrice d'adjacence A d'un réseau, alors :

$$d_i = \sum_{j=1}^{V} A_{ij}$$

En cas de réseau orienté, on distingue degré entrant (ensemble des liens reçus - *in-degree*) et degré sortant (ensemble des liens émis - *out-degree*) [2].

Le degré moyen dans un réseau non orienté est égal à $\frac{2E}{V}$ (chaque lien a deux extrémités, le nombre total d'extrémités est égal à la somme des degrés).

Dans un réseau orienté, le degré sortant moyen est égal au degré entrant moyen ainsi qu'à E/V.

Dans un réseau valué, il est possible de sommer les intensités des liens adjacents afin d'obtenir un degré pondéré (*weighted degree* ou *strength*). Si ce réseau est par ailleurs orienté, on peut distinguer degré pondéré entrant et sortant.

La normalisation du degré est le plus souvent faite par le degré maximal possible ($V-1$ dans un réseau unimodal simple). Si le réseau est bimodal, la normalisation est faite en fonction de l'ordre de l'autre ensemble de sommets. Si une contrainte est donnée au nombre de liens (questionnaire sociométrique par exemple), il est nécessaire d'adapter le dénominateur pour normaliser le degré.

1. Je renvoie aux pages très claires du manuel de Newman sur ces indicateurs.
2. On trouve dans certains textes les expressions « demi-degré intérieur » et « demi-degré extérieur ».

Densité (*density*)

Mesure portant sur le réseau dans son ensemble. Elle varie entre 0 (réseau vide) et 1 (réseau complet).

Soit G un réseau simple non orienté d'ordre V et de taille E :

$$\delta = \frac{2E}{V(V-1)}$$

Soit G est un réseau simple orienté :

$$\delta = \frac{E}{V(V-1)}$$

Soit G un réseau bimodal, $G = \{V_1, V_2, E\}$:

$$\delta = \frac{E}{V_1 \times V_2}$$

Soit G un réseau planaire :

$$\delta = \frac{E}{3(V-2)}$$

La densité est sensible à l'ordre du réseau : plus ce dernier augmente, plus la densité baisse.

Diversité d'Agresti (*Agresti diversity*)

Indice utilisé pour mesurer l'homogénéité d'un réseau personnel. Soit p_i la proportion d'alters de la catégorie i et k le nombre total de catégories, l'indice est égal à

$$\frac{1 - \sum_{i=1}^{k} p_i^2}{1 - 1/k}$$

Le numérateur peut être interprété comme la probabilité que deux alters pris au hasard appartiennent à deux catégories différentes. Le dénominateur est la valeur maximale que peut prendre cette mesure étant donné le nombre de catégories. L'indice varie entre 0 (homogénéité maximale) et 1 (hétérogénéité maximale).

Modularité (*modularity*)

Indicateur souvent utilisé [1] pour évaluer la qualité d'une partition P en C communautés. Elle peut donc être utilisée pour mesurer l'homophilie relative à un critère catégoriel. En effet, vérifier si beaucoup de liens ont pour extrémités des sommets appartenant à la même partition P est un cas particulier de l'appartenance des deux sommets à une même catégorie.

1. Qu'il soit souvent utilisé ne signifie pas qu'il est le plus pertinent et/ou le plus efficace. Cette remarque est généralisable à tous les indicateurs évoqués dans ce guide.

La direction des liens n'est généralement pas prise en compte. La formule donnée ici est celle proposée en 2004 par Clauset *et al.* et implémentée dans le logiciel *igraph* ; des variantes existent.

$$Q = \frac{1}{2E} \sum_{ij} \left(A_{ij} - \frac{d_i d_j}{2E} \right) \delta(c_i, c_j)$$

où E est le nombre de liens, A_{ij} la valeur présente dans la matrice d'adjacence A pour la ligne i et la colonne j, d le degré et δ une fonction booléenne valant 1 si i et j appartiennent à la même communauté c et 0 dans le cas contraire.

Cette formule fonctionne pour les réseaux valués : l'intensité de la relation remplace 0 ou 1 (A_{ij}) et le degré pondéré est utilisé (d_i, d_j). Le point de comparaison pour mesurer la modularité est un graphe aléatoire (terme $\frac{d_i d_j}{2E}$). Une valeur supérieure à 0.3 indique une bonne modularité, une valeur inférieure à 0 signale un réseau disassortatif.

Cette mesure peut être normalisée par la modularité maximale possible dans un réseau présentant la même distribution de degrés. Si le logiciel que vous utilisez propose plusieurs algorithmes de détection de communautés, comparer la modularité obtenue avec les différents algorithmes peut être utile.

Diverses propositions ont été faites pour adapter cette mesure aux réseaux bimodaux (voir l'article de Barber (2007) et les travaux qui le citent et amendent sa proposition). Aucune à ma connaissance ne s'est imposée.

Shimbel (indices de)

Shimbel dans un article de 1953 propose deux indicateurs (accessibilité et dispersion) permettant de décrire des réseaux de communication, indicateurs utilisés depuis notamment en géographie des transports.

La dispersion D d'un réseau correspond à la somme de tous les plus courts chemins L entre sommets au sein d'un réseau :

$$D = \sum_{i=1}^{V} \sum_{j=1}^{V} L(i, j)$$

Plus la valeur est basse, plus le réseau est considéré comme compact.

L'accessibilité d'un sommet i depuis l'ensemble des autres sommets S [1] est égale à :

$$A(i, S) = \sum_{j=1}^{V} L(i, j)$$

L'inverse de l'accessibilité de Shimbel est égale à la centralité de proximité [2].

1. Je reprends ici la notation utilisée par l'auteur dans son article.
2. J'ai repris ici les formules données par Shimbel dans son article de 1953 ; il serait plus logique que l'accessibilité soit mesurée pour j variant de 1 à S ; la distance d'un sommet à lui-même étant nulle, faire varier j et 1 à V donne le même résultat.

Transivité (*transitivity*)

Indicateur pouvant être calculé sur le réseau dans son ensemble (T) et pour chaque sommet (t). Parfois appelé coefficient de clustering (*clustering coefficient*).

La mesure au niveau du réseau se calcule ainsi :

$$T = \frac{\text{nombre de triades fermées} \times 3}{\text{nombre de triades connexes}}$$

Une triade fermée est un ensemble de 3 sommets ijk où chaque sommet est lié aux deux autres. Une triade connexe désigne 3 sommets ijk où les liens ij et jk sont présents, le lien ik pouvant être présent (triade fermée) ou absent (triade ouverte).

L'indicateur varie entre 0 (arbre, grille rectangulaire, réseau en étoile) et 1 (réseau complet ou réseau où chaque composante est une clique).

Dans leur article de 1998, Watts et Strogatz nomment *global clustering coefficient* la valeur moyenne des transitivités mesurées pour chaque sommet ; les résultats obtenus seront différents si leur formule est utilisée.

Au niveau d'un sommet, l'indicateur est :

$$t_i = \frac{\text{nombre de triades fermées contenant } i \times 3}{\text{nombre de triades connexes contenant } i}$$

Il varie entre 0 (nombre de voisins < 2 ou étoile) et 1.

Différentes propositions existent pour les réseaux valués. Opsahl et Panzarasa (2009) proposent ainsi de calculer une transivité globale T_w où

$$T_w = \frac{\text{Somme des intensités des triades fermées}}{\text{Somme des intensités des triades connexes}}$$

Il est possible d'adapter cet indicateur pour chaque sommet (les auteurs n'évoquent pas cette possibilité dans leur article).

Plus surprenant, l'indicateur a été adapté pour les réseaux bimodaux où, par définition, il ne peut exister de cycle et donc pas de triades fermées. L'intérêt dans ce cas se porte sur les cycles de longueur 4 fermés rapportés aux cycles de longueur 4 ouverts (voir le site de Tore Opsahl pour plus de détails).

Bibliographie

L'url est indiquée uniquement pour les articles librement accessibles en ligne. Les ouvrages anglophones - et parfois francophones - sont disponibles sur libgen. Les articles autres sont sur sci-hub.

Arfaoui, Mehdi, 2021, « Du laboratoire à la plateforme. Enquête sur le Twitter des sociologues en France », *Tracés* 21 : 85-106. [En ligne]

Bahoken, Françoise, Laurent Beauguitte et Serge Lhomme, 2013, *La visualisation des réseaux. Principes, enjeux et perspectives*, Groupe fmr. [En ligne]

Barabási, Albert-László et Réka Albert, 1999, « Emergence of scaling in random networks », *Science* 286(5439) : 509-512. [En ligne]

Barber, Michael J., 2007, « Modularity and community detection in bipartite networks », *Physical Review E*, 76(6) : 066102. [En ligne]

Battiston, Federico, Vincenzo Nicosia et Vito Latora, 2014, « Structural measures for multiplex networks », *Physical Review E*, 89(3) : 032804. [En ligne]

Beauguitte, Laurent, 2011, *L'Assemblée générale de l'ONU de 1985 à nos jours : acteur et reflet du Système-Monde. Essai de géographie politique quantitative*, Thèse de géographie, Université Paris VII. [En ligne]

Beauguitte, Laurent, 2020, Indices Kansky, *Hypergeo*. [En ligne]

Beauguitte, Laurent, 2022, « Théorie des graphes et analyse de réseau en géographie : histoire d'un lien faible (1950-1963) », *PasserelleSHS*, 1 [En ligne]

Beauguitte, Laurent, 2023, « L'analyse de réseaux bimodaux en sciences sociales. Aperçus épistémologiques et historiques », Document de travail, 10 p. [En ligne]

Berge, Claude, 1958, *La théorie des graphes et ses applications*, Dunod.

Berge, Claude, 1970, *Graphes et hypergraphes*, Paris, Dunod.

Berroir, Sandrine, Nadine Cattan, Frédéric Dobruszkes, Marianne Guérois, Fabien Paulus et Céline Vacchiani-Marcuzzo, 2017, « Les systèmes urbains français : une approche relationnelle », *Cybergeo : European journal of geography*. [En ligne]

Bersier, Louis-Félix, 2007, « A history of the study of ecological networks », *Biological networks*, 3 :365 421. [En ligne]

Bidart, Claire, Alain Degenne et Michel Grossetti, 2011, *La vie en réseau. Dynamique des relations sociales*, PUF.

Bonacich, Phillip, 1987, « Power and centrality : A family of measures », *American Journal of Sociology*, 92(5) : 1170-1182. [En ligne]

Bonacich, Phillip, 1991, « Simultaneous group and individual centralities », *Social Networks*, 13(2) : 155-168.

Bott, Elizabeth, 1971, *Family and Social Network*, Tavistock.

Brin, Sergey et Lawrence Page, 1998, The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Computer networks and ISDN systems* 30(1-7) : 107-117. [En ligne]

Briot, Ninon, 2021, *Villes en réseaux. Les coopérations internationales des villes françaises : spatialisation, internationalisation, européanisation*, Thèse de géographie, Université de Lyon. [En ligne]

Burt, Ronald S., 1992, *Structural holes : the social structure of competition*, Harvard University Press.

Cattan, Nadine, 1990, « Une image du réseau des métropoles européennes par le trafic aérien », *L'Espace géographique*, 19–20(2) : 1051-116. [En ligne]

Cazabet, Rémy, 2013, *Détection de communautés dynamiques dans des réseaux temporels*, Thèse d'informatique, Université Paul Sabatier-Toulouse III. [En ligne]

Cazabet, Remy, et Giulio Rossetti, 2019, « Challenges in community discovery on temporal networks », in Petter Holme et Jari Saramäki (dir.), *Temporal network theory*, Springer, 181-197. [En ligne]

Clauset, Aaron, Mark Newman et Moore Cristopher, 2004, « Finding community structure in very large networks », *Physical review E*, 70(6) : 066111. [En ligne]

Coleman, James, Elihu Katz et Herbert Menzel, 1957, « The Diffusion of an Innovation among Physicians », *Sociometry*, 20(4) : 253-270. [En ligne]

Collar, Anna, Tom Brughmans, Fiona Coward et Claire Lemercier, 2014, « Analyser les réseaux du passé en archéologie et en histoire », *Les nouvelles de l'archéologie*, 135 : 9-13. [En ligne]

Crossley, Nick, Elisa Bellotti, Gemma Edwards, Martin G. Everett, Johan Koskinen et Mark Tranmer, 2015, *Social network analysis for ego-nets*, SAGE Publications.

Dancoisne, Pascale et Karel Kansky, 1989, « Measures of network structure », *Flux*, 5(1) : 89-121. [En ligne]

Degenne, Alain et Michel Forsé, 2004 (3e éd.), *Les réseaux sociaux*, Armand Colin.

Di Giacomo, Emilio, Walter Didimo, Fabrizio Montecchiani et Alessandra Tappini, 2021, « A user study on hybrid graph visualizations », *Graph Drawing and Network Visualization : 29th International Symposium*. [En ligne]

Elton, Charles S., 2001[1927], *Animal ecology*, University of Chicago Press. [En ligne]

Eve, Michael, 2002, « Deux traditions d'analyse des réseaux sociaux ». *Réseaux*, 5 : 183-212. [En ligne]

Forsyth, Elaine et Leo Katz, 1946, « A Matrix Approach to the Analysis of Sociometric Data : Preliminary Report », *Sociometry*, 9(4) : 340-347. [En ligne]

Freeman, Linton C., 2004, *The Development of Social Network Analysis. A Study in the Sociology of Science*, Empirical Press.

Gourdon, Paul, 2021, *La coopération entre villes européennes : convergences dans l'action publique urbaine par la circulation transnationale de modèles*, Thèse de géographie, Paris 1. [En ligne]

Granovetter, Mark S.,1973, « The strength of weak ties », *American Journal of Sociology*, 78(6) : 1360-1380. [En ligne]

Grataloup, Christian, 1996, *Lieux d'histoire. Essai de géohistoire systématique*, GIP RECLUS.

Gribaudi, Maurizio (dir.), 1998, *Espaces, temporalités, stratifications*, Éditions de l'EHESS.

Haggett, Peter et Richard J. Chorley, 1969, *Network Analysis in Geography*, Edward Arnold.

Harary, Frank, Robert Zane Norman et Dorwin Cartwright, 1965, *Structural models : An Introduction to the Theory of Directed Graphs*, John Wiley & Sons.

Hennig, Marina, Ulrik Brandes, Jürgen Pfeffer et Ines Mergel, 2012, *Studying Social Networks*, Campus Verlag.

Ings, Thomas C. et Joseph E. Hawes, 2018, « The history of ecological networks », *Ecological networks in the tropics*, 15-28.

Kansky, Karel J., 1963, *Structure of Transportation Networks : Relationships between Network Geometry and Regional Characteristics*, University of Chicago Press.

Katz, Leo, 1953, « A New Status Index Derived from Sociometric Analysis », *Psychometrika*, 18(1) : 39-43. [En ligne]

Kivelä, Mikko, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno et Mason A. Porter, 2014, « Multilayer networks », *Journal of complex networks*, 2(3) : 203-271. [En ligne]

Kleinfeld, Judith S., 2002, « Could it be a big world after all ? The six degrees of separation myth », *Society*, 12, 5. [En ligne]

Klovdahl, Alden S., 1981, « A note on images of networks », *Social Networks*, 3(3) : 197-214. [En ligne]

Lazega, Emmanuel, 2014, *Réseaux sociaux et structures relationnelles*, PUF.

Lee, Clement et Darren J. Wilkinson, 2019, « A review of stochastic block models and extensions for graph clustering », *Applied Network Science*, 4(1) : 1-50. [En ligne]

Le Texier, Thibault, 2018, *Histoire d'un mensonge. Enquête sur l'expérience de Stanford*, La Découverte.

Lemercier, Claire, 2005, « Analyse de réseaux et histoire », *Revue d'histoire moderne contemporaine*, 522(2) : 88-112. [En ligne]

Li Vigni, Fabrizio, 2018, *Les systèmes complexes et la digitalisation des sciences. Histoire et sociologie des instituts de la complexité aux États-Unis et en France*, Thèse de sociologie, EHESS-PSL. [En ligne]

, Johan Koskinen et Garry Robins (dir.), 2013, *Exponential random graph models for social networks : Theory, methods, and applications*, Cambridge University Press.

Maisonobe, Marion, 2015, *Étudier la géographie des activités et des collectifs scientifiques dans le monde : de la croissance du système de production contemporain aux dynamiques d'une spécialité, la réparation de l'ADN*, Thèse de géographie, Université Toulouse 2. [En ligne]

Maisonobe, Marion, Laurent Jégou, Nikita Yakimovich et Guillaume Cabanac, 2019, « NETSCITY : a geospatial application to analyse and map world scale production and collaboration data between cities », *International Conference on Scientometrics and Informetrics*. [En ligne]

Masuda, Naoki et Renaud Lambiotte,Lambiotte, 2020 (2e éd.), *A Guide to Temoral Networks*, World Scientific.

McCarty, Christopher, Miranda J. Lubbers, Raffaele Vacca, José Luis Molina, 2019, *Conducting Personal Network Research. A Practical Guide*, The Guilford Press.

Mercklé, Pierre, 2011, *Sociologie des réseaux sociaux*, La Découverte.

Mitchell, J. Clyde (dir.), 1969, *Social networks in urban situations*, Manchester University Press.

Moreno, Jacob L., 1934, *Who shall survive ? A New Approach to the Problem of Human Interrelations*, Nervous and Mental Disease Publishing Co. [En ligne]

Moreno, Jacob L. et Helen H. Jennings, 1938, « Statistics of social configurations », *Sociometry*, 1(3/4) : 342-374. [En ligne]

Muraco, William A., 1972, « Intraurban Accessibility », *Economic Geography*, 48(4) : 388-405. [En ligne]

Newman, Mark, 2018 [1e éd., 2013], *Networks*, Oxford University Press.

Northway, Mary L., 1940, « A method for depicting social relationships obtained by sociometric testing », *Sociometry*, 3(2) : 144-150. [En ligne]

Nystuen, John D. et Michael F. Dacey, 1961, « A graph theory interpretation of nodal regions », *Papers of the Regional Science Association*, 7(1) : 29-42. [En ligne]

Opsahl, Tore et Pietro Panzarasa, 2009, « Clustering in weighted networks », *Social Networks*, 31.2 : 155-163. [En ligne]

Opsahl, Tore, 2013, « Triadic closure in two-mode networks : Redefining the global and local clustering coefficients », *Social Networks*, 35(2) : 159–167. [disponible sur sci-hub]

Padgett, John F. et Christopher K. Ansell, 1993, « Robust Action and the Rise of the Medici, 1400-1434 », *American journal of sociology*, 98(6) : 1259-1319. [En ligne]

Perry, Brea L., Bernice A. Pescosolido et Stephen P. Borgatti, 2018, *Egocentric Network Analysis. Foundations, Methods, and Models*, Cambridge University Press.

Shimbel, Alfonso, 1953, « Structural Parameters of Communication Networks », *The Bulletin of Mathematical Biophysics*, 15(4) : 501-507. Version bilingue et commentée. [En ligne]

Snyder, David et Edward L. Kick, 1979, « Structural position in the world system and economic growth, 1955-1970 : A multiple-network analysis of transnational interactions », *American Journal of Sociology*, 84(5) : 1096-1126. [En ligne]

Sade, Donald Stone, 1965, « Some aspects of parent-offspring and sibling relations in a group of rhesus monkeys », *American Journal of Physical Anthropology*, 23(1) : 1-17. [En ligne]

Sueur, Cédric (dir.), 2015, *Analyse des réseaux sociaux appliquée à l'éthologie et l'écologie*, Éditions Matériologiques.

Tufte, Edward R., 2001, *The visual display of quantitative information*, Graphics Press.

Valdivia, Paola, Paola Buono, Catherine Plaisant, Nicole Dufournaud et Jean-Daniel Fekete, 2018, Using Dynamic Hypergraphs to Reveal the Evolution of the Business Network of a 17th Century French Woman Merchant, *Proceeding of the VIS4DH Workshop*, 1-5. [En ligne]

Wasserman, Stanley et Katherine Faust, 1994, *Social Network Analysis. Methods and Applications*, Cambridge University Press.

Watts, Duncan J., 2003, *Six Degrees. The Science of a Connected Age*, William Heinemann.

Watts, Duncan J. et Steven H. Strogatz, 1998, « Collective dynamics of 'small-world' networks », *Nature* 393(6684) : 440-442. [En ligne]

Wilke, 2019,*Fundamentals of Data Vizualization*, O'Reilly. [En ligne]

Wyngaerden, François, Marie Tempels, Jean-Louis Feys, Vincent Dubois et Vincent Lorant, 2020, « The personal social network of psychiatric service users », *International Journal of Social Psychiatry*, 66(7) : 682-692. [En ligne]

Zachary, Wayne W., 1977, « An Information Flow Model for Conflict and Fission in Small Groups », *Journal of Anthropological Research*, 33(4) : 452-473. [En ligne]

Table des figures

Index

Table des matières