L'analyse des réseaux - exemple du métro parisien

Un réseau est un ensemble de points rattachés entre eux par des liens. Une carte routière, un organigramme d'entreprise ou encore un arbre généalogique constituent des réseaux, pierre angulaire des travaux qui seront présentés dans la suite. Objets d'études en informatique, et plus précisément en théorie des graphes, ils sont un outil de modélisation utilisé par de nombreuses disciplines de sciences appliquées. En sociologie, le succès des réseaux, aussi bien dans la recherche que dans un certain imaginaire collectif, leur a valu de prêter leur nom aux plateformes de mise en relation de personnes via internet, telle Facebook, une des plus célèbres, dont les données qui sont étudiées ici sont issues.

Formellement, un réseau (ou un graphe) est donc formé par un ensemble de points qu'on appelle des *nœuds*, qui sont reliés les uns aux autres par des *liens*. Deux nœuds

ainsi reliés entre eux sont appelés des *voisins* ou sont dits voisins l'un de l'autre. Les réseaux sont extrêmement utiles car ils permettent de modéliser de nombreux objets ou situations qui décrivent des relations entre des éléments. Une liste non exhaustive d'exemples de telles situations sera décrite dans la section 1.3.3.

Avant de poursuivre, il est à noter qu'un réseau est souvent confondu avec sa visualisation, alors qu'en pratique un réseau est généralement défini comme un ensemble de relations, par exemple [*a–b a–c a–d b–c d–e*]. Dans ce cas, le réseau est composé de 5 sommets, aux noms de *a*, *b*, *c*, *d* et *e*. *a* est relié à *b*, *c* et *d*, *b* est également relié à *c* et *d* et *e* sont voisins. C'est seulement à partir de cette description qu'un algorithme dit de visualisation (il en existe d'ailleurs une grande variété, voir [Battista et al., 1998] pour une large présentation de ceux-ci) permet de dessiner le réseau. La forme dessinée dépend donc autant de l'algorithme choisi que du réseau lui-même, comme l'illustre la Figure 1.1 et dont les choix ont été explorés par [Henry, 2008]. On conjugue en général la visualisation d'un réseau, utile à l'interprétation par l'œil humain et à l'émission d'hypothèses, avec sa représentation structurale qui permet sa manipulation par des ordinateurs, de manière plus rapide et précise que le dessin.

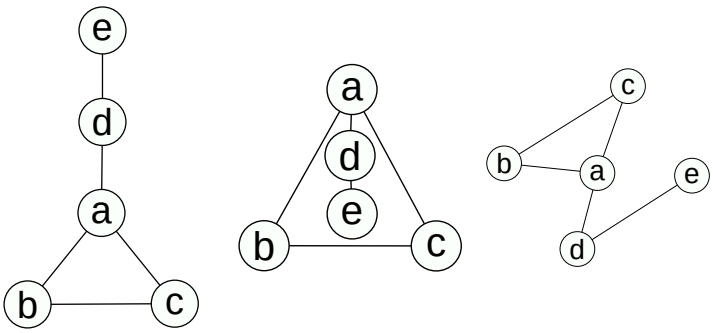

FIGURE 1.1 – Plusieurs visualisations possibles et toutes structuralement équivalentes du réseau [*a–b a–c a–d b–c d–e*].

La Figure 1.2 présente ainsi une visualisation du réseau du métro parisien calculée par l'algorithme Force Atlas 2. Les stations y sont positionnées de telle manière que celles qui sont connectées entre elles sont rapprochées l'une de l'autre. Dans un réseau géographiquement contraint, comme celui-ci, on obtient un plan qui semble à peu de choses près calqué sur celui de Paris, bien que l'algorithme de visualisation utilisé ne connaisse pas la notion de points cardinaux ni la longueur des tunnels entre deux stations. Un réseau des lignes aériennes, par exemple, serait certainement moins proche de la réalité, puisqu'un aéroport connecté aux quatre coins du monde peut très bien être géographiquement proche d'un aéroport régional, sans pour autant être relié à ce dernier.

Un exemple tiré du réseau du métro va servir de dernière remarque pour souligner la différence entre structure du réseau et visualisation. En haut de la carte, la station Marcadet-Poissonniers est reliée à deux branches de deux stations chacune qui partent vers le nord. L'une des deux correspond à la ligne 12 du métro (ce plan est issu de données anciennes et la ligne a depuis été prolongée) et l'autre à la ligne 4. Ici l'algorithme a choisi aléatoirement quelle serait la ligne qui serait dessinée la plus à droite de la figure et laquelle serait à gauche, sans savoir ce qu'il en est en réalité. En l'absence de l'affichage du nom des stations, il est alors impossible de déterminer si la réalité a été respectée.

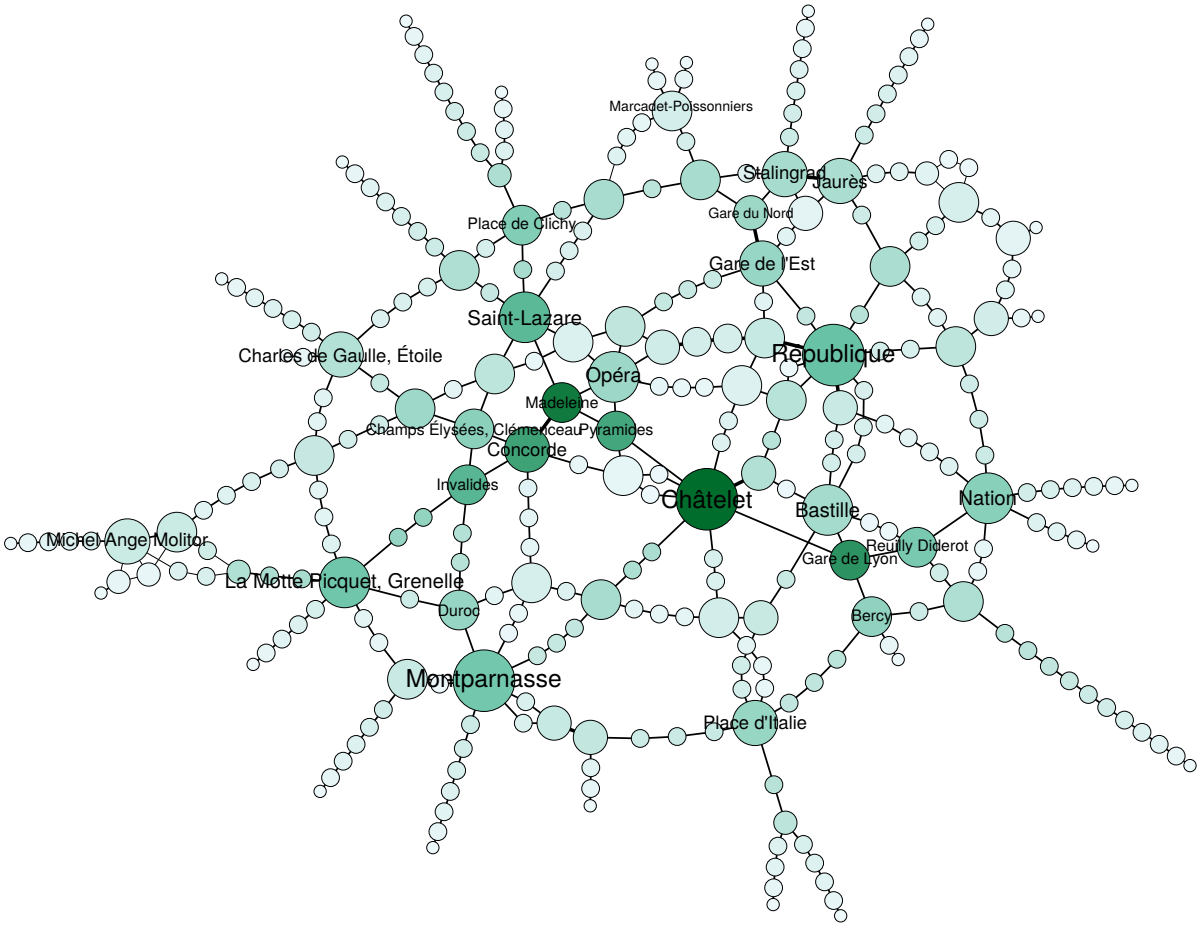

FIGURE 1.2 – Un réseau représentant les stations du métro parisien. Le réseau, comme les autres présentés dans ce manuscrit, est exporté depuis le logiciel de visualisation et de manipulation de graphes Gephi.

La similitude entre le réseau du métro et le positionnement effectif des stations illustre néanmoins en partie l'intérêt porté aux réseaux : une connaissance peu élaborée d'un

système éventuellement complexe, soit simplement la liste des connexions deux à deux qui le composent permet la construction d'un objet dont l'étude peut mener à une interprétation fidèle à la réalité. Au delà de leur visualisation, les réseaux permettent en effet de mener des analyses mathématiques au travers des outils proposés par la *théorie des graphes* évoquée plus en détail en section 1.3.1, *graphe* étant le nom de l'objet mathématique représentant la structure du réseau. Cette théorie, portée par des mathématiciens ou des informaticiens spécialisés, ainsi que par des chercheurs issus d'autres disciplines variées et ayant utilisé les réseaux pour leurs recherches, a favorisé l'émergence de mesures permettant l'identification automatisée de la forme du réseau, de la centralité de chacun de ses nœuds ou encore de l'importance des liens, pour ne citer que quelques exemples.

Dans le cas du plan du métro, par exemple, pour décider de la taille de chacun des nœuds, j'ai utilisé le nombre de liens qui leur sont adjacents, soit leur nombre de voisins. Cette mesure du nombre de voisins, qu'on appelle le *degré*, vaut 1 pour la plupart des terminus, généralement seulement reliés à la station qui les précède sur la ligne, et qui sont donc les plus petits du dessin. Les stations qui ne sont pas des terminus mais qui n'ont pas de correspondance ont un degré de 2, car elles sont reliées à une station de chaque côté. À l'inverse, les stations par lesquelles passent le plus de lignes de métro, comme Châtelet, République ou Montparnasse ont un degré plus important et apparaissent en gros sur la carte.

De manière analogue, la couleur des nœuds a été choisie en fonction de leur *centralité*. S'il existe de nombreuses définitions de la centralité qui seront évoquées dans la section 1.3.1, c'est ici la centralité dite d'*intermédiarité* que j'ai choisi d'utiliser. Le score de centralité d'intermédiarité d'une station est proportionnel au nombre de trajets pour lesquels elle se trouve sur le chemin (la succession de liens entre un point et un autre) le plus court. Plus un nœud est vert foncé et plus la station qu'il représente est centrale au sens de l'intermédiarité. Les stations les plus centrales du réseau parisien sont Châtelet, Madeleine et Gare de Lyon. Au contraire, aucun plus court chemin ne passe par les stations en bout de ligne qui sont donc les plus claires.

Munis simplement de ces deux métriques, on peut déjà remarquer plusieurs catégories de stations.

	Faible degré	Fort degré
Centralité faible	La majorité	Michel-Ange Molitor, Stalingrad, Jaurès, Charles de Gaulle Étoile, ...
Centralité forte	Madeleine, Gare de Lyon, ...	Châtelet, République, Montparnasse, ...

Cette classification succincte suggère d'ores et déjà des similitudes entre les stations de mêmes catégories. Par exemple, les stations peu centrales mais à degré important sont des stations périphériques qui offrent des correspondances entre des lignes circulaires et des lignes reliant le centre de la capitale aux stations extérieures. Les stations à faible degré mais centrales sont situées entre des stations qui sont elles-mêmes plutôt centrales et à degrés importants et profitent ainsi en quelque sorte de l'importance de celles-ci puisqu'elles permettent de les atteindre.

Loin d'être exhaustive, cette présentation du potentiel offert par la science des réseaux et la théorie des graphes est en un avant-goût de ce que nous allons aborder au long de cette thèse en appliquant, cette fois, ce genre de méthodes d'analyse aux réseaux sociaux.

1.2 Le réseau social

Si les réseaux servent à la représentation, visuelle et mathématique, de toutes sortes d'objets, comme un plan de métro, c'est sur les relations sociales que se penche ce travail de doctorat. Polysémique, le terme de réseau social désigne en sociologie des réseaux un réseau dont les nœuds sont des individus ou des organisations qui sont reliés selon des critères de connaissance, de relations d'échanges, ... L'étude du réseau social vise à analyse comment le positionnement structurel des agents en son sein permet d'interpréter leur influence dans l'environnement observé.

Dans cette section, je vais présenter un réseau social et mettre en avant quelques notions et quelques questions autour desquelles repose ce travail. La figure 1.3 propose donc une visualisation du réseau social. Chaque nœud représente un individu et deux nœuds sont reliés entre eux si les deux personnes qu'ils représentent se connaissent entre elles.

Comment a-t-on construit ce réseau ? Toutes les personnes qui y figurent sont en fait les « amis » Facebook, le terme utilisé par la plate-forme pour désigner les contacts, d'un répondant à l'enquête Algopol, à laquelle j'ai collaboré au cours de mon doctorat. Cette personne est une jeune femme de 25 ans qui vivait dans les Yvelines au moment de l'enquête. Elle n'apparaît pas dans le réseau qui est pourtant celui formé par ses relations car le nœud qui la représenterait serait alors relié à tous les autres. Il n'apporterait aucune information supplémentaire et « écraserait » le reste par son omniprésence. On appelle *réseau personnel* ou bien *réseau égocentré* un tel réseau, composé par les contacts d'une personne. On reviendra plus en détail sur cette notion dans la section 1.4.3. Dans le reste du manuscrit, pour chaque réseau personnel qu'on rencontrera et selon les termes usuels, j'appellerai ainsi *ego* l'enquêté auquel il

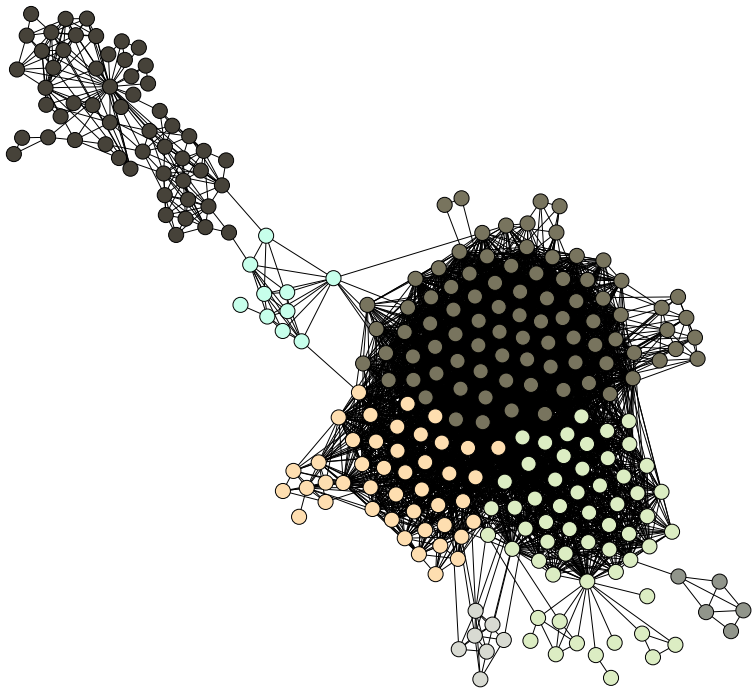

FIGURE 1.3 – Un réseau social

appartient, tandis que les individus qui le composent seront nommés les *alters*.

Comme on peut l'observer, l'organisation de ce réseau est bien différente de celle du métro parisien : il possède plusieurs groupes de nœuds très interconnectés entre eux. Ces groupes, assez différents les uns des autres, sont reliés entre eux par quelques individus tandis que le réseau précédent présentait une structure d'ensemble nettement plus homogène. La forme des réseaux est en fait très dépendante de leur nature, géographique, sociale, trophique, ..., et on verra d'ailleurs dans les sections 4.4 et 4.6.3 que des méthodes permettent de déterminer la discipline dont est issu un réseau à partir des caractéristiques structurales de celui-ci.

Ici, les couleurs des nœuds représentent les communautés d'individus, les nœuds d'une même couleur appartenant à la même communauté. Ces dernières ont été calculées par un algorithme dit de détection de communautés en regroupant les nœuds qui sont plus connectés ensemble qu'avec les autres, ce ne sont donc *a priori* pas des communautés connues ou observées mais bien détectées structuralement. Selon les critères de l'algorithme, le réseau est composé de huit communautés. On repère de visu deux groupes principaux d'individus :

— le premier, en bas à droite est très dense, composé de trois communautés ;

— en haut à gauche, le second groupe important (en gris foncé) est moins dense, et on observe qu'un de ses alters semble être ami avec la majorité des membres de la communauté.

Un petit groupe d'alters, en bleu, opère la jonction entre ces deux groupes principaux.

Le fait que le groupe très dense à droite soit découpé en trois sous-communautés distinctes semble contre-intuitif car on aurait probablement imaginé qu'il forme une unique communauté. En fait, on sait grâce aux questions auxquelles a répondu l'enquêté qu'il correspond aux étudiants et enseignants des trois classes de son établissement d'études : de nombreux élèves ont des liens avec d'autres classes mais la majorité de leurs connaissances sont dans la même classe qu'eux. De plus, les professeurs intégrés au groupe créent encore plus de liens entre les élèves des différentes classes.

Si les réseaux personnels nous apprennent peu concernant les alters, ceux-ci étant vus au prisme restreint de leur relation commune avec égo, ils offrent néanmoins, comme on le verra dans la suite, une grille de lecture pertinente de la sociabilité de ce dernier. Comment alors interpréter ce réseau ? Et est-il possible d'imaginer quels sont les amis proches d'*ego* à partir de son observation ? On imagine souvent que plus on a d'amis ou de relations en commun avec une personne et plus il y a de chance que cette personne soit importante pour nous. Ce réseau suggère le contraire : les alters issus de la communauté estudiantine sont bien ceux qui ont les plus importants degrés (le nombre de liens avec d'autres alters, et donc le nombre d'amis communs avec égo, l'enquêté) mais il semble aussi peu probable qu'*ego* ait de fort liens interpersonnels avec autant d'individus et la forte interconnexion serait alors plus probablement la cause d'un effet structurant des écoles.

La littérature montre, et on aura l'occasion d'explorer ces résultats dans la section 1.4.2 qu'il est en fait plus pertinent de se pencher sur les *alters* ayant des liens avec des gens qui sont par ailleurs peu reliés entre eux. Effectivement, dans ce cas, c'est soit *ego* qui a présenté l'*alter* en question à d'autres de ses connaissances, soit c'est l'*alter* lui même qui a introduit *ego* à des amis à lui, amis qui sont par la suite devenus des contacts Facebook d'*ego* puisqu'ils apparaissent dans ce réseau. Dans les deux cas, une relation, sinon forte au moins réelle, existe entre *ego* et cet *alter*.

Au cours de mon travail de thèse, j'ai pu explorer des méthodes de qualification des amis de nos enquêtés, qu'on verra en section 3.3 et dans le Chapitre 6.2. Il est néanmoins nécessaire d'aborder dans un premier temps les outils d'analyse des graphes pour bien les appréhender.

1.3 Les réseaux et les graphes

Si dans toutes les disciplines scientifiques ou presque, des chercheurs utilisent maintenant les réseaux pour modéliser les interactions qu'entretiennent les éléments de leurs objets d'études, les mathématiciens et les informaticiens qui proposent des méthodes fondamentales, non appliquées à des problèmes concrets, travaillent eux sur un objet similaire qu'ils nomment plus volontiers des *graphes*. Le graphe est l'objet mathématique qui représente un réseau. Il n'est pas composé de nœuds et de liens mais d'arêtes et de sommets. Les deux terminologies diffèrent mais sont équivalentes et, comme beaucoup, je les utiliserai indistinctement au long de ce manuscrit.

1.3.1 Théorie des graphes

La *théorie des graphes* est la discipline mathématique et informatique qui traite de l'étude des graphes. C'est au mathématicien suisse Leonhard Euler qu'on attribue la paternité du concept et le premier résultat de la discipline à la suite de sa résolution, en 1736, d'une énigme populaire de l'époque.

Au début du 18^{eme} siècle, Königsberg faisait partie du Royaume de Prusse, avant de devenir Kaliningrad la capitale de l'enclave russe en Europe. Elle est construite autour de deux îles et traversée par la Pregolia, un fleuve qui se jette dans la mer Baltique, 7 ponts permettant aux habitants de la ville de rejoindre les îles ou l'autre rive du fleuve. L'énigme consistait à savoir s'il leur était possible de traverser tous les ponts sans passer deux fois par le même. S'emparant de la question, Euler propose une représentation en réseau de la ville dans laquelle chaque rive, ainsi que les deux îles qui se dressent sur le fleuve, est représentée par un sommet et chaque pont par une arête comme illustré par la figure 1.4. Afin de traverser chaque pont une seule fois, Euler indique qu'il est nécessaire que chaque sommet du graphe ait un nombre pair d'arêtes adjacentes, à l'exception du point de départ et du point d'arrivée. Puisque chaque sommet a un nombre impair d'arêtes, il n'est donc pas possible aux habitants de Königberg de faire une telle promenade.

Depuis Euler et ses premiers travaux, la théorie des graphes a servi toile de fond d'une utilisation de plus en plus importante des réseaux à travers les différents champs de la science. C'est pourquoi ont été construits de nouveaux types de graphes, permettant de modéliser fidèlement par les réseaux des phénomènes plus nombreux, nécessitant par ailleurs les adaptations des algorithmes de calcul des mesures déjà existantes, qu'elles aient été développées par les chercheurs en théorie des graphes eux-mêmes ou bien par les spécialistes de disciplines où les réseaux sont utilisés comme outil d'analyse. Ces algorithmes n'ont en même temps eu de cesse d'être améliorés, produisant donc

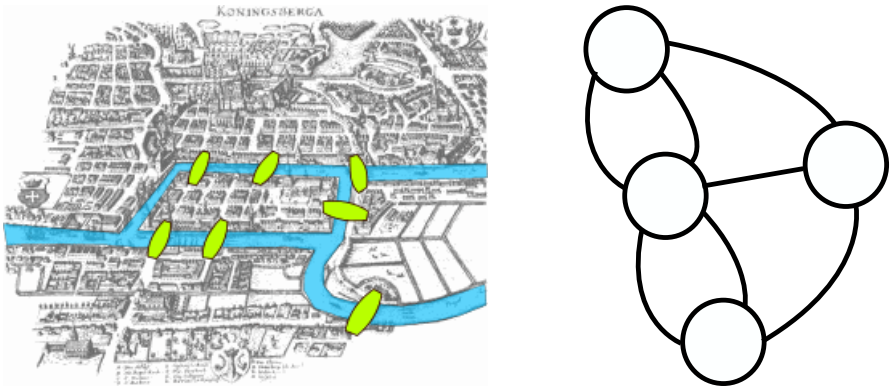

FIGURE 1.4 – Le graphe représentant les ponts de Königsberg

leurs résultats de plus en plus rapidement et pour des graphes aux tailles de plus en plus importantes. Dans cette partie, je propose un passage en revue plus formel mathématiquement de méthodes et résultats de la théorie des graphes qui vont permettre de formaliser les premières intuitions d'analyses qui ont été présentées en amont et qui sont un prérequis pour une partie des analyses proposées dans ce manuscrit.

Un **graphe** est un couple $G = (V, E)$ composé d'un ensemble de sommets V (pour *vertices* en anglais) et d'un ensemble d'arêtes E (pour *edges*). Chaque arête de E est représentée par un couple de sommets qu'elle relie entre eux. On note généralement pour un graphe G donné, $V(G)$ l'ensemble de ses sommets et $E(G)$ celui de ses arêtes.

En poursuivant l'exemple de Königsberg, si on nomme n et s respectivement les rives nord et sud de la ville et i_o et i_e la petite île à l'ouest et la plus grande à l'est, alors le graphe qu'on appelle K représentant les ponts est :

$$\begin{aligned} K = (& \\ & V : (n, s, i_o, i_e), \\ & E : ((n, i_o), (n, i_o), (n, i_e), (i_o, s), (i_o, s), (i_o, i_e), (i_e, s)) \\ ) & \end{aligned}$$

On note généralement n le nombre de sommets d'un graphe et m le nombre de ses arêtes. Ici n vaut 4 et m 7. J'utiliserai ces notations dans l'ensemble du manuscrit.

Depuis Euler, plusieurs types de graphes ont donc été construits afin d'étudier des réseaux à même de modéliser des phénomènes relationnels variés.

Les graphes **simples** n'ont pas de boucle, c'est-à-dire d'arête allant d'un sommet vers lui même et chaque couple de sommets est relié par au plus une arête. Le réseau construit par Euler pour modéliser le problème des 7 ponts ne peut donc pas être représenté par un graphe simple puisque deux arêtes relient l'île ouest à la rive sud et deux autres à la rive nord. En le réduisant à un graphe simple de K, on obtiendrait :

$$
\begin{aligned}
K\text{simple} = (& \\
& V : (n, s, i_o, i_e), \\
& E : ((n, i_o), (n, i_e), (i_o, s), (i_o, i_e), (i_e, s)) \\
)&
\end{aligned}
$$

Ce graphe ne permet plus de répondre à l'énigme mais nous apprend toujours qu'il faut passer par une des îles pour traverser le fleuve. Dans la suite de ce travail, on utilisera exclusivement des graphes simples. Je vais néanmoins présenter quelques autres types de graphes à titre d'exemples.

Les **arbres** sont des graphes sans cycles, c'est-à-dire qu'il n'existe pas plus d'un chemin simple (soit sans passer plusieurs fois par la même arête) entre deux sommets.

Les arêtes d'un graphe **orienté** sont dirigées d'un sommet vers l'autre, contrairement au cas **non orienté** où elles indiquent une relation réciproque entre deux sommets. Une modélisation par un réseau de Twitter nécessiterait l'emploi d'un graphe orienté puisqu'y « suivre » Barack Obama n'indique pas que lui même nous « suive ». À l'inverse, un graphe non orienté est adapté à la modélisation de réseaux d'« amitiés » Facebook où la relation est mutuelle. Notons que la contrainte de la simplicité n'est pas exactement la même pour un graphe orienté que dans le cas non orienté et qu'il y est possible pour un couple de deux sommets d'être reliés par deux arêtes, à condition qu'elles soient dans deux sens différents.

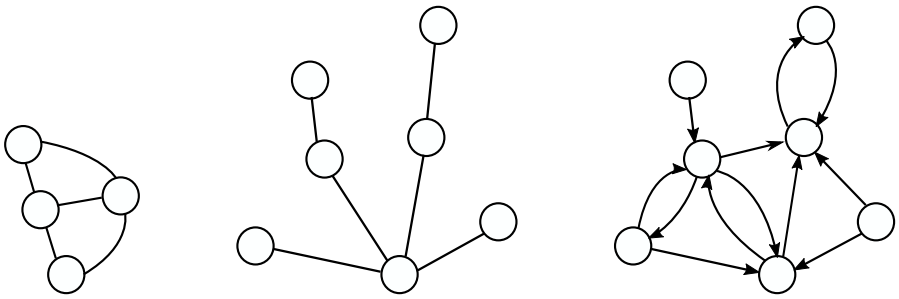

FIGURE 1.5 – De gauche à droite : le graphe de Konigsberg simplifié, un arbre et un graphe orienté. Un graphe orienté simple, comme celui-ci, n'a pas plus d'une arête d'un sens donné entre deux sommets.

Les graphes **pondérés** attribuent un poids à chacune de leurs arêtes, ce qui est très utile pour représenter des distances, dans le cas, par exemple, de réseaux de chemins de fer ou bien l'intensité des relations, dans le cas des réseaux sociaux. Si on avait ajouté au réseau du métro l'information de la longueur des tunnels pour chaque arête, on aurait probablement eu l'exacte carte de Paris, à symétrie près, en utilisant un algorithme de visualisation tenant compte de la pondération des liens.

Les graphes **multi-niveaux** ont la particularité d'avoir plusieurs types de sommets et d'arêtes. Ils sont utilisés dans le cas où l'ont veut par exemple représenter des relations entre individus appartenant à plusieurs organisations, elles mêmes reliées entre elles. Dans ce cas on aurait deux types de sommets (individus et organisations) et trois types d'arêtes (inter-individus, inter-organisations, entre individus et leurs organisations) qu'il faudrait éventuellement traiter différemment. On peut notamment citer [Lazega et al., 2007] comme exemple d'utilisation dans le champ des réseaux sociaux.

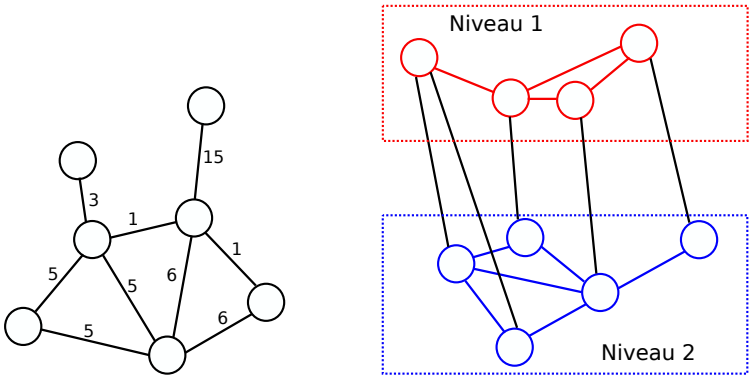

FIGURE 1.6 – De gauche à droite : un graphe pondéré et un graphe multi-niveaux, dont on voit les liens entre les éléments de même niveau et entre les éléments de niveaux distincts.

Dans ce manuscrit, on manipule donc des graphes simples, non orientés et non pondérés. Les quelques définitions usuelles qui suivent concernant ces graphes seront régulièrement employées dans la suite :

Voisinage : Le voisinage $N(v)$, pour *neighborhood*, d'un sommet v est l'ensemble de ses voisins.

$$N(v) = \{u \in V | (u, v) \in E\}.$$

On utilise également parfois le voisinage d'un ensemble de sommets.

$$N(V' \subset V) = \{u \in V \setminus V' | \exists v \in V' \text{ tel que } (u, v) \in E\}$$

On considère ici uniquement les sommets de $V \setminus V'$ afin de ne pas avoir de sommets de V' dans le voisinage de V' mais on peut aussi trouver dans la littérature la définition l'autorisant, auquel cas le voisinage tel que je le définit est parfois appelé **voisinage ouvert**.

Degré : Le degré d'un sommet v, $d(v) = |N(v)|$ est son nombre de voisins, ou la taille de son voisinage.

Distribution des degrés : La distribution des degrés d'un graphe est la liste (éventuellement ordonnée) des degrés de ses sommets.

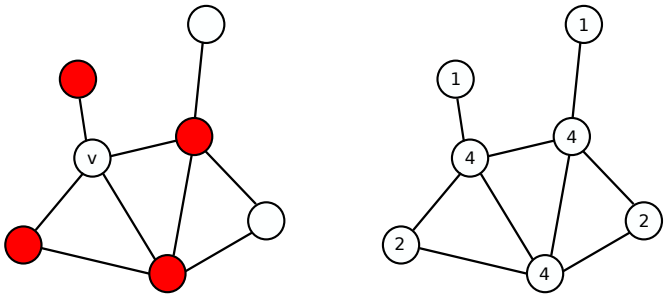

FIGURE 1.7 – À gauche, un réseau dans lequel le voisinage du sommet v est indiqué en rouge. À droite, le degré de chaque sommet. La distribution des degrés triée du graphe est donc 1,1,2,2,4,4,4

Pour calculer l'un de ces trois éléments (voisinage d'un sommet, degré d'un sommet, distribution des degrés) dans un graphe, il faut parcourir l'ensemble de ses arêtes. Par exemple, pour le degré d'un sommet, lors du parcours des arêtes, on ajoute 1 à la variable représentant son résultat à chaque fois qu'on trouve le sommet en question dans l'une d'entre elles. Pour le voisinage, on ajouterait l'autre sommet de l'arête à la liste représentant le voisinage. On dit qu'un algorithme calculant ces valeurs a une complexité en $O(m)$ car le nombre d'opérations qu'il devra réaliser est proportionnel au nombre m d'arêtes du graphe.

Densité : Pour un nombre donné de sommets, la densité d'un réseau augmente avec le nombre de liens entre eux. Elle est formellement définie par la formule :

$$\text{densité} = \frac{2 \times m}{n \times (n-1)}$$

où m est le nombre d'arêtes du réseau et n est son nombre de sommets. C'est en fait le rapport entre le nombre d'arêtes qui existent dans le réseau et le nombre maximal qu'il aurait pu en compter. En effet, le nombre de liens possibles entre n sommets est $\frac{n \times (n-1)}{2}$ puisque chacun des n sommets a dans ce cas $n-1$ voisins. La division par 2 vient du fait que dans ce cas, chaque arête est comptée deux fois au lieu d'une. Un graphe dont tous les sommets sont connectés entre eux est appelé un *graphe complet*. Puisque m varie entre 0 et $\frac{n \times (n-1)}{2}$, la densité varie elle entre 0, dans le cas d'un réseau où il n'existe aucune connexion et 1, dans le cas d'un réseau complet.

Coefficient de clustering ou **transitivité** : Le coefficient de clustering est une mesure de la densité locale du réseau. Il traduit, pour chacun de ses sommets, le degré de

connectivité de ses voisins. Il en existe plusieurs définitions. Watts et Strogatz proposent de prendre la moyenne, pour chaque sommet du réseau de la proportion de ses voisins qui se connaissent entre eux [Watts and Strogatz, 1998] tandis que Barrat et Weigt calculent le rapport entre le nombre de triplets de sommets fermés (c'est à dire avec trois arêtes, où chacun des sommets est relié aux deux autres) et le nombre de triplets connectés (les triplets fermés ou ouverts, un triplet ouvert étant composé d'un sommet central relié aux deux autres qui ne sont pas eux mêmes reliés entre eux) [Barrat and Weigt, 2000]. Le coefficient de clustering est consécutif de la découverte du concept de *petit monde* sur lequel je reviendrai en section 1.4.2.

Chemin : Un chemin est une suite de sommets tels que deux sommets successifs sont voisins l'un de l'autre. n–i_o–s–i_e–s est par exemple un chemin valide pour une promenade à Königsberg tandis que n–s–i_o–i_e–n ne l'est pas puisqu'il n'y a pas d'arêtes entre n et s.

Longueur d'un chemin : C'est le nombre d'arêtes empruntées par le chemin. Par exemple, n–i_o–s–i_e–s a pour longueur 4.

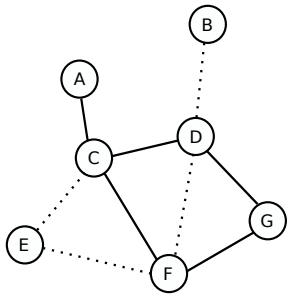

FIGURE 1.8 – Les arêtes pleines correspondent à celles qui sont sur l'un de deux plus courts chemins A-C-F-G et A-C-D-G entre A et G. Ils sont de longueur 3

Composante connexe : On dit d'un ensemble de sommets d'un graphe qu'ils forment une composante connexe si, pour n'importe quel couple de sommets différents pris parmi eux, il existe au moins un chemin allant de l'un à l'autre. On appelle taille d'une composante connexe son nombre de sommets.

Graphe connexe : Un graphe est connexe si tous ses sommets appartiennent à la même composante connexe. La figure 1.9 montre deux graphes avec plusieurs composantes connexes.

Sommet isolé : Un sommet isolé est un sommet qui n'a aucun voisin. Il a donc un degré nul. Un sommet isolé est donc une composante connexe de taille 1. La figure 1.9 contient notamment un graphe avec des sommets isolés. Dans la suite, on va noter **isolés**(G) l'ensemble des sommets isolés d'un réseau G.

FIGURE 1.9 – À gauche un graphe avec deux composantes connexes ACE et BDFG et à droite un graphe à trois composantes connexes où E et B sont des sommets isolés.

Sous-graphe : un sous-graphe $G' = (V', E')$ de G est un graphe composé d'une partie des sommets de G $V' \subset V$ et d'arêtes prises parmi celles reliant deux de ces sommets $E' \subset \{(v_1, v_2) \in E | v1 \in V', v_2 \in V'\}$. Quelques sous-graphes d'un graphe sont présentés en Figure 1.10.

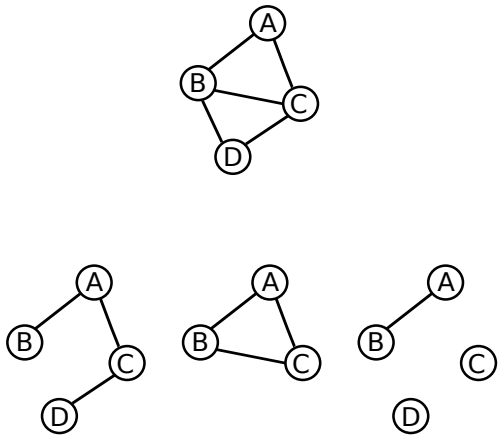

FIGURE 1.10 – Un graphe et trois de ses sous-graphes parmi la multitude de possibles.

On dit d'un sous-graphe G' de G qu'il est un **sous-graphe induit** si et seulement si $E' = \{(v_1, v_2) \in E | v1 \in V', v_2 \in V'\} \subset E$ contient toutes les arêtes de E qui sont entre deux des sommets de V'. Toutes les arêtes possibles sont donc prises dans le sous-graphe induit. La figure 1.11 illustre le concept de sous-graphe induit. Dans la suite, pour un graphe $G = (V, E)$ et un sous-ensemble V' de V, on notera $G|V'$ le sous-graphe de G induit par V'.

Les sous-graphes étant eux-mêmes des graphes, les mesures relatives aux graphes ou aux réseaux qu'on a préalablement définies peuvent également leur être appliquées.

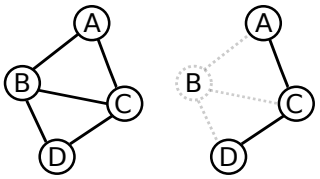

FIGURE 1.11 – Un graphe et son unique sous-graphe induit à A, C et D.

Ils ont donc un diamètre, une connexité, des sommets plus ou moins centraux, etc. Les propriétés structurales d'un graphe n'ont pas une influence très importante sur un ou quelques-uns de ses sous-graphes induits pris au hasard : par exemple, un graphe connexe peut très bien avoir un grand nombre de sous-graphes induits non connexes.

Communauté ou **cluster** : La définition de communauté est ambiguë et peut dépendre de modèles mathématiques comme des connaissances empiriques des chercheurs sur les données modélisées par leurs réseaux. Elle correspond néanmoins intuitivement à des groupes de sommets qui sont plus fortement reliés entre eux qu'avec le reste du réseau, comme c'est le cas dans la Figure 1.12. De nombreuses méthodes de détection des communautés ont été proposées dans la littérature [Clauset et al., 2004, Flake et al., 2002]. Un exemple que je trouve particulièrement élégant est celui de Newman et Girvan qui supprime tour à tour du réseau les arêtes ayant les plus fortes centralités d'intermédiarité (de manière analogue aux sommets, on peut trouver des arêtes du réseau qui sont empruntées par de nombreux plus courts chemins) jusqu'à séparer le réseau en composantes connexes qui forment alors ses communautés [Newman and Girvan, 2004]. Pour une revue de littérature récente des algorithmes de détection de communauté, il est possible de se référer à [Javed et al., 2018]. L'une des méthodes les plus utilisées, et notamment dans ce manuscrit, est celle proposée par Blondel et son équipe et qui se base sur la *modularité*, définie juste après [Blondel et al., 2008].

Modularité : La modularité, introduite par Newman et Girvan [Newman and Girvan, 2004] est un indicateur de la qualité du découpage d'un graphe en communautés de sommets. Elle est définie comme étant la différence entre les proportions d'arêtes incluses dans chaque communauté et celles qui auraient été obtenues pour un graphe aléatoire de même nombre d'arêtes et de sommets. On verra plus en détail les graphes aléatoires en section 1.3.3. La modularité a été l'inspiration de nombreux algorithmes de détection de communautés tels que la méthode dite de Louvain proposée par Vincent Blondel et son équipe, très utilisée et sur laquelle je m'appuie abondamment. Elle évalue de nombreux découpages possibles en communautés en cherchant à maximiser la valeur de modularité [Blondel et al., 2008].

Distance : La distance entre deux sommets est la longueur du plus court chemin entre

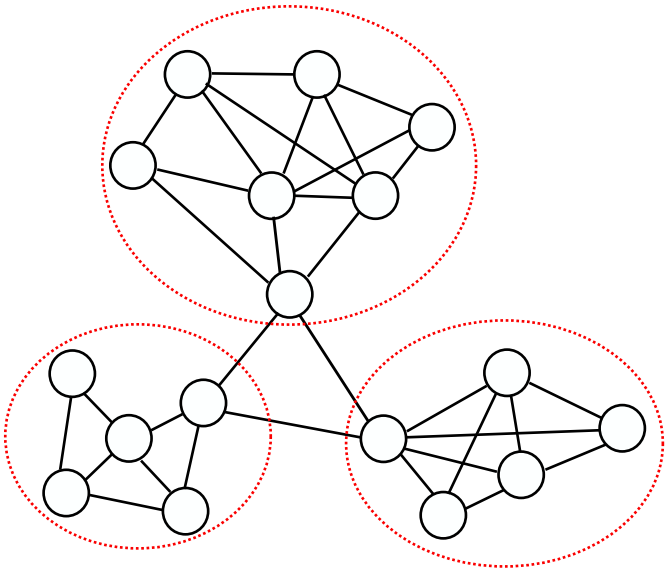

FIGURE 1.12 – Un réseau découpé en trois communautés de sommets

ces deux sommets. Deux sommets voisins sont à distance 1 l'un de l'autre, ce qui est le cas de tous les sommets de Königsberg à l'exception de n et s. Les plus courts chemins pour passer de n à s sont n–i–s et n–e–s qui sont donc à distance 2. Dans le cas de distance entre deux sommets appartenant à deux composantes connexes différentes, plusieurs conventions existent : considérer que la distance est $+\infty$, ou bien qu'elle est égale à 1 + la distance maximales entre deux sommets d'une même composante connexe. La distance entre deux sommets u et v d'un graphe est généralement notée $d(u, v)$, notation qu'on conservera dans ce manuscrit.

Excentricité : la valeur d'excentricité d'un sommet est la distance qui le sépare de son sommet le plus lointain.

Diamètre : Le diamètre d'un graphe est la plus grande distance entre deux de ses sommets, soit l'excentricité maximale d'un sommet du graphe. Le diamètre du graphe de Königsberg est donc de 2. On note le diamètre d'un graphe G, diam(G).

Centralité : La centralité des nœuds d'un graphe est, comme la communauté, une définition fluctuante mais néanmoins importante. Chaque mesure de centralité procure un score aux différents sommets qui se retrouvent ainsi être plus ou moins centraux, plus ou moins périphériques. Une mesure de centralité naïve est le degré : plus un sommet a un degré important et plus il est considéré comme central [Shaw, 1954]. Parmi les centralités les plus utilisées, on peut citer la *centralité de proximité* (closeness en anglais) introduite par Alex Bavelas et qui est inversement proportionelle à l'excentricité

de chaque sommet [Sabidussi, 1966] ou la *centralité d'intermédiarité* proposée par Linton Freeman qui traduit le fait qu'un sommet est un point de passage des plus courts chemins entre beaucoup d'autres nœuds du réseau [Freeman, 1977]. Le célèbre algorithme *PageRank* mis au point par Larry Page, le co-fondateur de Google, est également une mesure de centralité dans les graphes. Il attribue aux pages web un score déterminé par la probabilité d'y passer en naviguant aléatoirement, et donc au nombre de liens hypertextes y amenant.

1.3.2 Algorithmes et complexité

Revenons sur le diamètre. Il me semble être une bonne transition pour présenter le concept d'algorithme et de complexité algorithmique qui, s'ils ne sont pas centraux dans mon travail lui sont néanmoins sous-jacents et méritent d'être abordés.

Un exemple d'algorithme

S'il est relativement aisé de trouver à l'œil nu le diamètre d'un petit graphe comme celui des 7 ponts, la question devient rapidement plus complexe quand le nombre de sommets et d'arêtes du graphe augmente. Dans le cas du réseau du métro parisien, par exemple, il est difficile de deviner quelles sont les deux stations qui sont les plus éloignées ou quel est le chemin le plus court pour passer de l'une à l'autre. Comme pour beaucoup d'autres questions, la théorie des graphes a fait émerger de nombreux algorithmes pour répondre à celle-ci.

Un **algorithme** est une suite d'instructions qui permettent, à partir de n'importe quel objet d'un type donné, ici n'importe quel graphe, d'obtenir la réponse à une question spécifique. L'algorithme qui calcule le diamètre d'un graphe doit donc être capable de retourner la réponse quelque soit le graphe qui lui est donné en entrée. Un exemple d'algorithme bien connu est l'algorithme d'Euclide, qui permet de trouver le plus grand diviseur commun à deux nombres entiers positifs quelconques.

Comment alors calculer le diamètre d'un graphe ? On l'a vu, le diamètre est la distance la plus longue entre deux sommets du réseau. Calculer pour chaque sommet la distance qui le sépare de son sommet le plus éloigné permet donc, dans le cas où le graphe est connexe, ce qu'on va supposer ici, d'obtenir son diamètre.

On appelle **parcours en largeur**, souvent raccourci en *bfs* pour *breadth-first search*, le parcours à partir d'un sommet de départ quelconque, de l'ensemble des sommets qui sont dans la même composante connexe que lui. Ce parcours est dit en largeur, en opposition au parcours en profondeur, car les voisins du sommet de départ sont tous

visités en premiers, puis les voisins de ces voisins, etc. L'algorithme procède comme suit :

(1) On crée une file ne comportant que le sommet de départ.
(2) On retire de la file le sommet qui y est depuis le plus longtemps et on le marque.
(3) Parmi tous les voisins de ce sommet, on ajoute à la file ceux qui n'ont pas encore été marqués.
(4) Si la file n'est pas vide, on recommence l'étape (2)

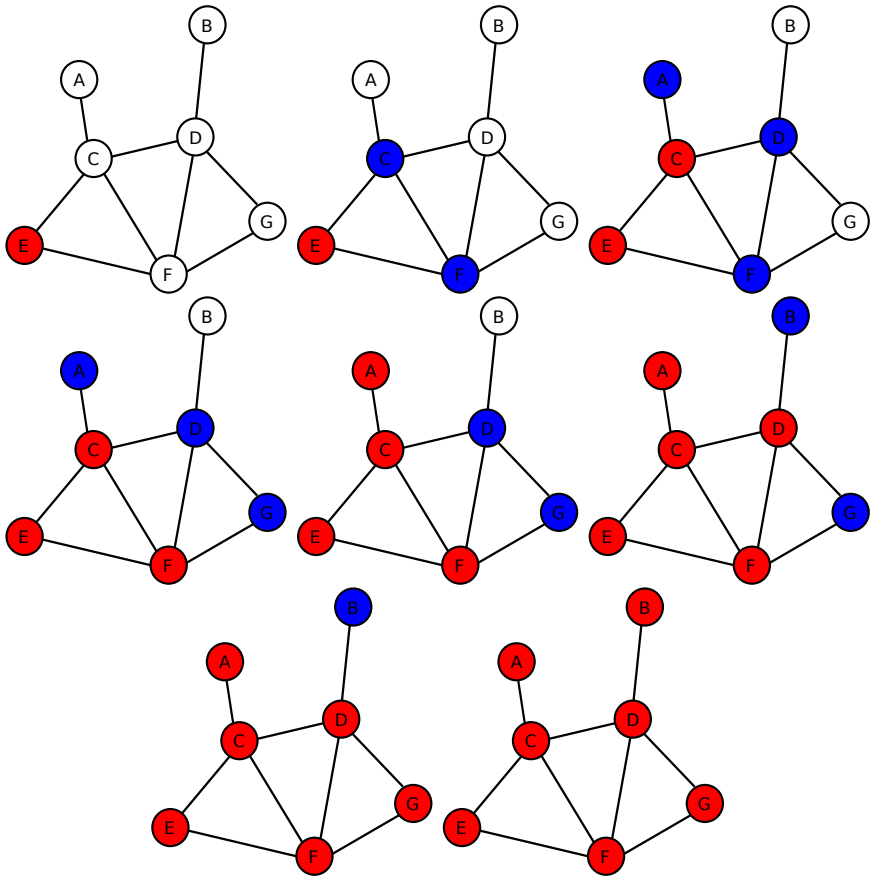

FIGURE 1.13 – Les étapes successives d'un parcours en profondeur à partir de E. En rouge, les sommets visités et en bleu ceux qui sont inclus dans la file.

Puisqu'on marque les sommets au fur et à mesure qu'on les rencontre, il n'est pas possible de parcourir plusieurs fois le même sommet durant le processus. Supposons maintenant qu'à chaque sommet qu'on ajoute à la file, on note quel est son sommet père, c'est-à-dire le sommet qu'on a retiré de la file à l'étape (2) et dont il est le voisin.

Dans cette configuration, le père d'un sommet est en fait le voisin de ce sommet qui est le plus proche du sommet de départ de l'algorithme. On peut, en faisant cela, construire en même temps qu'on parcourt le graphe une liste des distances de chaque sommet au sommet de départ, en considérant que cette distance vaut 1 + la distance entre le sommet de départ et son sommet père. Le sommet de départ étant à distance 0 de lui même et étant le père de tous ses voisins, l'algorithme leur attribuera une distance de 1, puis une distance de 2 à l'ensemble de leurs voisins, etc. Par exemple, dans la figure 1.13, le sommet F a pour père E dont il est à distance 1 et le sommet G a pour père F. Il est donc à distance $1 + 1 = 2$ de E. À la fin du processus on pourra donc connaître l'excentricité du sommet de départ, sa distance à son sommet le plus lointain.

En réitérant l'opération à partir de chaque sommet, on obtient donc l'excentricité de chacun d'eux. Le diamètre du graphe est la valeur maximale parmi ces résultats. Dans cet exemple, les sommets à la plus grande excentricité sont A, B, E et G qui ont une excentricité de 3, c'est donc le diamètre du graphe.

La complexité algorithmique

L'analyse de la complexité algorithmique est l'étude de la quantité de temps, ou d'espace mémoire nécessaires au fonctionnement d'un algorithme. Dans le cas de l'analyse de la complexité temporelle, plus étudiée, on s'intéresse au nombre d'étapes que l'algorithme va devoir exécuter en fonction de la taille de l'entrée. En théorie des graphes, les paramètres qui vont le plus souvent être pris en compte pour caractériser la taille de l'entrée sont le nombre de sommets n et le nombre d'arêtes m, mais certains algorithmes peuvent parfois avoir une complexité exprimée par des paramètres plus subtils comme le diamètre ou le degré maximal.

L'algorithme du calcul de l'excentricité d'un sommet, qui est ici adapté du parcours en largeur du graphe va répéter l'étape (2) autant de fois que le graphe possède de sommets, soit n fois. En effet, chacun des sommets va être visité et donc ajouté à la file exactement une fois. De plus, pour chacun d'entre eux, le procédé vérifie pour l'ensemble de ses voisins s'ils ont bien été marqués comme déjà visités ou pas lors de l'étape (3), ce qui signifie que toutes les arêtes vont également être parcourues, soit autant d'étapes supplémentaires que le nombre m de liens du graphe. Finalement, l'ordre de grandeur du nombre d'étapes que va effectuer l'algorithme est de $n + m$, ce qu'on note usuellement $O(n + m)$.

De plus, puisqu'on cherche le diamètre du réseau et non pas l'excentricité d'un seul de ses sommets, il faut reproduire l'opération pour chaque nœud afin de prendre la plus grande valeur obtenue comme diamètre. L'opération totale se fait ainsi en $O(n^2+nm)$.

La complexité algorithmique est en pratique directement liée au temps que met un algorithme à fournir une réponse. Certains de ceux qu'on présentera à partir du chapitre 4 sont d'ailleurs si complexes qu'ils n'ont pas pu être appliqués sur les plus grands réseaux à notre disposition.

1.3.3 Les graphes de terrain

Les études des algorithmes de graphes se basent souvent sur des graphes quelconques ou ayant à l'inverse des propriétés structurales très particulières et qui permettent alors de construire des méthodes *ad hoc* plus rapides. À l'inverse, les réseaux qu'on construit à partir de la modélisation d'un objet ou d'un ensemble d'objets observés, qu'on appelle également des graphes de terrain, ont généralement des caractéristiques communes particulières, plus ou moins similaires selon leur discipline d'origine, et sur lesquelles il est intéressant de s'arrêter.

Les réseaux, un objet transversal

Mise à part la sociologie, à laquelle est consacrée la section 1.4, de nombreuses disciplines scientifiques se sont, comme on l'a dit, emparées des réseaux et des graphes pour mettre en œuvre de nouvelles méthodologies de recherches. Un revue plus complète que celle qui suit peut être trouvée dans [Newman, 2010].

Parmi ces disciplines, c'est probablement la psychologie qui la première a mis en œuvre des analyses à partir de réseaux. En s'appuyant sur des petites structures relationnelles pour caractériser les relations entre quelques individus [Moreno et al., 1934, Bavelas, 1950, Shaw, 1954], les chercheurs impliqués ont rapidement ouvert la voie à des applications en sociologie.

En biologie, les réseaux sont régulièrement utilisés pour représenter les interactions de protéines entre elles [Vazquez et al., 2003, Sharan et al., 2005, Rual et al., 2005]. Pour ces réseaux, souvent orientés et parfois pondérés, les nœuds représentent les protéines tandis qu'un lien modélise l'inhibition ou la production d'une protéine par une autre. L'étude de la structure de ces réseaux peut permettre de prédire l'influence de mutations, ou de l'action d'un médicament sur l'organisme. Les réseaux d'interactions ne se limitent pas aux protéines mais sont également utilisés pour étudier les gènes [Hecker et al., 2009] ou les réseaux métaboliques, qui permettent de déterminer les propriétés des cellules [Jeong et al., 2000]. Les réseaux sont également propices à la représentation et à l'étude d'écosystèmes, et certains chercheurs proposent des définitions, en termes de structure de réseaux, de la prédation, de l'omnivorisme ou encore de la concurrence entre espèces [Paine, 1966, Pimm et al., 1991, Dunne et al., 2002]. Ces réseaux offrent

notamment la possibilité de prévoir l'impact de la disparition ou de l'intégration d'une espèce sur un écosystème par exemple. Les réseaux sont également utilisés pour observer des phénomènes de transmission de gènes, décisifs dans les processus d'évolution des espèces [Corel et al., 2016].

En épidémiologie, l'analyse structurale des réseaux permet de concentrer les efforts de vaccination ou de gestion des transports, dans la prévention de la diffusion de maladies. Les recherches se penchent aussi bien sur les relations entre individus [Bansal et al., 2007] que sur les liens entre des fermes d'élevage, par exemple [Eubank et al., 2004]. On imagine effectivement bien qu'un nœud à forte centralité d'intermédiarité favoriserait la dispersion d'une maladie s'il venait à être contaminé.

En histoire, les réseaux sont généralement utilisés de manière assez semblable à ce qu'on trouve en sociologie. Les historiens cherchent en effet à reconstruire des réseaux d'interaction entre individus à partir d'archives pour cerner les influenceurs ou les coutumes dans d'anciennes cultures [Lemercier, 2005]. On peut par exemple citer une étude sur le rôle central qu'ont les Médicis dans le réseau des familles florentines juste avant leur accession au pouvoir [Padgett and Ansell, 1993].

Les réseaux de sites internet, reliés les uns aux autres par des liens hypertextuels ont également profité de la modélisation en réseaux, et on verra d'ailleurs en section 4.4 qu'ils sont similaires aux réseaux sociaux. Une utilisation célèbre de tels réseaux est celle faite par l'algorithme de recherche de Google, PageRank, qui produit un score de centralité pour chaque site en effectuant des marches aléatoires à travers les pages internet, en fournissant ainsi un résultat satisfaisant aux recherches des internautes [Page et al., 1997].

La géographie est un domaine très approprié à la modélisation par les réseaux puisque le réseau est également un objet particulièrement adapté à la représentation de réseaux de transport, comme on l'a déjà vu dans le cas du métro parisien [Barthélemy, 2011]. Parmi les études mobilisant les réseaux, on retrouve ainsi des recherches appliquées aux réseaux ferrés [Jeong et al., 2007, Kreutzberger, 2008], routiers [Xie and Levinson, 2007], aériens [Guimera et al., 2005], viaires [Lagesse et al., 2016] et même multi-modaux [Janic, 2007] permettant de mettre en avant les villes centrales dans le transport de fret, de caractériser les stratégies de développement de ville ou encore d'analyser la structure des quartiers d'une agglomération.

Pour terminer cette énumération non exhaustive, on peut également mentionner des applications en chimie via la représentation de réactions sous la forme de réseaux dirigés, liants les réactifs aux produits [Graovac et al., 2012].

Des propriétés communes

La diversité des domaines d'utilisation des réseaux ont conduit, au crépuscule du vingtième siècle, à remarquer les similitudes spectaculaires qu'ils partagent [Watts and Strogatz, 1998, Barabási and Albert, 1999]. On appelle les graphes modélisant ces réseaux des *graphes de terrain* en référence aux terrains d'observation des chercheurs fournissant les données à partir desquelles ils sont construits. Dans ce cas, on est proche de réellement pouvoir confondre graphes et réseaux. En anglais on les nomme d'ailleurs *complex networks*, network signifiant réseau. Ils ont, depuis cette mise en lumière, donné lieu à la rédaction de nombreux ouvrages de référence [Newman, 2003, Boccaletti et al., 2006].

Les graphes de terrain ont généralement un diamètre faible, une densité faible et une distribution hétérogène des degrés de leurs sommets, et ce de façon indépendante de leur taille, ce qui est vrai pour un petit réseau l'étant aussi dans le cas de réseaux plus grands.

La propriété des graphes de terrain d'avoir un faible diamètre est connue sous le nom d'effet de *petit monde* en référence aux travaux, que je présenterai dans la section 1.4, du psychologue Stanley Milgram. Un graphe ayant la propriété de petit monde se trouve mécaniquement avoir un fort coefficient de clustering.

La faible densité des réseaux s'explique par des contraintes spécifiques mais néanmoins comparables selon les domaines. Dans le cas des réseaux d'interconnaissances entre individus, le fameux *nombre de Dunbar*, du nom de l'anthropologue l'ayant proposé, postule ainsi que les compétences cognitives liées à la sociabilité réduisent à environs 150 le nombre de relations stables qu'une personne peut entretenir simultanément [Dunbar, 1992]. De manière analogue, des contraintes géographiques, économiques, écologiques empêchent certainement des aéroports d'avoir trop de destinations possibles, des prédateurs d'avoir un très grand nombre de proies différentes ou des protéines d'avoir une influence trop importante sur l'organisme.

Concernant la distribution hétérogène des degrés, elle suit ce qu'on appelle une loi de puissance. Quelques sommets centraux qu'on trouve souvent sous l'appellation anglaise de hubs sont reliés à beaucoup d'autres qui sont beaucoup plus faiblement reliés entre eux et ont donc un degré moins important. Ces hubs sont par exemple les aéroports internationaux des réseaux de transport aérien ou les personnalités influentes d'un réseau de relations issu de Twitter tandis que la majorité des utilisateurs n'ont pas autant de contact qu'un Obama [Adamic et al., 2001, Stephen and Toubia, 2009]. Les réseaux respectant cette propriété sont qualifiés de *sans échelle*.

Génération aléatoire de graphes

Cette découverte a relancé la question de la génération de graphes aléatoires, d'abord initiée par les travaux de Paul Erdős et Alfréd Rényi, deux mathématiciens hongrois, à la fin des années 50 puis au cours des années 60. L'étude des graphes aléatoires permet d'évaluer l'efficacité d'algorithmes ainsi que de générer des réseaux semblables à des graphes de terrain, utiles pour la construction de métriques basées sur la comparaison entre un réseau observé et de tels réseaux générés pour l'occasion. On a déjà parlé de la modularité en section 1.3.1 et on verra en section 4.6.4 une autre méthode, proche de celle proposée dans ce manuscrit pour l'étude des réseaux, qui s'appuie également sur la comparaison avec des réseaux aléatoires. Cette section propose un bref aperçu des différents modèles de génération aléatoire qui ont jalonné la théorie des réseaux.

Le premier modèle historique est donc connu sous le nom de modèle d'Erdős–Rényi. Il produit des réseaux au sein desquels, pour chaque couple de sommets, celui-ci a une probabilité p d'être relié par une arête et donc $1-p$ de ne pas être relié. La densité du réseau augmente donc proportionnellement à la valeur de p [Erdos and Rényi, 1960]. La découverte des propriétés communes des graphes de terrain a par la suite poussé les chercheurs à revoir ce modèle, finalement peu satisfaisant, afin de l'adapter. Il ne respectait effectivement pas certaines propriétés comme l'hétérogénéité des degrés ou le coefficient de clustering élevé.

Le modèle proposé par Watts et Strogatz [Watts and Strogatz, 1998], respectivement sociologue australien et mathématicien américain (ce qui souligne encore la pluridisciplinarité des études du domaine), vise spécifiquement à produire des graphes respectant la propriété de petit monde. L'algorithme de génération commence par produire un graphe dans lequel chaque sommet est relié à un certain nombre de voisins donné en paramètre. Les voisins de chaque sommet sont sélectionnés de telle façon qu'un réseau circulaire régulier est obtenu, tel que celui à gauche de la figure 1.14. Après quoi, pour chaque arête, un de ses deux sommets est échangé avec un autre selon une certaine probabilité, elle aussi donnée en paramètre, qui a haute valeur rend le graphe totalement aléatoire mais qui permet d'obtenir, jusqu'à un certain point, un réseau respectant la propriété de petit monde comme l'illustre encore la figure 1.14.

Les réseaux générés par ce modèle ne respectent pas la propriété d'être sans échelle, leurs sommets étant de degrés homogènes. Deux chercheurs d'origine roumaine, Albert-László Barabási et Réka Albert, impliqués dans des recherches en physique et en biologie, vont proposer un modèle de génération de graphes sans échelle, le modèle Barabási-Albert dit d'attachement préférentiel [Barabási and Albert, 1999]. Ce modèle fonctionne par l'ajout successif de sommets à un réseau d'abord vide. Chaque nouveau sommet est relié à ceux précédemment ajoutés avec une probabilité qui dépend du nombre de voisins de ces sommets. Formellement, la probabilité qu'un nouveau

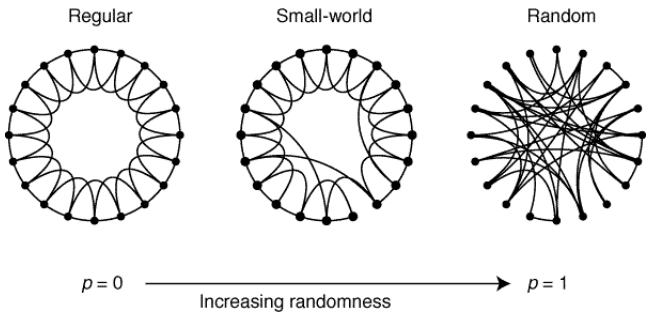

FIGURE 1.14 – Illustration issue de [Watts and Strogatz, 1998]. Trois réseaux dont le degré d'aléatoire diffère. Celui de gauche est régulier, celui de droite totalement aléatoire tandis que celui du centre, intermédiaire, présente les caractéristiques du petit monde.

sommet soit relié à un sommet s au moment où il est ajouté au graphe est de

$$p = \frac{d(s)}{\sum_{v \in V} d(v)}$$

où $d(x)$ est le degré du sommet x et V est la liste des sommets déjà contenus dans le graphe. La probabilité est donc plus grande d'être connecté à un sommet déjà parmi les plus connectés du réseau.

Comme déjà évoquées, de nombreuses métriques des réseaux utilisent d'une manière ou d'une autre les graphes aléatoires, et notamment pour les comparer à des graphes de terrain afin de caractériser ceux-ci. Dans ce genre de situations, le modèle aléatoire choisi est déterminant puisqu'il influe sur le résultat de l'indicateur. On verra dans la suite que les méthodes développées au long de ce doctorat s'affranchissent des réseaux aléatoires grâce à la quantité importante de données à notre disposition.

1.4 Les réseaux en sociologie

Si le sens vernaculaire de réseau social se rapporte aujourd'hui à des plateformes comme Twitter, Facebook et bien d'autres, d'échanges en ligne, l'origine de l'expression vient de la sociologie. Cette section explore les premières représentations de la sociabilité par des graphes ainsi que les évolutions successives qui, entre autres externalités, ont abouti à ce projet de thèse.

1.4.1 La question de la relation sociale dans la sociologie

La sociologie des réseaux sociaux, qui emprunte aux travaux de Georg Simmel postule que l'étude de la société passe autant (ou plus, ou uniquement, ou différemment, selon les écoles) par la prise en compte de la structure des relations qui existent entre les objets (individus ou groupes d'individus, généralement) que par leurs attributs propres.

Cette analyse structurale des réseaux sociaux se positionne pour certains entre les deux grandes traditions de la sociologie, l'individualisme et le holisme. Dans le premier cas, l'étude part de l'analyse des actions individuelles pour expliquer les phénomènes sociaux, tandis que dans l'autre, le holisme explique les actions individuelles par les contraintes sociales auxquelles chaque agent est exposé. Elle est ainsi régulièrement qualifiée de méso-analyse, alors située entre les niveaux micro et macro.

On comprend aisément que le réseau est l'objet idoine pour représenter les relations entre les composantes d'une structure sociale donnée et on retrouve leur première utilisation à cet effet dès 1934 dans les travaux du psychologue J. L. Moreno. Ce dernier étudiait alors les relations d'attirance et de répulsion entre élèves au sein de classes d'enfants en les modélisant sous la forme de réseaux, alors sous le nom de sociogrammes, dont il extrayait des formes représentatives (voir la figure 1.15) [Moreno et al., 1934]. On peut également interpréter cet effort comme une première approche de l'analyse des réseaux par leurs sous-structures, champ qui a connu par la suite de nombreuses variations, comme on le verra en section 4.4. La principale méthode que je déploie dans l'analyse des réseaux, à partir du chapitre 4, fait partie de cette famille.

Notons cependant prudemment que la sociologie des réseaux sociaux est avant tout l'étude de la sociabilité des personnes comme individus intégrés à une structure sus-jacente, contrainte par elle mais également agissant sur elle. Si de nombreux travaux s'appuient alors sur l'objet « réseau », les deux concepts ne doivent pas pour autant être assimilés l'un à l'autre. La fréquence des contacts d'un individu avec ses différentes relations sociales peut, par exemple, être interprétée dans le cadre de la sociologie des réseaux, sans qu'elle ne donne pour autant d'information sur ni n'utilise directement la structure relationnelle établie entre lesdits contacts.

1.4.2 La sociologie des réseaux sociaux

On cela a déjà été dit, la visualisation est souvent l'accès premier à l'analyse d'une structure relationnelle, autre nom couramment donné aux réseaux de sociabilité, mais elle est changeante et la difficulté de son interprétation augmente lorsque le nombre d'individus et de liens entre eux croît. Les sociologues des réseaux sociaux ont donc été amenés à proposer, comme je l'ai moi-même fait au cours de mon travail de doctorat,

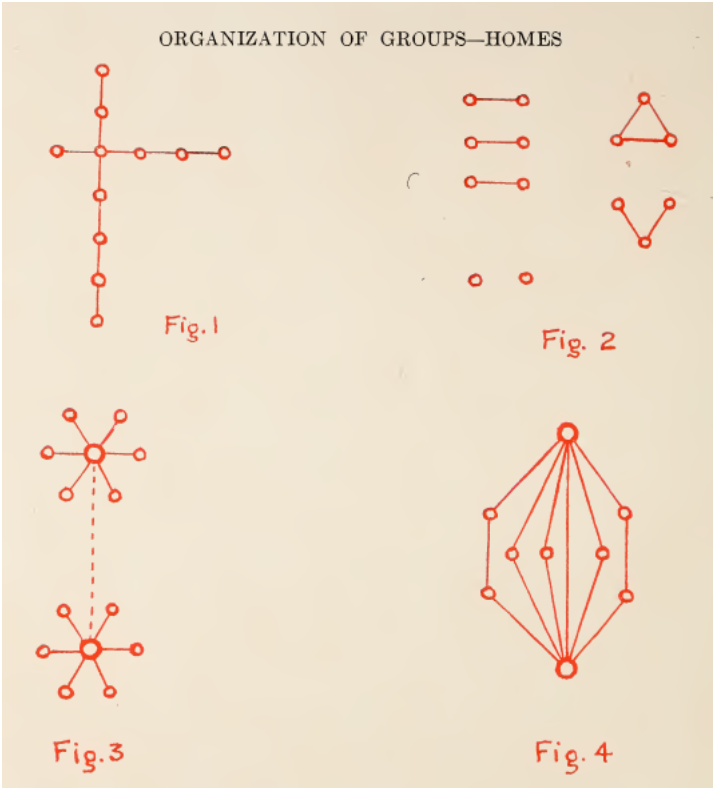

FIGURE 1.15 – Une figure de [Moreno et al., 1934]. Il s'intéressait déjà aux petites structures représentatives des réseaux

une myriade de méthodes, de métriques et d'indicateurs afin d'extraire l'information des réseaux qu'ils ont pu collecter au cours de leurs enquêtes. Je vais dans cette partie présenter quelques résultats parmi les plus emblématiques qui ont jalonné l'évolution d'une discipline qui a débuté, « *bien avant les outils et les concepts qu'elle a produits* » [Cristofoli, 2008] par l'utilisation de formes imagées de réseaux pour décrire des organisations (réseaux en étoile, réseaux circulaires, en Y, etc) avant, petit à petit, de construire lesdits outils d'analyse de grands réseaux de terrain. Plusieurs ouvrages de référence traitent de la sociologie des réseaux sociaux, en anglais comme en français. On peut notamment citer [Freeman, 2004, Mercklé, 2011, Lazega, 2014, Scott, 2017].

L'école d'anthropologie de Manchester, et notamment autour des étudiants de Max Gluckman, a été dans les années 50 pionnière dans les travaux de sociologie liés aux réseaux. John A. Barnes aurait ainsi été le premier à employer le terme de réseau social (selon [Wellman, 1997]) pour décrire son article relatant son étude des classes sociales d'une île norvégienne [Barnes, 1954]. Bien que l'expression de réseau social ait pu être rencontrée dans quelques publications antérieures, il semble faire consensus dans la communauté que c'est bien ici que se trouve la première occurrence de l'expression