Sociabilités en ligne, usages et réseaux

Raphaël Charbey

▶ **To cite this version:**

Raphaël Charbey. Sociabilités en ligne, usages et réseaux. Réseaux sociaux et d'information [cs.SI]. Télécom ParisTech, 2018. Français. NNT : 2018ENST0049 . tel-02180543

HAL Id: tel-02180543
https://pastel.hal.science/tel-02180543

Submitted on 11 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2018-ENST-0049

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Informatique et réseaux »

présentée et soutenue publiquement par

Raphaël CHARBEY

le 7 novembre 2018

Sociabilités en ligne, usages et réseaux

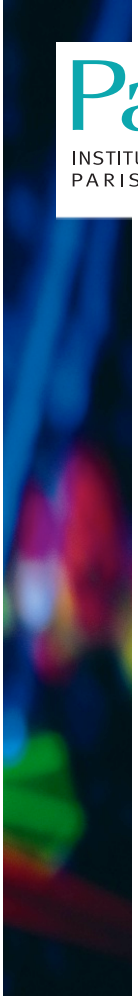
Directeur de thèse : **Christophe PRIEUR**
Co-encadrement de la thèse : **Antonio CASILLI**

Jury

Mme. Florence d'ALCHÉ-BUC, Professeure, Télécom ParisTech	Examinatrice
Mme. Claire BIDART, Directrice de recherche, Université Aix-Marseille	Examinatrice
M. Antonio CASILLI, Maître de conférence HDR, Télécom ParisTech	Co-directeur de thèse
M. Christophe PRIEUR, Maître de conférence HDR, Télécom ParisTech	Directeur de thèse
Mme. Nataša PRŽULJ, Professeure, University College of London	Rapportrice
M. Fabrice ROSSI, Professeur, Université Paris 1 Panthéon-Sorbonne	Rapporteur
M. Nicolas TROTIGNON, Directeur de recherche CNRS, ENS Lyon	Examinateur

THÈSE

TELECOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr



ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

Sociabilités en ligne, usages et réseaux

Raphaël CHARBEY

RESUME : Avec l'avènement du numérique, il est désormais possible aux chercheurs d'amasser des grandes quantités de données et les plateformes de réseaux sociaux en ligne ne font pas exception à cela. Les sociologues, comme d'autres, se sont emparés de ces nouvelles ressources afin de poursuivre leurs enquêtes sur les modalités de l'interaction entre individus et leur impact sur la structuration de la sociabilité.

Suivant cette voie, ce travail de thèse vise à l'analyse d'un grand nombre de comptes Facebook, aussi bien au travers des outils classiques de l'analyse de données que de la théorie des graphes, à laquelle des contributions méthodologiques sont apportées. Deux facteurs principaux encouragent l'étude de l'activité et de la sociabilité en ligne. D'une part, le temps important dédié à cette plateforme par de nombreux internautes justifie l'intérêt porté par les sociologues aux échanges qui s'y construisent. Par ailleurs, et contrairement à ce que l'on peut observer sur d'autre sites de réseaux sociaux en ligne, les liens entre individus sur Facebook sont proches de ceux hors-lignes.

Dans un premier temps, la thèse s'évertue à démêler les multiples facettes de ce à quoi "être sur Facebook" correspond. Distribués autour de pratiques normatives fabulées, les usages de nos enquêtés fluctuent au gré de leur appropriation ou non des composantes de l'importante variété de moyens de communication proposés par la plateforme. Ces usages, comme on le verra, sont ainsi différemment adoptés selon les catégories socio-professionnelles et influent par ailleurs sur les modalités d'échanges et d'interactions des enquêtés avec leurs amis en ligne. Ces modalités sont également explorées dans ce travail, tout comme le rôle du conjoint et sa place dans la structure relationnelle.

La seconde partie de la thèse se propose de construire une typologie de ces structures relationnelles dites égocentrées, c'est-à-dire depuis le point de vue de l'enquêté. Cette typologie des réseaux de sociabilité en ligne se base sur l'énumération de leurs sous-graphes induits, les *graphlets*, initialement développée par des chercheurs en bio-informatique. Cette approche offre une vision méso (entre micro et macro) des réseaux, propice à souligner des phénomènes inédits de sociologie des réseaux. À fort potentiel pluri-disciplinaire, la méthodologie *graphlets* elle même est également discutée et explorée.

MOTS-CLEFS : Sociologie des réseaux, théorie des graphes, analyse de données

THÈSE

ABSTRACT : With the digital advent, it is now possible for researchers to collect important amounts of data and online social network platforms are surely part of it. Sociologists, among others, seized those new resources to investigate over interaction modalities between individuals as well as their impact on the structure of sociability.

Following this lead, this thesis work aims at analyzing a large number of Facebook accounts, through data analysis and graph theory classical tools, and to bring methodological contributions. Two main factors encourage to study Facebook social activities. On one hand, the importance of time spent on this platform by many Internet users justifies by itself the sociologists interest. On the other, and contrarily to what we observe on other social network websites, ties between individuals are similar to the ones that appear offline.

First, the thesis proposes to detangle the multiple meanings that are behind the fact of "being on Facebook". The uses of our surveyed are not compacted in fantasized normative practices but vary depending on how they appropriate the different composers of the platform tools. These uses, as we will see it, do not concern all the socio-professional categories in the same way and they also influence how the respondents interact with their online friends. The manuscript also explores these interactions, as well as the lover role into the relational structure.

Second part of the thesis builds a typology of these relational structures. They are said as egocentred, which means that they are taken from the perspective of the respondent. This typology of social networks is based on their *graphlet* counts, that are the number of times each type of subnetwork appear in them. This approach offers a meso perspective (between micro and macro), that is propitious to underline some new social phenomena. With a high pluri-disciplinary potential, the *graphlet* methodology is also discussed and explored itself.

KEY-WORDS : Social networks, graph Theory, data analysis

Remerciements

C'est toujours avec gêne que j'ai reçu la compassion de ceux à qui il m'arrivait, au détour de rencontres, d'expliquer être en train de réaliser une thèse. Parfois masqué dans le regard ou l'intonation de la voix, mais plus souvent encore explicité de manière claire et sans atermoiement. Ce sentiment ne saurait être plus inapproprié tant les personnes qui m'ont accompagné de près ou de loin durant ces trois années ont su les rendre plus que chérissables à mes yeux. J'espère, en ces quelques lignes, leur témoigner une partie, au moins, de la gratitude que j'éprouve à leur égard.

Il serait probablement incongru de débuter ces remerciements autrement qu'en rendant hommage à mon directeur de thèse, Christophe Prieur, qui m'accompagne (à moins que ça ne soit l'inverse !) depuis plus de 4 ans. Merci pour tous tes encouragements ainsi que tes conseils qui, toujours prodigués dans une contagieuse bonne humeur ont, sans doute aucun, été les principaux moteurs de mes progrès en tant que chercheur.

Un grand merci également à Antonio Casilli, mon co-directeur, sans qui je serais à l'heure actuelle perdu dans les méandres de la rédaction d'une thèse en sociologie ! Ma propension à vouloir faire trop de choses en même temps nous a empêchés de plus collaborer ensemble et je m'en excuse. Merci également de m'avoir fait découvrir Burt, dont l'intérêt est (selon moi) aussi grand que son style est laborieux.

Merci aux membres du jury, qui ont accepté de prendre de leur temps pour s'intéresser à mon travail. J'ai déjà eu des échanges forts intéressants et extrêmement formateurs avec mes deux rapporteurs, Nataša Pržulj et Fabrice Rossi, ainsi qu'avec Claire Bidart (avec laquelle notre prometteuse collaboration, accompagnés notamment de Michel Grossetti, est toujours riche en rebondissements). Je suis certains que les remarques et questions de Florence d'Alché-Buc ainsi que de Nicolas Trotignon vont encore être source d'améliorations de la qualité de ma recherche et j'ai hâte de pouvoir échanger avec vous.

Mon aventure dans la recherche a débuté et s'est jusqu'à présent déroulée au sein de l'équipe Algopol, devenue Algodiv, dont j'ai eu plaisir à côtoyer et à apprendre auprès de tous les membres. Merci donc à Irène Bastard, Dominique Cardon, Jean-Philippe

Cointet, Baptiste Fontaine, Stéphane Raux, Camille Roth, et même à Joël Marchand !

J'ai passé au département de sciences économiques et sociales de Télécom ParisTech des moments tout aussi enthousiasmants sur le plan personnel que professionnel et j'en sais gré à l'ensemble des collègues qui m'y ont merveilleusement accueilli. Je repense parfois au soir où je postulais encore fébrilement aux offres de financement de thèse, Michaël Martial, le cadre administratif du laboratoire, a remué ciel et terre pour mettre la main sur la salvatrice enveloppe qui m'a permis d'arriver à la poste du Louvre avant sa fermeture et ainsi de pouvoir envoyer ma demande en temps et en heure (je n'avais pas obtenu le financement mais qu'importe ?). Corinne Chevalier et Marie-Josée Vatin complètent l'équipe administrative la plus à l'écoute et réactive qu'il m'ait été donné de croiser, je vous en suis très reconnaissant à tous les trois.

À ces trois indispensable co-laborantins, j'ajoute mille pensées à Florence Besnard et à Marianna Baziz qui, entre administration et formation, chapeautent la vie des doctorants de l'école. J'ai toujours trouvé l'aide que je venais chercher dans vos bureaux, ce qui vous a certainement coûté beaucoup de temps à toutes les deux. Merci donc encore.

L'accueil que j'ai pu recevoir au département SES de Télécom ParisTech a été le meilleur qu'il ait été imaginable. Les permanents du laboratoire ont tous participé à créer une atmosphère de travail propice aux belles rencontres. Je remercie néanmoins tout spécialement Caroline Rizza dont l'amitié m'a été très précieuse pendant ces trois années et Julien Morel pour ne pas m'avoir lâché la manche sur Ingress. Merci à tous les autres, bien entendu. Samuel pour les échanges de mangas, David d'avoir gardé son calme malgré tous mes passages dans son bureau, Sophie d'avoir essayé de m'apprendre la salsa, Dominique de toujours me faire un compliment sur mes chemises, Ulrich de m'avoir réinstallé mon linux (heureusement qu'il y a des économistes pour s'occuper des informaticiens dans ce labo ...), Valérie, Dana, Annie, Nicolas, Michael, Françoise, Christian, Stéphane et tous.

Merci également à Christian et Steven de m'avoir apporté mes doses quotidiennes de caféine et de lactose dans l'ambiance hip-hop de la cafet'.

Naturellement, j'ai assidûment fréquenté mes consœurs doctorantes et mes confrères doctorants (et les post-doc également, ne vous formalisez pas...) durant ces trois années rue Barrault et sans eux rien n'aurait été pareil. En premier lieu, merci à Antoine, mon voisin de bureau, en définitive la personne avec laquelle j'ai partagé le plus de temps ces trois dernières années. Merci à Constance d'avoir été presque jusqu'au bout la meilleure binôme de représentants et merci au reste des managers David, François, Giulia et Sabine, même si vous faites des recherches bizarres. Merci à Hélène-Marie d'avoir été la meilleure voisine du 18eme. Merci à Maude de m'avoir fait découvrir qu'il fallait être prudent avant de partir en post-doc au fin fond de la pampa. Merci à

Quentin pour le séminaire d'écriture, un grand moment sans aucun doute. Merci à Yann et Martin, Antoine et Mattias d'avoir été les meilleurs alliés, une fois 18h30 sonnée. Merci aux sociologues François, Gabriel, Jérémy, Marine, Thomas, Wilfried. Merci aux économistes Marc-iens Adrien, Ambre, Angela, Arrah, Jean-Marc et Vicente. Merci à la Team Catastrophe Florent, Sandrine et Robin d'avoir joyeusement accompagné la fin de ma rédaction. Merci aux ergonomes Angélique, Chloé, Jonathan et Mohini lorsque je suis arrivé, puis Anaïs, Anthony, Robin désormais Merci aux designers Émeline et Frédéric ainsi qu'à Dorian, Estelle, Justine, Laurent, Max et Samuel de prendre la relève. Pour terminer, merci à Ji-Yun pour les pauses cafés, et merci (et désolé...) à toutes celles et ceux que j'ai oublié.

J'ai eu l'occasion de visiter quelques autres départements pendant mon séjour à Télécom, ne serait-ce qu'en traversant l'école pour aller au restaurant. Merci à mes collègues de cours informaticiens, aux nombreux autres doctorants que j'ai pu croiser, et de manière générale à tous ceux qui ont fait que d'un point de vue égocentré (tout comme les réseaux que j'étudie, vous le verrez si d'aventure vous lisez plus que les remerciements), j'ai pu pleinement tirer profit de ces trois ans d'expérience.

Pour conclure cette interminable liste, j'aimerais terminer par tirer mon chapeau à tous mes ami-e-s et à ma famille qui m'ont toujours apporté un soutient indéfectible sans lequel le travail que j'ai eu la chance de réaliser n'aurait pas eu de sens pour moi. Merci, donc, d'avoir toujours été là pour moi, et ne croyez pas que ça s'arrêtera là !

Table des matières

Introduction **13**

1 Éléments contextuels **17**
1.1 L'analyse des réseaux - exemple du métro parisien 17
1.2 Le réseau social 21
1.3 Les réseaux et les graphes 24
1.3.1 Théorie des graphes 24
1.3.2 Algorithmes et complexité 33
1.3.3 Les graphes de terrain 36
1.4 Les réseaux en sociologie 40
1.4.1 La question de la relation sociale dans la sociologie 41
1.4.2 La sociologie des réseaux sociaux 41
1.4.3 Les réseaux personnels 49
1.5 Sociologie et plateformes de réseaux sociaux 51
1.5.1 Facebook, un mode de discussion parmi d'autres 52
1.5.2 Une sociabilité influencée par l'usage 52
1.5.3 Différences entre réseaux sociaux en ligne et hors ligne 53
1.6 Analyse de données 54
1.7 Informatique et sciences humaines 58

I Usages **61**

2 Algopol et usages **63**
2.1 L'enquête 63
2.1.1 Présentation de Facebook 64
2.1.2 Récupération des données 65
2.1.3 Recrutement et biais associés 67
2.1.4 Les métadonnées 69
2.2 Méthodes d'analyse 70

2.2.1 Accéder aux activités élémentaires . . . 70
2.2.2 La méthode du χ^2 . . . 71
2.3 Catégories socio-professionnelles . . . 72
2.4 Six profils d'usages de Facebook . . . 76
2.4.1 Publier chez soi . . . 77
2.4.2 Publier partout . . . 82
2.4.3 Qui sont les non-actifs ? . . . 86

3 Des usages intégrés dans des réseaux 89
3.1 Des réseaux particuliers . . . 89
3.1.1 Comparaison au panel de Caen . . . 90
3.1.2 Réseaux de commentateurs et likeurs . . . 91
3.2 Interroger les réseaux . . . 93
3.2.1 Clés de lecture des réactions du réseau aux publications . . . 93
3.2.2 Des réseaux qui confortent la variabilité des usages . . . 101
3.2.3 Vers une catégorisation des publications . . . 107
3.3 La nature du lien social . . . 108
3.3.1 Les alters qualifiés . . . 108
3.3.2 Des mesures structurelles et d'interactions difficilement conciliables . . . 109
3.3.3 La position des conjoints dans le réseau dépend de la nature de leur relation . . . 111
3.3.4 Cas des réseaux contenant un alter-ego . . . 115

II Réseaux 123

4 Graphlets 125
4.1 Définition . . . 126
4.2 Les positions au sein des graphlets . . . 130
4.2.1 Notions d'algèbre . . . 130
4.2.2 Positions et intérêts de la notion . . . 131
4.3 Limitations des graphlets . . . 132
4.4 Historique de la notion . . . 133
4.5 Considérations algorithmiques . . . 136
4.5.1 Parcours du réseau . . . 139
4.5.2 Détermination du graphlet . . . 142
4.5.3 Autres approches . . . 145
4.6 Méthodes d'interprétation . . . 145
4.6.1 Relative graphlet frequency . . . 146
4.6.2 Graphlet degree distribution agreement . . . 147

4.6.3 graphlet correlation distance . . . 149
4.6.4 NetDis . . . 150

5 Représentativité des graphlets 153
5.1 Les motivations derrière la représentativité . . . 153
5.1.1 Rapport aux graphes aléatoires . . . 153
5.1.2 Données uniformes . . . 154
5.1.3 Un indicateur simple . . . 155
5.2 Définition formelle . . . 155
5.3 Menaces à la validité . . . 160
5.3.1 Taille du corpus . . . 160
5.3.2 Outliers potentiels . . . 161
5.3.3 Réseaux hétérogènes . . . 161
5.4 Retour sur la représentation graphique . . . 161
5.5 Description générale du corpus . . . 162
5.6 Les groupes de graphlets . . . 164
5.6.1 Méthode d'analyse . . . 165
5.6.2 Les groupes des graphlets . . . 167
5.6.3 Deux niveaux de relations entre graphlets . . . 170
5.7 Les formes de réseaux personnels . . . 172
5.7.1 Le regroupement par les métriques de la littérature . . . 172
5.8 Les clusters de la représentativité . . . 174
5.8.1 Les clusters . . . 175
5.9 Quels individus pour quels réseaux ? . . . 183
5.9.1 Croisement avec les données socio-démographiques . . . 183
5.9.2 Graphlets et usages . . . 185

6 Approfondissements 189
6.1 Taille des graphlets . . . 189
6.1.1 Intuition : des groupes structuraux de graphlets . . . 189
6.1.2 Clustering depuis la taille 4 . . . 190
6.1.3 Similitude entre les clusterings . . . 192
6.1.4 Des clusterings aux intérêts complémentaires . . . 194
6.2 Positions et lien social . . . 194
6.3 Sélection du nombre de clusters du kMeans . . . 196

Conclusion 203

Introduction

Cette thèse d'informatique a été réalisée au sein du Département de Sciences Économiques et Sociales de Télécom ParisTech et vise à étudier la variabilité des utilisations faites de la plateforme de réseau social Facebook à travers une approche quantitative. Elle est organisée en deux parties. La première se concentre sur une différenciation des usages de la plateforme et sur l'analyse des fonctionnalités employées par chaque utilisateur, pour s'exprimer et échanger avec ses amis. La seconde partie est orientée vers l'étude des différentes structures relationnelles qui se construisent en ligne et à l'analyse d'une méthodologie, l'énumération des graphlets, que cette étude mobilise. On peut interpréter l'approche pluridisciplinaire de la thèse au travers du fait qu'y sont employées des méthodes informatiques pour analyser des données sociologiques, et que dans un même temps, ces données sont utilisées pour développer des outils informatiques.

Pour l'essentiel, ce travail de thèse repose sur les données collectées par le projet Algopol (pour politique des algorithmes), qui est un projet pluridisciplinaire porté par l'Agence Nationale de la Recherche (ANR) et par plusieurs établissements de recherche, universités ou centres de recherche. Il apporte une contribution à l'étude de l'influence des algorithmes sur les activités des internautes et à la construction d'une typologie des modes de partage de l'information en ligne. Dans le cadre de ce projet a ainsi été mis en place, en collaboration avec la Commission Nationale de l'Informatique et des Libertés (CNIL), une application Facebook qui a permis la collecte, anonymisée et avec le consentement explicite de ses répondants, des données d'usages et d'interactions de plus de 16 000 personnes. L'application, en fonction entre décembre 2013 et avril 2015, permettait alors à ses utilisateurs de visualiser leur réseau social personnel, ou égocentré, en échange de l'ensemble de leurs données d'usage : publications, commentaires de réponse, amis, etc.

Pour appréhender le concept de réseau personnel, il convient dans un premier temps de rappeler que les réseaux (ou graphes, en mathématiques et informatique fondamentales) sont des objets qui décrivent les relations entre des éléments. Ces éléments y sont généralement représentés par des points qui sont reliés deux à deux lorsqu'une relation

existe entre eux. Un réseau personnel est un réseau social, un réseau qui représente les interactions entre des individus, ici le fait qu'ils sont amis sur Facebook. La particularité du réseau personnel est que ses points représentent l'ensemble des amis (sur Facebook, toujours) d'une seule personne, sans que celle-ci soit elle-même représentée par un point du réseau.

À l'époque où ceux qu'on nomme communément les réseaux sociaux, Facebook, Twitter et d'autres, prennent une importance de plus en plus grandissante dans nos sociétés, il apparaît que ces plateformes offrent d'intéressantes perspectives d'exploration aux chercheurs en sciences sociales, et notamment aux sociologues des réseaux. Ces derniers étudient depuis maintenant plus de 60 ans les manières dont les relations ou les interconnaissances structurent la sociabilité de chaque individu, dans le cas d'une communauté insulaire [Barnes, 1954] ou bien en se focalisant sur l'indépendance des membres d'un couple [Bott, 1957]. Ces chercheurs postulent qu'un individu est justement plongé dans un réseau social sur lequel il influe et qui, réciproquement, a une influence sur lui.

Ce travail de doctorat s'articule autour de l'analyse de ce réseau social, aussi bien par l'étude des représentations mathématiques en points et liens que par celle d'indicateurs liés aux modes de conversation avec les membres de ce réseau. La représentation, sous la forme d'un réseau personnel, tel qu'on l'a décrite, permet ainsi d'élaborer des interprétations précises de la sociabilité de son propriétaire [Bidart et al., 2011]. En effet, ses relations qui proviennent de la même sphère sociale, ses collègues par exemple, vont très certainement être connectés entre eux. Le conjoint ou la conjointe de l'enquêté aura de fortes chances d'avoir déjà rencontré certains de ses amis d'enfance, quelques-uns de ses collègues, ainsi que les membres de sa famille, et sera donc connecté-e à beaucoup de points du réseau qui ne sont pas eux-même connectés entre eux [Freeman, 1977, Backstrom and Kleinberg, 2014], puisqu'il est plus rare de présenter ses amis d'enfance à ses collègues de travail. Mais dans ce cas, puisque tous les collègues de travail ne sont pas connectés au conjoint, on peut imaginer que ceux qui l'ont rencontré sont en fait ceux qui sont les plus proches de l'enquêté. Les possibilités sont nombreuses.

Depuis les premiers pionniers de l'analyse des réseaux sociaux qui, stylo en main, partaient mener des entretiens pour dessiner les réseaux personnels de leurs enquêtés, l'analyse de ces objets s'est structurée et les méthodes employées se sont diversifiées. Des sociologues ont continué à proposer de nouvelles mesures, toujours plus précises [Freeman, 1977], leur permettant d'interroger les réseaux sur certaines de ces questions de sociabilité, sur la structuration des relations dans des monastères [White et al., 1976] ou encore sur les perspectives de promotions et de primes dans des entreprises [Burt, 1992].

En parallèle de ce travail, mais avec une temporalité différente, la théorie des graphes,

la branche des mathématiques et de l'informatique spécialisée dans l'étude des réseaux, a permis d'en analyser de plus en plus et qui soient de plus en plus grands, par l'amélioration continue de l'efficacité des algorithmes employés à la production de ces résultats [Brandes, 2001, Hočevar and Demšar, 2014] et au développement de modèles de génération aléatoire de réseaux dont la diversité permet d'envisager de fortes similitudes avec des réseaux réels de différents types [Erdos and Rényi, 1960, Watts and Strogatz, 1998, Barabási and Albert, 1999]

Ces progrès sont les bienvenues puisque les traces d'usages des internautes sont désormais omniprésentes sur la toile, et qu'il est possible d'accéder en ligne à des milliers de réseaux sociaux. Si ces derniers sont moins précis que ceux construits à partir de longs entretiens qui permettent à l'enquêteur d'avoir une connaissance fine de ses membres et de leurs interactions, ces réseaux construits à partir de listes d'amis Facebook donnent en revanche une place plus visible aux liens faibles [Granovetter, 1977] et sont également moins dépendants de la perception de l'enquêté sur les relations entre ses amis [Rivière, 2000]. Ces réseaux qui sont plus nombreux et plus généralement grands sont donc difficiles à analyser sans passer par l'intermédiaire d'algorithmes efficaces.

Toutes ces avancées de la science des réseaux, appliquée d'un côté et fondamentale de l'autre, en ont petit à petit fait un outil commun à beaucoup de domaines de recherche. On peut par exemple citer la biologie, qui les emploie aussi bien pour représenter des interactions entre protéines [Vazquez et al., 2003, Sharan et al., 2005] que des chaînes trophiques [Paine, 1966], l'épidémiologie [Bansal et al., 2007, Eubank et al., 2004], l'histoire [Lemercier, 2005], la géographie [Lagesse et al., 2016], et bien d'autres.

Cette multiplication des champs d'application des réseaux confère naturellement à ce travail de thèse des potentialités d'utilisations au delà des frontières de la sociologie ou de l'informatique. L'énumération des graphlets, la méthode d'analyse des réseaux explorée dans ce manuscrit, est d'ailleurs majoritairement utilisée dans le cadre de la biologie [Pržulj et al., 2004, Pržulj et al., 2006] et l'article scientifique qui l'a popularisée insistait justement sur la possibilité de l'appliquer à des réseaux décrivant des jeux de données issus de domaines variés [Milo et al., 2002]. Malgré cela, les graphlets n'ont jusqu'à présent que peu été utilisés en sociologie [Cunningham et al., 2013]. Une nouvelle méthode d'interprétation du résultat de l'énumération des graphlets, la *représentatitivité des graphlets*, est donc proposée dans cette thèse. Elle est dans un premier temps utilisée pour mieux comprendre les graphlets et les relations qui existent entre eux, ce qui peut être utile à n'importe quel chercheur désireux de travailler avec ces outils. Une typologie des réseaux personnels de Facebook est ensuite construite en employant la même méthode qui en exhibe cinq familles. On verra que celles-ci offrent de nouvelles accroches à des interprétations en termes sociologiques, auxquelles des pistes d'études sont proposées.

Outre les perspectives de collectes massives de données à des buts de recherche, quand il ne s'agit pas de ciblage publicitaire ou politique, l'activité des usagers des plateformes de réseaux sociaux intéresse également la sociologie. À l'instar de la plupart des nouveaux dispositifs de communication et de télécommunication qui apparaissent, comme le téléphone [Licoppe and Smoreda, 2000, Stoica et al., 2013] ou les blogs [Herring et al., 2005], les sociologues se sont en effet emparés du phénomène Facebook pour essayer de comprendre ses enjeux. Si les auteurs réfutent assez largement l'idée selon laquelle les interactions qui s'y déroulent seraient déconnectées de la réalité [Ellison et al., 2007, Beaudouin, 2009, Dagiral and Martin, 2017], ils pointent au contraire l'influence réciproque qu'ont cette sociabilité en ligne et la sociabilité tangible [Cardon, 2010]. Un autre des thèmes récurrents de la recherche sur les plateformes de réseaux sociaux en ligne est la pluralité des usages [Ellison et al., 2007, Burke et al., 2011] et avec elle la difficulté de proposer des interprétations généralisées.

C'est dans cette optique qu'après une exploration plus précise de l'état de la recherche dans ces différents domaines proposée dans le chapitre 1, la partie I, qui englobe les chapitres 2 et 3, est dédiée à l'analyse des usages de la plateforme.

Le chapitre 2 décrit l'enquête Algopol, les méthodes choisies pour l'analyse des métadonnées et les six catégories d'usage, regroupées en trois familles, qu'on retrouve parmi les utilisateurs de Facebook. Dans le chapitre 3 et en repartant de ces catégories, ainsi que des catégories socio-professionnelles de nos enquêtés, je présente les différentes modalités d'interaction avec le réseau, qui selon les utilisateurs vont de l'inexistence à l'omniprésence.

La partie II opère une plongée dans les réseaux personnels et débute, avec le chapitre 4 par une présentation et un historique de l'énumération des graphlets. Ces derniers sont ensuite mobilisés dans le chapitre 5 au cours duquel je présente la représentativité des graphlets, les relations entre ceux-ci, et les cinq familles de réseaux qu'elle découpe. Finalement, le chapitre 6 conclut par plusieurs travaux d'ouvertures, en présentant des recherches sur la taille des graphlets, qui est déterminante pour le temps de calcul et pour leur interprétation, la position des amis de nos enquêtés dans leurs réseaux personnels, qui permet de caractériser leur lien social.

Chapitre 1

Éléments contextuels

Ce travail de thèse, inscrit en informatique, présente la spécificité d'avoir été réalisé au sein d'un laboratoire de recherche en sciences humaines et sociales. Si l'approche pluridisciplinaire apporte maintes opportunités en termes de collaborations et d'ouvertures, elle induit également un champ élargi de notions à aborder. Ces notions, à commencer par celle de réseaux, centrale dans ce travail, méritent une description précise qui permette à chaque lecteur, quelle que soit sa ou ses disciplines de prédilection, de pouvoir se les approprier aisément. Ce premier chapitre ambitionne donc de dresser un panorama, sinon complet, tout au moins suffisant, du contexte général de la théorie et de la sociologie des réseaux.

1.1 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.

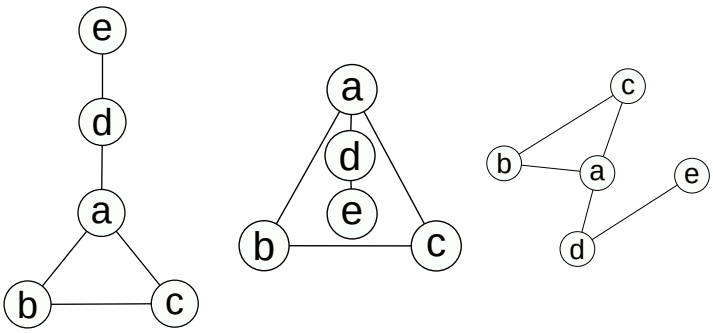


FIGURE 1.1 – Plusieurs visualisations possibles et toutes structuralement équivalentes du réseau [*a–b a–c a–d b–c d–e*].

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.

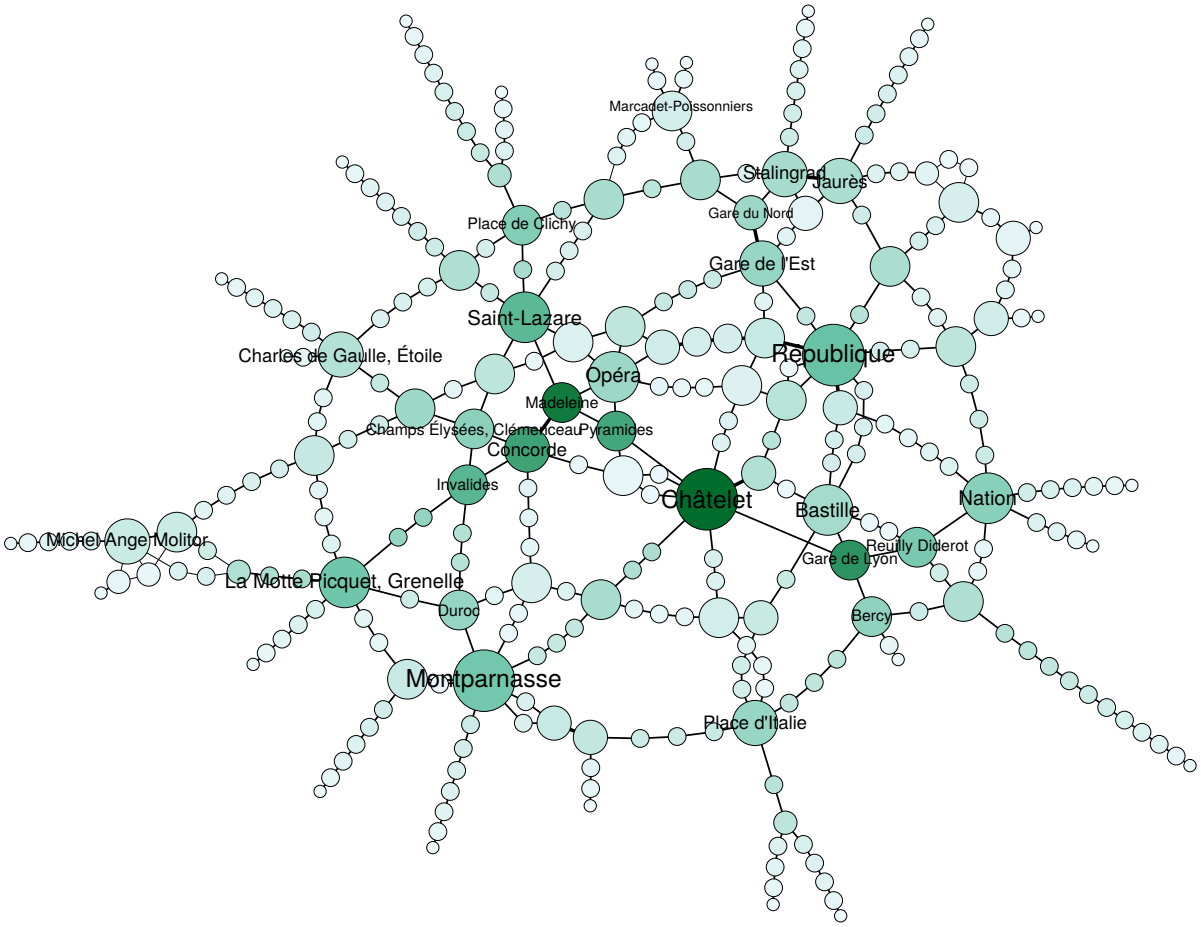


FIGURE 1.2 – Un réseau représentant les stations du métro parisien. Le réseau, comme les autres présentés dans ce manuscrit, est exporté depuis le logiciel de visualisation et de manipulation de graphes Gephi.

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

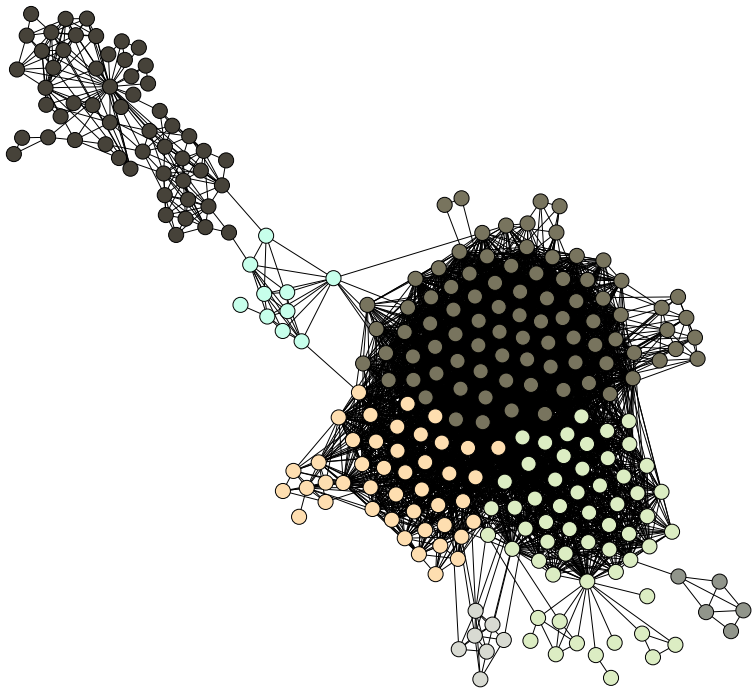


FIGURE 1.3 – Un réseau social

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

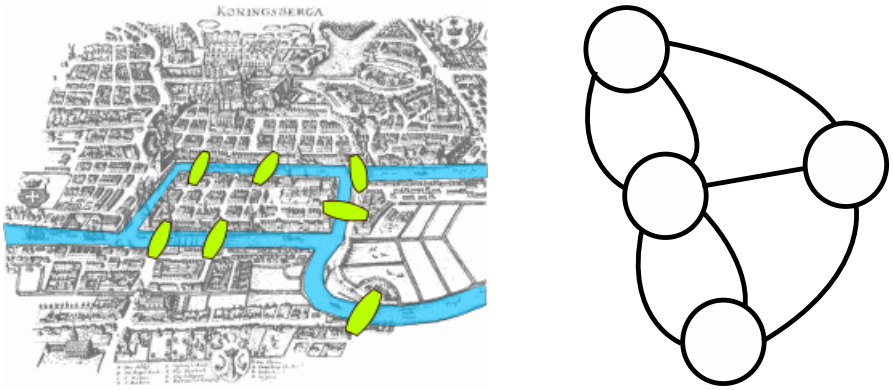


FIGURE 1.4 – Le graphe représentant les ponts de Königsberg

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.

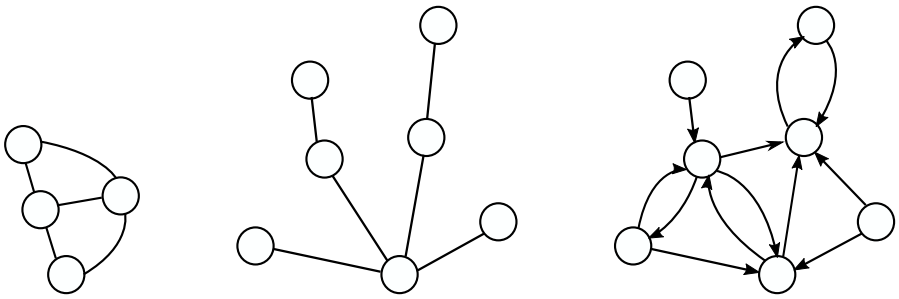


FIGURE 1.5 – De gauche à droite : le graphe de Konigsberg simplifié, un arbre et un graphe orienté. Un graphe orienté simple, comme celui-ci, n'a pas plus d'une arête d'un sens donné entre deux sommets.

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.

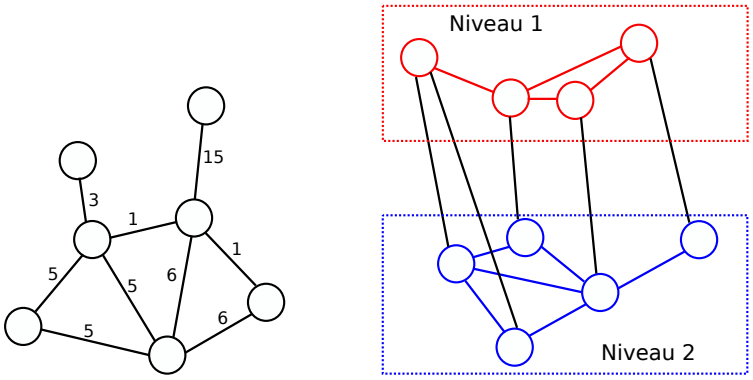


FIGURE 1.6 – De gauche à droite : un graphe pondéré et un graphe multi-niveaux, dont on voit les liens entre les éléments de même niveau et entre les éléments de niveaux distincts.

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.

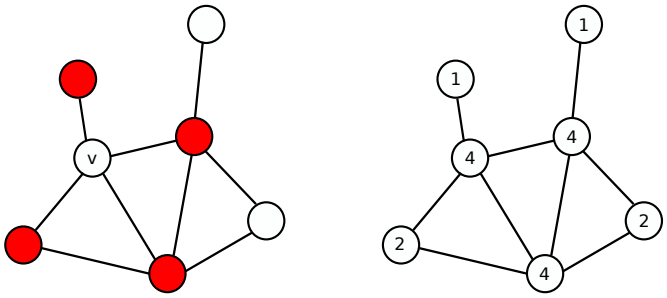


FIGURE 1.7 – À gauche, un réseau dans lequel le voisinage du sommet v est indiqué en rouge. À droite, le degré de chaque sommet. La distribution des degrés triée du graphe est donc 1,1,2,2,4,4,4

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.

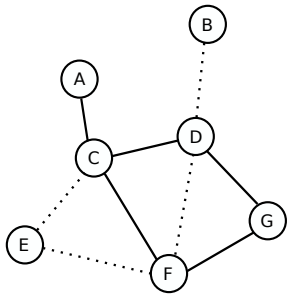


FIGURE 1.8 – Les arêtes pleines correspondent à celles qui sont sur l'un de deux plus courts chemins A-C-F-G et A-C-D-G entre A et G. Ils sont de longueur 3

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.

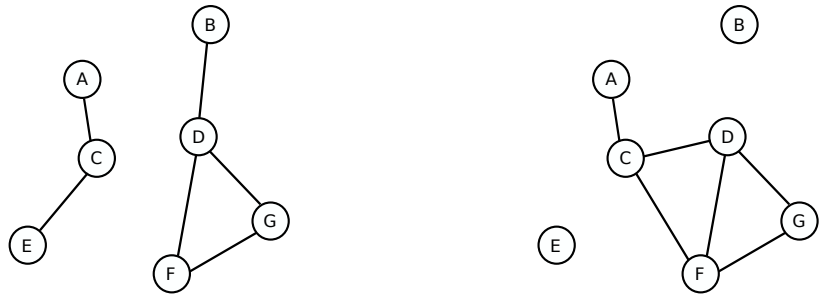


FIGURE 1.9 – À gauche un graphe avec deux composantes connexes ACE et BDFG et à droite un graphe à trois composantes connexes où E et B sont des sommets isolés.

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.

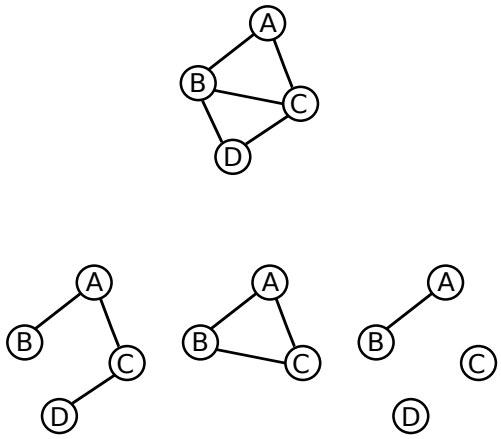


FIGURE 1.10 – Un graphe et trois de ses sous-graphes parmi la multitude de possibles.

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.

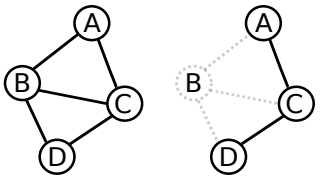


FIGURE 1.11 – Un graphe et son unique sous-graphe induit à A, C et D.

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

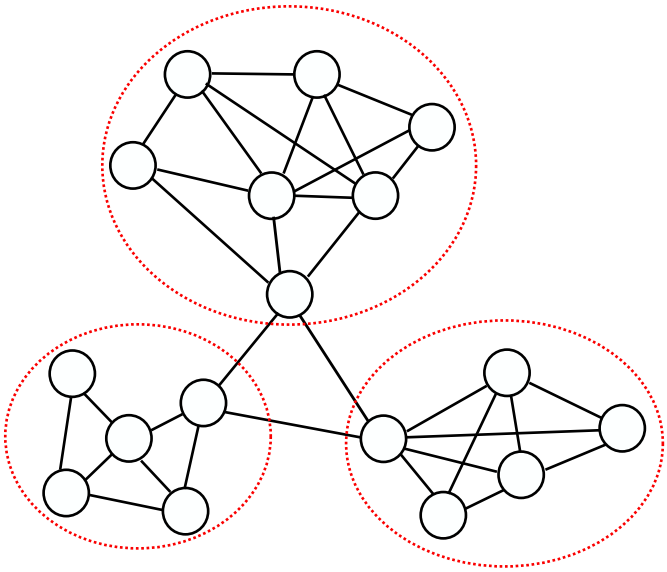


FIGURE 1.12 – Un réseau découpé en trois communautés de sommets

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)

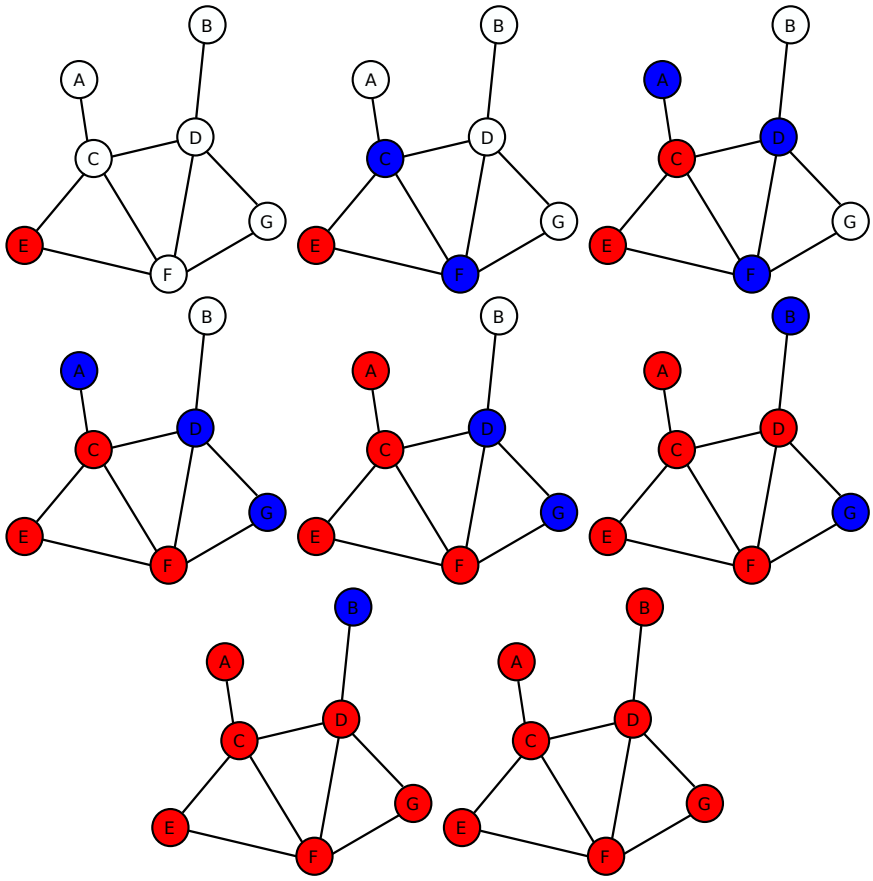


FIGURE 1.13 – Les étapes successives d'un parcours en profondeur à partir de E. En rouge, les sommets visités et en bleu ceux qui sont inclus dans la file.

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

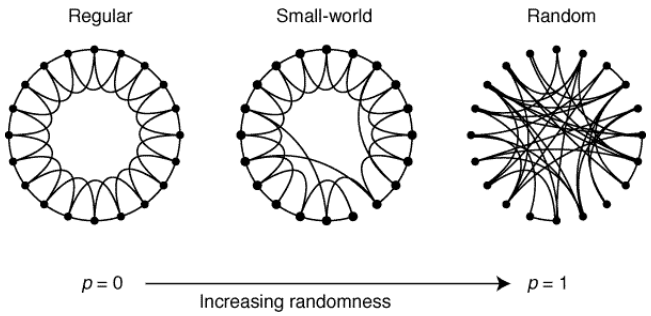


FIGURE 1.14 – Illustration issue de [Watts and Strogatz, 1998]. Trois réseaux dont le degré d'aléatoire diffère. Celui de gauche est régulier, celui de droite totalement aléatoire tandis que celui du centre, intermédiaire, présente les caractéristiques du petit monde.

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,

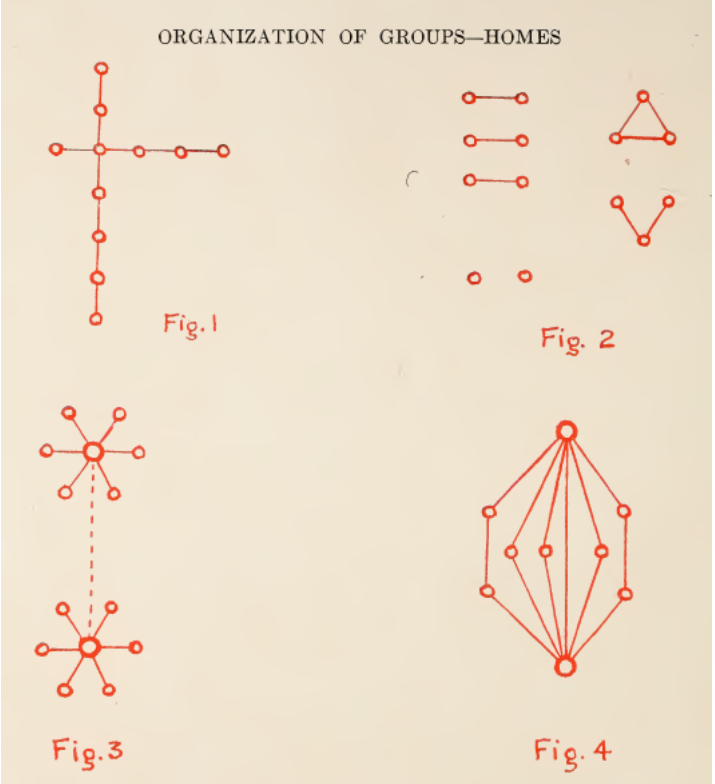


FIGURE 1.15 – Une figure de [Moreno et al., 1934]. Il s'intéressait déjà aux petites structures représentatives des réseaux

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

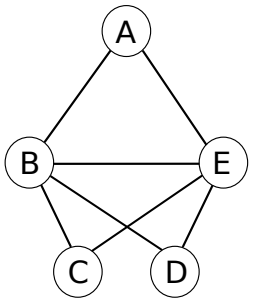
dans son sens actuel. Au long de son observation, Barnes remarque que la structuration sociale repose sur l'entrelacement de ce qu'il appelle des sphères sociales, qu'elles soient géographiques, économiques, ..., et que chacun des habitants de l'île appartient en fait simultanément à plusieurs de ces sphères. Sans directement préfigurer des suites, mêlant visualisation et analyse mathématique, qui seront données à l'étude des réseaux sociaux, l'article de Barnes a ouvert la voie aux nombreux sociologues pour qui les réseaux sont devenus un outil d'analyse, et en premier lieu à l'une de ses collègues.

Elizabeth Bott, également membre de l'école d'anthropologie de Manchester et autre précurseur de la sociologie des réseaux sociaux, a introduit la notion de *densité* du réseau parmi les indicateurs classiques. Dans une étude basée sur l'observation de 20 ménages londoniens, elle propose que la densité du réseau des relations de la famille est liée à la répartition des rôles au sein du couple. Pour elle, un couple qui aurait déménagé présenterait un réseau de relations moins dense qu'un couple ayant toujours vécu au même endroit. Dans ce cas, la tendance à rechercher plus de soutien auprès de leur partenaire ou encore à cultiver des relations communes encouragerait les deux époux à un plus équitable partage des tâches domestiques. À l'inverse, elle observe une plus importante et plus traditionnelle distinction des assignations ménagères chez les couples intégrés à des réseaux plus anciens et plus denses [Bott, 1957].

Signe de la montée en puissance de l'analyse des réseaux sociaux, en 1963, James Davis compile 56 questions sociologiques auprès d'une variété d'auteurs notables de la sociologie et qu'il juge être les questions phares du domaine à cette époque et les réinterprète selon les outils de la théorie des graphes [Davis, 1963].

Harrison White, alors à l'université de Harvard, est un chercheur qui a initialement étudié la physique théorique et qui va développer une méthode de représentation des réseaux qui fera date. Il propose qu'un réseau soit représenté sous la forme d'une matrice carrée au sein de laquelle chaque sommet est assigné à une ligne et à une colonne. Pour un réseau à n sommets, on a donc une matrice correspondante de taille $n \times n$. Les cases de cette matrice représentent alors les liens entre les sommets, chacune valant 0 ou 1 dans le contexte des réseaux simples. Une case notée 0 signifie qu'il n'y a pas d'arête entre les deux sommets à l'intersection desquels elle se trouve et 1 qu'il y a une arête. La figure 1.16 illustre un réseau et sa matrice associée.

Un exemple d'application de cette représentation est la recherche de sommets aux mêmes positions, correspondant ainsi aux mêmes rôles sociaux. En réorganisant la matrice de telle façon que les lignes et les colonnes similaires soient consécutives les unes aux autres (comme illustré dans la Figure 1.17), White obtient des groupes de sommets qui sont reliés au reste du réseau de la même manière. C'est ce qu'on appelle le *blockmodeling*, un terme qui est resté en français. On dit de sommets qui sont dans le même groupe qu'ils sont *structuralement équivalents*.



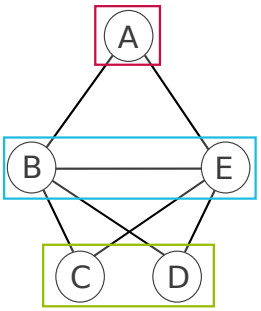
	A	B	C	D	E
A		1	0	0	1
B	1		1	1	1
C	0	1		0	1
D	0	1	0		1
E	1	1	1	1	

FIGURE 1.16 – À gauche, un réseau et à droite, la matrice correspondante dans le système de White. Notez bien qu'exactement les mêmes informations sont contenues dans chacune de ces deux représentations.

Pour bien comprendre ce que sont deux sommets structuralement équivalents, il faut imaginer qu'on supprime les labels des nœuds d'un réseau et qu'on modifie sa visualisation, comme on l'a fait dans la Figure 1.1. Il devient alors impossible de relabéliser les sommets à l'exception du sommet A qui n'est structuralement équivalent à aucun autre. Les sommets B et E, par exemple, deviendraient indiscernables l'un de l'autre, comme le montre la figure 1.18. Dans le cas du réseau du métro parisien, les quatre sommets des deux branches au nord de Marcadet-Poissonniers sont deux à deux structuralement équivalents, ce qui explique qu'on ne puisse pas déterminer quelle branche correspond à quel bout de ligne.

Selon White, les classes d'individus qui émergent par la méthode du blockmodeling sont les seules pertinentes et explicatives, à l'inverse des classes sociales usuelles de la sociologie et auxquelles les individus sont assignés *a priori* dans les études. White utilisera notamment cette modélisation dans le cadre de son étude des relations d'amitié et d'inimitié entre les moines d'un monastère américain. Il y montre, alors qu'un conflit avait éclaté au sein du monastère, que les groupes de moines qui avaient fini par le quitter, en avaient été écartés ou y étaient restés correspondaient aux groupes obtenus par l'équivalence structurale [White et al., 1976].

Il n'est cependant pas possible d'appliquer la méthode de l'analyse structurale à des réseaux dès lors que leur nombre de sommets augmente. Il n'existe en effet généralement pas, dans ce cas, de groupes de sommets qui aient exactement les mêmes groupes de voisins, autrement que localement, à l'image des bouts de ligne au nord du plan de métro. Plusieurs versions moins contraignantes de la méthode ont été proposées au fil des études et on aura l'occasion de revenir sur l'une d'entre elles, abordée dans la partie 4.5.2, sur laquelle s'appuient en partie les méthodes que nous avons appliquées à nos réseaux.



	A	B	E	C	D
A		1	1	0	0
B	1		1	1	1
E	1	1		1	1
C	0	1	1		0
D	0	1	1	0	

FIGURE 1.17 – Les sommets ont été regroupés en trois ensembles : [A], [B,E] et [C,D]. On remarque qu'à l'exception des intersections entre sommets du même groupe, ceux-ci ont exactement les mêmes lignes et les mêmes colonnes. En fait, A est connecté à tous les membres du groupe [B,E] qui sont en plus connectés aux membres du groupe [C,D]. B et E sont structuralement équivalents entre eux, tout comme C et D entre eux. Aucun sommet n'est par contre structuralement équivalent à A.

Parmi les travaux en sociologie des réseaux, celui qui est peut-être le plus emblématique, et en tous cas le plus connu au delà de la communauté des chercheurs, est l'œuvre de Stanley Milgram. Outre ses études comportementales, le psychologue a mené une enquête déjà évoquée dans la section 1.3.3, qui a permis de faire émerger le concept de petit monde. Son étude a en effet mis en lumière le faible diamètre des réseaux sociaux.

En 1967, il a distribué des lettres à des habitants de deux villes américaines en donnant à ses enquêtés la consigne de faire parvenir cette lettre à une personne-cible, la même pour tous, résidant dans une autre ville du pays, éloignée du point de départ. Pour ce faire, chacun devait remettre ou envoyer la lettre à une personne qu'il connaissait personnellement. Cette personne devait alors noter son nom sur la lettre et réitérer l'opération en envoyant la lettre à quelqu'un d'autre qu'elle jugerait être plus proche de la cible [Travers and Milgram, 1967].

Il a alors été possible de retracer la petite minorité des chaînes ayant atteint leur destination et c'est en remarquant que ces chaînes sont courtes, quelques intermédiaires à peine, que Milgram a pu conclure que les réseaux sociaux ont un faible diamètre. On parle encore des fameux 6 degrés de séparation moyens qu'il y a entre deux personnes. Cette valeur n'a d'ailleurs cessé de diminuer avec les connexions toujours plus nombreuses opérées par les plateformes de réseau social [Kwak et al., 2010, Backstrom et al., 2012].

Mark Granovetter, un sociologue américain, professeur à Stanford et ancien étudiant

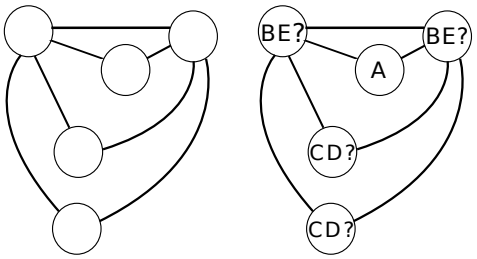


FIGURE 1.18 – Impossible de déterminer les labels des sommets structuralement équivalents dans cette autre visualisation possible du réseau précédent.

de Harrison White, va introduire en 1973 un autre concept important de la sociologie des réseaux avec la notion de force des liens entre deux personnes. Il la définit en fonction d'une combinaison de facteurs comme le temps passé ensemble ou l'intensité émotionnelle de la relation. Il postule que les liens forts sont nécessairement circonscrits à l'intérieur de *cliques* ou au moins à l'intérieur d'une communauté dense, une clique étant un réseau ou une partie d'un réseau dans lequel tous les sommets sont reliés entre eux, c'est à dire où tout le monde se connaît. Le corollaire de ce postulat est que les liens forts ne créeraient donc jamais de pont entre les communautés, ce qui s'explique comme suit.

Imaginons que deux individus, disons Harrison et Elizabeth, aient un lien fort. La probabilité est alors grande pour qu'Elizabeth ait présenté certains de ses amis à Harrison, ou qu'inversement, elle ait rencontré des proches de lui. Dans tous les cas, le nombre d'arêtes entre le groupe d'Elizabeth et celui d'Harrison augmente et la relation entre eux deux ne peut ainsi plus former un pont entre ces deux communautés puisqu'elles sont reliées par plusieurs arêtes comme l'illustre la Figure 1.19.

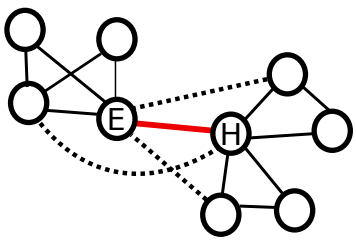


FIGURE 1.19 – Dans le réseau, Elizabeth (E) et Harrison (H) ont un lien fort. Au fil du temps, il est très probable que des liens, comme ceux en pointillés, apparaissent entre Elizabeth et des amis de Harisson ou l'inverse. Si ces liens existent, celui entre E et H ne forme plus un pont entre deux groupes.

Notant qu'Elizabeth aurait préférentiellement présenté à Harrison ceux de ses amis avec qui elle partage également une relation forte, Granovetter nomme *triade interdite* la triade, c'est-à-dire le réseau composé de trois individus, parmi lesquels l'un entretiendrait un lien fort avec chacun des deux autres tandis que ceux-ci n'auraient aucun lien entre eux.

Cette hypothèse a par la suite été étayée et validée par de nombreuses études empiriques réalisées par Granovetter lui même concernant la recherche d'emplois [Granovetter, 1974] ou bien par d'autres ([Friedkin, 1980] par exemple). Les chercheurs étudiant un réseau social cherchent alors à accéder à la qualification de la force des liens le composant, le plus souvent à l'aide de questionnaires qu'ils font passer aux enquêtés. On parle depuis de la *force des liens faibles*, l'existence de ponts entre les communautés favorisant la diffusion d'informations, d'avis, etc. plus variés.

Les études de Granovetter ont notamment amené les chercheurs en sociologie des réseaux à interpréter la position des individus qui forment leurs structures relationnelles d'études comme un facteur favorisant ou défavorisant leur accès à différentes ressources. Le concept de *trou structural* introduit par Ronald Burt, un autre sociologue américain, au début des années 90, a été particulièrement remarqué au sein de la communauté des sociologues des réseaux [Burt, 1992]. Son nom dérive du fait qu'il est plus intéressant pour un manager au sein d'une grande entreprise (un des terrains d'études privilégiés de Burt), en termes d'opportunités d'accès aux ressources, d'avoir des contacts non-redondants les uns par rapport aux autres. Deux contacts d'une même personne étant non-redondants s'ils ne se connaissent pas entre eux et *a fortiori* si leurs liens avec l'enquêté lui permettent d'accéder à des communautés différentes du réseau.

Burt a mis en œuvre des enquêtes de terrain en se penchant sur les primes et promotions de ses enquêtés et il a observé qu'elles étaient corrélées à ce qu'il nomme le *nombre effectif* de contacts d'un individu i au sein d'un réseau [Burt, 1992]. Pour mesurer ce nombre, il introduit une première métrique qui correspond à la redondance d'un contact j dans le réseau du point de vue de i. Pour l'obtenir, il faut calculer, pour chaque contact k de j la redondance particulière des relations avec j et k pour i, selon la formule :

$$p_{ik}m_{jk}$$

où p_{ik} est le degré d'investissement de i à sa relation à k tandis que m_{jk} représente à quel point sa relation avec k est pour j importante, en comparaison de la relation la plus forte qu'il ait parmi ses propres contacts. Notons que si la relation de i à k ou de j à k est nulle, alors k n'apporte pas de redondance à j dans sa relation avec i. À l'inverse du cas où k a une relation privilégiée avec i comme j.

En prenant la somme de cette expression pour chaque contact de j, on obtient la valeur de redondance R_{ij} de j pour i qui varie entre 0 et 1. Une valeur inverse de la

non-redondance de j pour i est donc $nR_{ij} = 1 - R_{ij}$ et le nombre effectif de i dans le réseau est ainsi décrit par la somme des valeurs de non-redondances de chaque contact de i :

$$\sum_j nR_{ij}$$

qui varie donc entre 1, dans le cas où tous les contacts de i sont redondants jusqu'au nombre total de contacts de i dans le cas inverse.

Cet indicateur nécessite d'avoir accès à la force des relations entre les individus qui composent le réseau et qu'il faut récupérer par l'observation, l'entretien, ou toute autre méthode d'évaluation plus ou moins précise et difficile à mettre en œuvre. Il a par la suite connu une simplification proposée par Steve Borgatti [Borgatti, 1997] qui, dans une acception très structurale, propose de considérer que tous les liens du réseau ont une force identique. Puisque nous travaillons nous-mêmes avec des réseaux non pondérés, comme on l'a dit dans la partie 1.3.1, on se place également dans le cas où les liens sont tous d'égale valeur.

Cette mesure de non-redondance appartient à la grande famille des indicateurs de centralité des sommets des réseaux qui, comme on l'a brièvement vu dans le cas du métro parisien, permet de classer les sommets d'un réseau en fonction d'un score. Les premières mentions de centralité dans les réseaux sociaux remontent jusqu'à la fin des années 40 et au début des années 50 avec les travaux précurseurs de Bavelas et Leavitt qui travaillaient sur des réseaux de communication [Bavelas, 1950, Leavitt, 1951]. Au gré de leurs terrains, de leurs hypothèses et des rôles spécifiques qu'ils voulaient souligner au sein des graphes, les sociologues des réseaux ont proposé de nombreux critères pour différencier un nœud plus central d'un nœud plus périphérique. Plusieurs travaux recensant et caractérisant théoriquement ces mesures de centralité ont d'ailleurs été proposés au fur et à mesure que de nouvelles apparaissaient dans la littérature [Freeman, 1978, Borgatti and Everett, 2006]

L'une des centralités les plus utilisées et les plus étudiées [Brandes, 2001] en sociologie des réseaux est la *centralité d'intermédiarité* dont j'ai déjà parlé et sur laquelle je m'appuie largement dans ce manuscrit. Elle a été proposée par Linton Freeman, un sociologue américain, en 1977. Pour chaque couple de sommets d'une même composante connexe du réseau, il existe un ou plusieurs plus courts chemins, autrement appelés chemins géodésiques, entre ces deux sommets. La centralité d'intermédiarité (le terme anglais de *betweenness* est également couramment utilisé) d'un sommet s est défini par la somme, pour chaque couple d'autres sommets a et b du réseau, du rapport entre le nombre de chemins géodésiques entre a et b passant par s, qu'on note $\sigma_{ab}(s)$ du le nombre total de chemins géodésiques allant de l'un à l'autre et qu'on note cette fois

σ_{ab}. Cette mesure de centralité est donc définie formellement comme suit :

$$\sum_{a \neq s \neq b} \frac{\sigma_{ab}(s)}{\sigma_{ab}}$$

Elle procure un score élevé aux sommets qui font le pont entre les communautés du réseau et, parmi les autres sommets, à ceux qui sont au centre de ces communautés.

Freeman propose également un indicateur qui utilise les différents scores de centralité des sommets d'un réseau en un indicateur directement lié au graphe [Freeman, 1978]. Sa métrique permet de déterminer à quel point la centralité du sommet le plus central du réseau excède celle des autres, avec une assertion générale de la centralité qu'il applique dans cette publication aux centralités de degré (plus un sommet a un degré élevé et plus il est central), de proximité (plus un sommet est à une distance moyenne courte des autres et plus il est central), et donc d'intermédiarité. Pour cela il calcule d'abord la somme des différences entre la valeur de centralité du sommet le plus central et la centralité des autres sommets, puis fait le ratio entre cette valeur et la valeur maximum que cette somme peut avoir. Dans le cas qui nous intéresse de la centralité d'intermédiarité, la formule de ce que je vais appeler dans le reste du document la **centralisation d'intermédiarité** est la suivante :

$$\frac{\sum_{i=1}^{n} \big(C_B(p*) - C_B(p_i) \big)}{n^3 - 4n^2 + 5n - 2}$$

Ici, $p*$ est le sommet le plus central du réseau, tandis que C_B représente sa valeur de centralité d'intermédiarité. Finalement n est le nombre de sommets du réseau et $n^3 - 4n^2 + 5n - 2$ est la plus haute valeur de la somme des différences de centralité d'intermédiarité d'un réseau à n sommets, c'est-à-dire dans le cas d'une étoile composée d'un sommet central et de $n - 1$ branches. La centralisation d'intermédiarité est comprise entre 0 et 1, 0 étant le cas où tous les sommets sont de centralités égales, comme dans le cas d'une clique ou d'un cycle et 1 étant le cas d'une étoile.

1.4.3 Les réseaux personnels

Tous ces travaux s'appuient sur l'observation et l'analyse de réseaux construits soit à partir de l'observation d'un groupe d'individus, soit, comme c'est le cas pour mes travaux, sur des réseaux centrés autour d'un individu. L'analyse des réseaux personnels, comme on les nomme, représente une branche de la sociologie des réseaux dont cette section présente les spécificités.

Les réseaux personnels, ou égocentrés, sont ceux sur lesquels s'appuie ce travail de thèse. Leur étude remonte à la fin des années 70 et au début des années 80 et ils

sont opposés aux réseaux dits complets, au sein desquels l'ensemble des interactions dans un groupe donné, comme une entreprise, un club sportif, etc. sont intégrées. Si plusieurs écoles existent, certaines étudiant les réseaux personnels et d'autres les réseaux complets, il semble pertinent d'utiliser les premiers dans le cadre d'études sur les plateformes de réseaux sociaux, où les questions de protection des données personnelles ainsi que la quantité de celles-ci rendent délicates des études de réseaux complets.

Plusieurs méthodes de construction des réseaux personnels coexistent et la plus utilisée est probablement celle des *générateurs de noms*, des questionnaires qu'on fait passer aux enquêtés et au travers desquels ils sont amenés à décrire les personnes qu'ils voient tous les jours, auprès desquels ils seraient prêts à demander de l'aide ou encore d'autres types d'échanges. On obtient ainsi ce qu'on appelle des réseaux d'échanges. On peut citer les travaux de Claude Fischer comme exemple d'étude sur les réseaux d'échange [Fischer, 1982]. Les réseaux obtenus par des générateurs de noms sont généralement denses, car les répondants décrivent souvent leurs relations les plus proches, et notamment les membres de la cellule familiale [Rivière, 2000], qui s'entre-connaissent fortement, ainsi que l'a montré Granovetter [Granovetter, 1977]. Cette méthodologie repose également sur l'interprétation de l'enquêté dans la décision d'inclure une arête du réseau ou pas, ce qui peut parfois amener à des imprécisions [McCallister and Fischer, 1978, Milardo, 1988].

Les réseaux d'interactions, qui consistent à faire remplir au fur et à mesure à des enquêtés l'ensemble des contacts qu'ils ont, accordent un rôle plus important aux liens faibles. L'enquête « Contacts », lancée en 1986 par l'INSEE utilise ce procédé. Elle est l'une des premières enquêtes françaises de vaste ampleur proposant des études directement liées aux questions de la sociologie des réseaux personnels. Un panel d'enquêtés devait ainsi noter, tout au long de la journée, les interactions sociales qu'ils avaient eues, en précisant leur relation avec ces personnes ou encore l'âge approximatif de celles-ci. Avec cette étude, François Héran a notamment montré que plus un individu a de relations sociales et plus il a de relations sociales avec toutes les catégories de personnes, de manière indifférenciée, à l'exception de la cellule familiale, seul groupe qui semble empiéter sur la sociabilité globale. Il propose également un croisement entre types de sociabilités (catégorisés par le temps d'interactions avec les différents groupes de relations : familles, amis, voisins, collègues...) et les catégories sociales [Héran, 1988].

Les réseaux personnels ont également été étudiés à la lumière de leur dynamique et de leur évolution. L'enquête de Claire Bidart, dans la région de Caen, est particulièrement représentative de ce type de recherches. Ses répondants, des jeunes adultes au début de l'enquête, ont été interrogés à intervalles de 4 ou 5 ans au cours de longs entretiens durant lesquels ils ont décrit leur réseau personnel et ses membres. Les chercheurs ont

ainsi pu étudier comment ces réseaux évoluent alors que les enquêtés traversent les différentes étapes de la vie [Bidart et al., 2011]. Le corpus construit, connu sous le nom de panel de Caen, est riche de nombreux réseaux personnels et on en aura utilité dans ce travail à titre de comparaison avec les réseaux personnels de Facebook sur lesquels nous travaillons dans la partie 3.1.1 et au chapitre 4.

L'analyse d'un réseau personnel ne peut pas se conduire tout à fait de la même manière que celle d'un réseau complet. Les alters, aussi éloignés qu'ils puissent l'être dans un tel réseau seraient en effet à distance 2 l'un de l'autre si un sommet représentant égo, par définition relié à chacun, était ajouté au graphe. Ainsi, Brandon Brooks, Bernie Hogan, Nicole Ellison, Cliff Lampe et Jessica Vitak proposent que les triades fermées sont en fait un marqueur de fragmentation du réseau personnel, organisé en groupes distincts les uns des autres, alors que les triades ouvertes sont justement la preuve que des liens existent entre les sphères sociales de l'enquêté [Brooks et al., 2014].

1.5 Sociologie et plateformes de réseaux sociaux

La sociologie, et notamment la sociologie des réseaux s'est toujours intéressée aux interactions médiées, aussi bien pour étudier les usages de ces médias que pour observer les impacts de leur apparition sur la sociabilité en général. Pour des cas antérieurs aux études sur les réseaux sociaux, on peut par exemple citer les relations téléphoniques étudiées par Licoppe et Smoreda qui indiquent que plus on se voit plus on s'appelle [Licoppe and Smoreda, 2000], ce que d'autres chercheurs ont par la suite nuancé [Stoica and Prieur, 2009]. Les relations via internet, d'abord au travers des échanges de mails puis des interactions via les sites de réseaux sociaux ont ainsi, avant celle proposée ici, été à l'origine de nombreuses enquêtes.

Si les données y sont beaucoup moins facilement accessibles que sur un réseau public comme Twitter, Facebook tient néanmoins une place particulière parmi les terrains d'enquête sur le net. Son grand nombre d'abonnés ainsi que les nombreuses critiques et louanges que reçoit la plateforme motivent en effet les sociologues à entreprendre des études liées à ses usages et à ses impacts. Une partie de ces études cherchent à établir l'apparition ou le ressenti en gain/perte en capital social de leurs enquêtés à partir du questionnaire proposé par Dimitri Williams [Williams, 2006], qui permet une analyse fine des différentes modalités de capital social.

Les deux formes de capital social généralement étudiées dans la littérature récente sont proposées par Robert Putnam dans une étude qui le voit conclure que la sociabilité va diminuant depuis les années 60 aux États-Unis [Putnam, 2000]. Il propose de décomposer la notion introduite par Bourdieu en capital social de *bonding* et de *brid-*

ging. Le premier correspond aux liens forts, qui sont généralement très interconnectés tandis que le second est celui qui définie la possibilité pour l'enquêté d'accéder à de nouveaux groupes sociaux, ce qui ramène aux travaux de Grannoveter dont on a parlé en section 1.4.2.

1.5.1 Facebook, un mode de discussion parmi d'autres

La majorité des travaux concernant l'influence de Facebook sur la sociabilité de ses membres remet en cause la lecture binaire qui consiste à décrire la plateforme soit comme l'outil idoine de rencontres, d'échanges et de retrouvailles qui n'auraient pas été permises sans lui, soit à l'inverse comme un lieu numérique de perte de vue de la vie réelle, des relations de tous les jours, ainsi que de déconnexion au cœur d'un monde égocentré où la mise en lumière de soi gouverne.

En 2009, Valérie Beaudoin [Beaudouin, 2009] note, dans une proposition qu'on retrouve largement dans la littérature, que les liens dits réels et ceux numériques se recoupent et recouvrent largement. Il est alors, selon elle, trompeur d'y voir une rupture des liens sociaux à même de révolutionner la vie sociale mais plutôt un entrelacement des formes de contacts, renforcé par des possibilités d'interactions multimodales [Nguyen and Lethiais, 2016]. Il serait en fait vain d'opposer « vrais liens », plus authentiques et « liens virtuels » techniquement médiés d'une manière ou d'une autre [Dagiral and Martin, 2017]. Des études menées outre-Atlantique dénotent ainsi qu'une minorité de nouvelles rencontres sont faites sur le réseau social qui est donc nettement plus utilisé dans le cadre de discussions entre connaissances préalables [Lampe et al., 2006, Ellison et al., 2007].

Une enquête menée par le Pew Research Center, un think tank états-unien, montre ainsi, après des questionnaires téléphoniques auprès de 2 277 adultes habitants aux États-Unis, qu'approximativement deux tiers des usagers déclarent utiliser les réseaux sociaux pour être en contact avec leurs amis et leur famille [Smith, 2011]. Cela entretient encore l'idée d'un renforcement par la plateforme des relations entre personnes déjà connectées. Les sociologues ne notent généralement ainsi pas de différence fondamentale dans la sociabilité en et hors ligne, bien que certains montrent que la communication via les réseaux sociaux permet d'entretenir avec plus de succès certains liens faibles [Donath and Boyd, 2004].

1.5.2 Une sociabilité influencée par l'usage

Les publications citées dans la section précédente relèvent d'études menées sur des sous-populations spécifiques, comme des étudiants, ou bien sur un panel d'enquêtés large

mais sans distinction faite, lors de l'analyse des données obtenues auprès des inscrits, des différents usages que ceux-ci font de la plateforme. C'est pourquoi, ayant l'objectif de mieux cadrer les études basées sur l'observation des internautes, des recherches se sont portées sur le déchiffrement des différents usages qu'ils font d'internet, des plateformes de réseaux sociaux et notamment de Facebook, à la manière de ce que propose le chapitre 2 de cette thèse.

Dans un premier temps, les différentes formes d'usages de la plateforme étaient séparées de manière sommaire. Dans leur étude sur la l'apport, en terme de capital social, de Facebook sur un panel d'étudiants, Nicole B. Ellison, Charles Steinfield et Cliff Lampe de la Michigan State University analysent les externalités de Facebook selon les différents usages. Ceux-ci dépendent de l'intensité d'utilisation de Facebook, une mesure qui agrège le temps passé par jour sur la plateforme et le nombre d'amis de l'enquêté, et une mesure du fait que celui-ci vient sur Facebook pour y retrouver des amis ou bien pour y rencontrer de nouveaux, construite selon des déclarations des répondants. Si ce dernier indicateur ne semble pas être très discriminant, l'intensité d'utilisation influe, elle, fortement sur les conclusions de l'étude [Ellison et al., 2007]. Bien que celle-ci soit liée de manière positive au bonding social capital, il est possible qu'une étude plus fine de l'utilisation des différentes fonctionnalités offertes par la plate-forme permette d'encore préciser les modalités d'accès au capital social en ligne.

De façon plus précise, Moira Burke et Robert Kraut de la Canergie Mellon University avec Cameron Marlow, un chercheur de Facebook distinguent trois usages du réseau social : la communication directe et individualisée avec les amis, la consommation passive des publications des autres et l'émission de contenus [Burke et al., 2011]. Cette étude, menée à partir d'analyses de l'activité sur la plateforme des répondants ainsi qu'au travers de questionnaires précis concernant leur sociabilité, fait dire à ses auteurs que seul le cas de la communication directe et individualisée permet d'augmenter le capital social dit de *bridging*, c'est-à-dire qui fait émerger des relations entre les sphères sociales, ce qu'une étude antérieure avait déjà permis de mettre en avant [Burke et al., 2010]. Les auteurs de celle-ci remarquent également une forte variabilité des usages, ainsi que de leur impact, qui dépendent de l'estime de soi et de l'aisance relationnelle des enquêtés. Ils plaident alors pour une différenciation claire des usages et des utilisateurs dans les analyses des études des réseaux sociaux.

1.5.3 Différences entre réseaux sociaux en ligne et hors ligne

Les relations hors-ligne se recoupent peut-être avec celles qu'on partage sur les réseaux sociaux en ligne mais les modalités d'observation ainsi que de construction de celles-ci ne sont pas neutres pour autant et les réseaux sociaux qu'on peut obtenir par l'agrégation de données issues de ces médias ont leurs propres spécificités. Les réseaux

du panel de Caen permettront une comparaison avec ceux de notre enquête mais voyons déjà quelques pistes offertes par la littérature.

On peut déjà noter que, contrairement aux réseaux issus de questionnaires, il est beaucoup plus probable qu'un alter central manque dans un réseau construit à partir de données collectées en ligne. Dans le cas de Facebook rien n'indique, par exemple, que le conjoint de l'enquêté ait lui même un compte sur la plateforme. Il n'apparaîtrait tout simplement pas dans le réseau dans le cas contraire.

Michel Grossetti note que les réseaux sociaux tirés de l'observation des plateformes numériques amènent un léger regain du nombre de liens faibles, qu'il est désormais plus simple de maintenir sporadiquement [Grossetti, 2014]. Il fait également l'hypothèse que la destruction de certaines barrières géographiques, entre autres, simplifie la recherche de contacts aux goûts ou habitudes similaires, favorisant ainsi, à la manière de l'urbanisation sur laquelle Fischer a travaillé [Fischer, 1982], l'homophilie au sein de ces réseaux.

S'il existe bien une sociologie des technologies de l'information et de la communication, certains chercheurs militent pour que les études ne se limitent pas à une approche via les médias ou sans eux mais aient plutôt une acception plurielle des liens sociaux [Dagiral and Martin, 2017]. Si cette thèse porte sur l'étude de réseaux issus de relations en ligne, les travaux antérieurs [Cardon, 2010, Casilli, 2010] indiquent bel et bien que les conclusions auxquelles elle aboutira peuvent porter aussi bien sur l'activité numérique des enquêtés que, dans une certaine mesure, sur leur sociabilité de tous les jours, en et hors ligne inclus.

1.6 Analyse de données

Comme beaucoup d'autres disciplines, la sociologie emprunte à l'informatique ses méthodes d'analyse de gros corpus de données pour mener ses enquêtes et cette thèse s'appuie abondamment sur certaines d'entre elles. On parle généralement de *big data* pour qualifier ce genre de corpus qui peut être composé d'une grande quantité d'objets qui peuvent eux même être décrits de manière très fine, par beaucoup de variables, créant ainsi deux niveaux de complexité.

Dans notre cas, par exemple, on étudiera la sociabilité d'individus à travers leur compte Facebook. Cette sociabilité sera décrite relativement à leur usage de la plateforme ainsi que par l'intermédiaire d'indicateurs liés à leur réseau personnel, le tout regroupant, comme on le verra tout au long du manuscrit, plus d'une centaine de métriques. Les méthodes de l'analyse des gros corpus de données se basent sur l'analyse de vecteurs,

dans des espaces de grande dimension, construits à partir de ces indicateurs. Chacun d'entre eux représentent ainsi un individu.

Généralités

L'une des méthodes les plus utilisées consiste à construire des groupes de tels vecteurs, soit d'individus pour nous, basés sur leur similarité. On appelle ces méthodes des méthodes de partitionnement ou clustering et les groupes qu'ils produisent des clusters. Elles permettent de mettre en avant des comportements ou des situations semblables à partir de modèles mathématiques. Croiser des regroupements en clusters de mêmes individus construits à partir d'indicateurs différents, par exemple catégories d'usages et catégories socio-professionnelles, permet d'étudier les relations entre ces deux données.

De nombreuses méthodes de clustering existent, basées sur différents critères ou procédures. Dans ce manuscrit, la méthode de partitionnement la plus utilisée est la méthode des kMeans, que je vais présenter rapidement.

La méthode des k-moyennes ou kMeans regroupe les données en k groupes, la valeur de k devant être fournie en entrée. Pour ce faire il place k points aléatoires qui vont varier à chaque nouvelle étape selon la moyenne des éléments les plus proches d'eux, jusqu'à former les centres des différents groupes stables comme la figure 1.20. Au final, chaque individu sera affecté au groupe dont le centre est le plus proche de lui. Cette méthode est couramment employée et très efficace mais nécessite souvent de lancer plusieurs générations de groupes car le résultat est très dépendant des places initiales des centres. Elle oblige également le chercheur à décider du nombre de groupes qu'il souhaite obtenir. Dans la section 6.3, on propose une méthode pour contourner ce dernier point et construire des groupes indépendamment du choix de k.

Choix des positions initiales des centres

Dans l'ensemble du manuscrit, les clusterings présentés qui résultent de l'invocation d'un kMeans emploient la méthode de positionnement initial des centre dite du kMeans++ qui vise à écarter les centres initiaux [Arthur and Vassilvitskii, 2007]. Cette méthode commence par placer aléatoirement un premier centre puis place successivement et aléatoirement les suivants selon une probabilité qui augmente avec le carré de la distance au plus proche centre déjà placé.

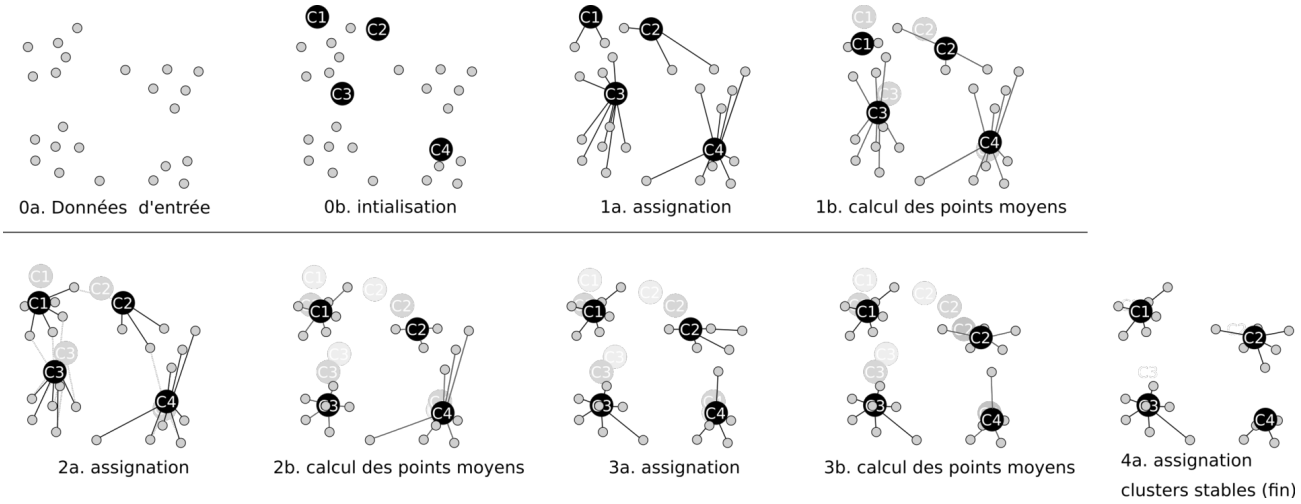


FIGURE 1.20 – Illustration des étapes d'un algorithme des k-moyennes de la page Wikipédia consacrée

Choix du nombre de groupes

De la même manière qu'il existe des méthodes qui permettent en règle générale d'améliorer le positionnement initial des centres, certaines permettent de guider le choix de la valeur de k, le nombre de groupes que le kMeans doit obtenir. Deux d'entre elles parmi les plus fréquemment utilisées sont la méthode de la silhouette et la méthode du coude. La *silhouette* [Rousseeuw, 1987] d'un regroupement en classe d'un ensemble de données permet de capturer son efficacité. De manière assez semblable à la modularité dans les réseaux, elle augmente lorsque les éléments qui sont dans le même groupe sont proches les uns des autres et que les éléments qui sont dans des groupes différents sont distants les uns des autres. La silhouette d'un clustering donné est la moyenne des silhouettes de l'ensemble de ses éléments. Considérons que, dans un espace $\mathcal{G}$ d'éléments, on a un clustering C, qu'on définit par $C = \{C_1, C_2, ...C_k\}$ où k est le nombre de groupes construits par le kMeans. Chaque groupe C_i est alors une collection d'objets (réseaux, individus, ...) $\{r_{i1}, r_{i2}, ...r_{i|C|}\}$. Pour chaque élément r_{ij} du clustering, on note $a(r_{ij})$ la mesure de bonne assignation à son cluster :

$$a(r_{ij}) = \frac{\sum_{l=1}^{|C_j|} d(r_{ij}, r_{il})}{|C_j| - 1}$$

qui est en fait la distance moyenne entre cet élément et ceux du même groupe, en notant $d(a, b)$ la distance euclidienne. À l'inverse,

$$b(r_{ij}) = \frac{\sum\limits_{\substack{1 \leqslant m \leqslant k \\ m \neq j}} \sum\limits_{l=1}^{|C_m|} d(r_{ij}, r_{il})}{|\mathcal{G}| - |C_j|}$$

est la moyenne des distances entre r_{ij} et les objets des autres groupes. À partir de ces deux valeurs, la silhouette de r_{ij} est calculée comme suit :

$$s(i) = \frac{b(r_{ij}) - a(r_{ij})}{\max\{a(r_{ij}), b(r_{ij})\}}$$

Le regroupement avec la silhouette la plus importante est généralement considéré comme étant le meilleur. La silhouette permet également, puisque le kMeans est un algorithme non déterministe, de choisir le meilleur regroupement parmi plusieurs essais à une même valeur de k.

La méthode du coude est une autre méthode couramment utilisée pour sélectionner le nombre maximal de clusters. Elle consiste à calculer une valeur strictement croissante ou strictement décroissante représentant l'efficacité du regroupement pour chaque nombre de clusters. Dans l'exemple de la figure 1.21, c'est le pourcentage de variance expliquée qui est utilisée. L'objectif est ensuite d'identifier la valeur de k à laquelle la variation de la valeur absolue de la tangente à la courbe est le plus important, ce qui correspond au fait qu'ajouter un cluster supplémentaire n'a plus beaucoup d'intérêt pour l'interprétation.

Dans mon cas, j'utilise l'*inertie* du regroupement pour représenter le coude. Soit un regroupement de i éléments notés $r_1, ..., r_i$. On note $C(r_j)$ le cluster de r_j et $c(C)$ le centroïde du cluster C. L'inertie d'un regroupement :

$$I = \sum_i (d(r_i, c(C(r_i)))^2$$

est la somme des carrés des distances entre chaque élément et le centre de son cluster. Bien qu'empirique, cette mesure est couramment utilisée pour choisir le nombre de clusters. En produisant une courbe qui passe par chaque point de coordonnées (k, silhouette du clustering pour k clusters), on peut observer à partir de quelle valeur l'augmentation de la précision ne justifie plus d'augmenter le nombre de clusters. C'est le coude, illustré par la figure 1.21, qui marque une baisse importante de la tangente de la courbe, qui permet de repérer ce seuil.

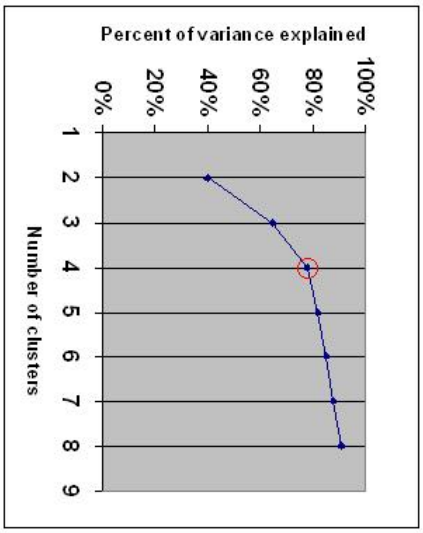


FIGURE 1.21 – L'illustration de la page Wikipédia de la méthode du coude. Le coude, la valeur la plus naturelle du nombre de clusters, est entouré sur la figure.

1.7 Informatique et sciences humaines

Je dois la chance d'avoir pu travailler dans un laboratoire de sciences économiques et sociales malgré ma formation en informatique théorique à l'importance grandissante des outils d'analyse des gros corpus de données. Avant cela, je n'avais jamais étudié l'ethnographie ou la sociologie et je suis loin de prétendre désormais les maîtriser. La recherche pluri-disciplinaire demeure un exercice délicat qui, faute d'un parcours antérieur lui même pluriel, nécessite du temps puisqu'il demande une compréhension des concepts de plusieurs champs. Malgré ces difficultés, j'ai eu la chance d'être confronté aux questions que se posent les chercheurs de ces disciplines humaines, à leurs méthodes d'enquêtes qu'ils mobilisent, aux raisonnements qu'ils entreprennent.

J'ai notamment pu collaborer, avec Julien Morel et Christian Licoppe, tous deux également chercheurs au sein de mon laboratoire, à des études sur la mobilité ludique, initialement éloignées de mes propres recherches. J'y ai contribué par ma maîtrise de certaines méthodes d'analyse de gros corpus de données qui apportaient ainsi une nouvelle dimension à leur étude mixant les approches. L'étude consistait en une analyse des logs, des traces numériques, laissées par les joueurs sur Ingress, un jeu collaboratif et géolocalisé où le but est de se rendre physiquement à certains lieux statiques afin d'y activer des portails. Nous avons récupéré les traces laissées par chaque action effectuée dans le jeu et j'ai pu construire une typologie des types de déplacements qui, à la manière de la sociabilité en ligne, se trouve être en relation directe avec

les déplacements non ludiques, découlant de l'activité journalière contrainte de nos enquêtés. Cette typologie qui décrit des modes de déplacement bipolaires de gens profitant de leur trajet jusqu'à leur travail pour activer des portails, hétérogènes de joueurs empruntant leur voiture, le week-end, pour aller jouer dans des zones auxquels ils n'ont pas le temps de se rendre d'habitude ou encore des actions inattendues de joueurs dont on ne peut savoir s'ils profitaient de leurs vacances éloignées ou d'un emploi les encourageant à beaucoup se déplacer a pu être mise en relation avec les observations faites, utilisant les outils de l'ethnologie de la mobilité dans un article présenté en annexe.

Sans m'appesantir sur ces travaux qui ne concernent pas directement la suite, je souhaitais souligner l'enrichissement et l'intérêt qui résultent du travail dans un cadre nouveau et différent de ce que j'avais connu jusqu'alors. L'objectif de mes recherches, combinant utilisation des méthodes que j'ai pu apporter à mes collègues et développement de ces mêmes méthodes à travers ce qu'eux m'ont permis d'explorer, correspond à ce que ce manuscrit tente de souligner.

Première partie

Usages

Chapitre 2

L'enquête Algopol et les usages de Facebook

L'étude des réseaux de sociabilité que j'ai menée au cours de mon travail de doctorat repose sur l'extraction d'une grande quantité de données issues de nombreux comptes Facebook. Malgré les similitudes mises en lumière par la littérature entre réseaux sociaux en ligne et hors ligne (voir section 1.5.3), on conviendra aisément que le réseau de sociabilité sur Facebook n'est pas l'exact pendant de celui qui se tisse à travers les rencontres de la vie tangible. Tout le monde ne possédant heureusement pas de compte sur la plateforme, il est de prime abord impossible de postuler qu'il n'y ait aucun individu (éventuellement même très central) du réseau social de l'enquêté qui ne soit pas manquant parmi ses contacts Facebook. De plus, les usages particuliers que font chacun de nos enquêtés des possibilités offertes par Facebook influencent les choix qu'ils y feront en terme d'ajouts et d'acceptations de contacts, rendant ainsi périlleuse une vision normative des structures relationnelles que nous avons pu construire.

Ce chapitre présente l'enquête mise en place pour récupérer les données ainsi qu'une typologie des usages de Facebook construite à partir des données qu'elle a permis de recueillir.

2.1 L'enquête

Le projet Algopol est financé par l'Agence Nationale de la Recherche et vise à l'étude de la politique des algorithmes via une approche pluridisciplinaire [1]. Étudier la politique des algorithme revient pour nous à essayer de comprendre comment ils classent les

1. algopol.fr - ANR-12-CORD-018

informations, les personnes et influent sur la circulation des informations au travers du web. Le projet regroupe des chercheurs de plusieurs universités et instituts de recherche qui, dans ce cadre, ont travaillé sur des données de Facebook et Twitter. À titre personnel, j'ai été recruté pour participer à l'analyse de données recueillies par une application Facebook.

2.1.1 Présentation de Facebook

Facebook est une plateforme de réseau social en ligne, c'est-à-dire un site qui permet aux personnes ayant créé un compte d'interagir entre elles. Généralement considéré comme un média social basé sur les contacts entre individus, plutôt que sur le partage de contenus, Facebook propose néanmoins une vaste palette de fonctionnalités, comme par exemple la pratique de jeux vidéo hébergés sur la plateforme, le partage de photos, la communication privée avec d'autres inscrits ou encore la lecture des publications des contacts, de personnalités publiques ou des nombreux organes de presse possédant un compte.

En pratique, chaque utilisateur possède une page dédiée, destinée à afficher les informations le concernant, ainsi que ses publications. Les publications, qu'on nomme également des statuts, peuvent être produites par l'individu lui-même, mais il est également possible qu'il les fasse suivre depuis la page d'un de ses contacts, ou bien d'un article trouvé sur la toile. Ils sont en effet nombreux à offrir cette fonctionnalité puisque, à l'instar de certains de ses concurrents comme Twitter, Facebook est omniprésent sur les pages du net et beaucoup de sites souhaitent profiter de l'audience importante qui s'y trouve en encourageant les re-publications sur le réseau social.

Les contacts d'un individu sur Facebook se créent par consentement mutuel et sur proposition d'un des deux intéressés. Une fois la proposition reçue et acceptée, chacun est ajouté à la liste d'amis (le terme employé par le site) de l'autre, leur permettant généralement d'interagir plus facilement. Les utilisateurs ont en effet accès au News Feed, qui correspond à la liste dynamique des publications auxquels ils sont abonnés, dont celles de leurs amis. Chacune de ces publications est susceptible de recevoir à tout moment des commentaires des personnes y ayant accès, menant parfois à de longues discussions qui lui sont rattachées. Une autre manière, plus informelle, de réagir est de laisser un *like* (ou un *j'aime* en français), qui permet aux amis du publiant de montrer qu'ils ont remarqué la publication et que, pour une raison ou une autre, elle a plu. Dans la suite de ce manuscrit, on sera ainsi amenés à étudier le réseau des commentateurs, les amis de nos enquêtés ayant publié des commentaires sur ses publications, et des likeurs, ceux qui y ont publié des likes. On dit alors des likeurs qu'ils ont liké une publication, et on aura l'occasion de réemployer ces deux néologismes qui semblent satisfaisants.

Il y a plusieurs intérêts principaux à l'étude de Facebook. D'une part, ses plus de 2 milliards de comptes (Facebook serait le troisième site le plus visité d'internet après Google et Youtube, selon le site d'analyse du trafic web Alexa) et le temps important consacré à Facebook par nombre d'utilisateurs justifient que la sociologie se penche sur les activités qui s'y pratiquent. D'autre part, le caractère mutuel des relations ainsi qu'encore une fois le fait qu'une part importante de la population s'y retrouve (près de 2/3 des internautes français sont utilisateurs de Facebook) en font certainement un des plus fidèles miroirs de la sociabilité hors-ligne.

2.1.2 Récupération des données

Une application Facebook est un outil mis à disposition des utilisateurs de la plateforme par un tiers. Il existe plusieurs milliers d'applications qui participent au succès du site lui même en offrant une riche variété de possibilités aux inscrits telles que jeux, utilitaires de gestion de compte, journaux, etc. Pour sa part, l'application Algopol permettait à ses utilisateurs de naviguer de manière interactive dans une carte du réseau égocentré construit à partir de leurs contacts Facebook ainsi que d'accéder à certaines statistiques concernant leurs interactions avec ceux-ci.

En contrepartie de cette visualisation, l'application demandait à ses abonnés de participer à nos recherches selon une procédure établie conjointement avec la CNIL. Après avoir obtenu un accord explicite de la part des abonnés, l'application récupérait de manière anonyme l'ensemble de leurs données personnelles. La politique de confidentialité de Facebook encadrant cette pratique, il était par exemple impossible de récupérer les conversations de la messagerie privée. Un court questionnaire était également soumis à l'enquêté afin de connaître sa catégorie socio-professionnelle et de qualifier avec son aide au moins 5 de ses amis. Pour ces 5 contacts, sélectionnés parmi les plus centraux de son réseau et parmi ceux avec lesquels il interagit le plus, il était demandé à l'enquêté de renseigner l'ancienneté de la relation, la fréquence des contacts en face à face et au téléphone, ou encore la force du lien.

Lancée en novembre 2013, l'application a permis la constitution d'un corpus de 16 410 enquêtés, jusqu'à sa fermeture en avril 2015 suite à un changement de la politique de Facebook concernant la vie privée de ses utilisateurs. Il a en effet été décidé qu'il ne serait désormais plus possible aux applications d'accéder à la liste des liens entre les amis de leurs abonnés, à moins que ceux-ci ne soient eux mêmes utilisateurs de l'application. Cette information est justement à la base de la construction des réseaux égocentrés au cœur d'une grande partie de nos analyses. Ce changement de politique n'a *a priori* rien de critiquable, bien au contraire, puisqu'il protège en fait les données des amis de l'utilisateur d'une application.

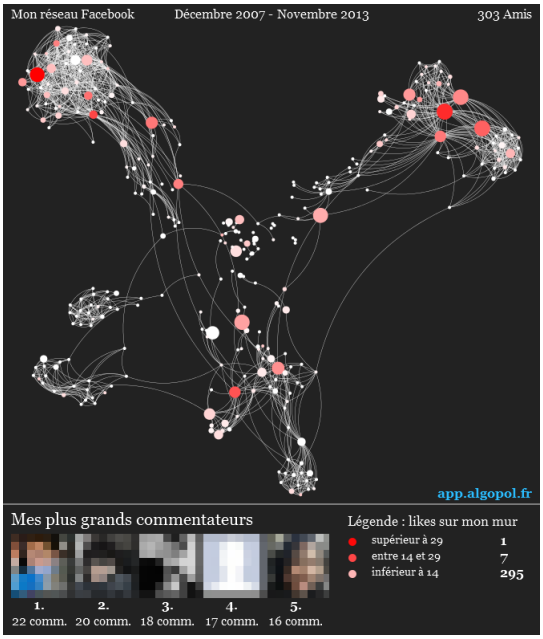


FIGURE 2.1 – Capture d'écran prise depuis l'application Algopol. La partie principale représente le réseau égocentré dont la couleur et la taille des sommets peut varier selon plusieurs critères au choix de l'enquêté. En bas on voit la liste des amis interagissant le plus avec ce dernier.

Nous avons néanmoins pu montrer à partir de nos données que les liens du réseau pouvaient être reconstruits avec une grande précision, en analysant les discussions, toujours accessibles, sur la page de profil de l'enquêté et en les inférant à partir des discussions qui s'y trouvent [Nasim et al., 2016]. On observe que les commentaires d'une publication sont en effet souvent issus d'une partie précise du réseau. Les alters qui répondent souvent aux mêmes statuts ont donc de grandes chances d'être amis entre eux.

2.1.3 Recrutement et biais associés

Bien que l'enquête Algopol ait été diffusée grâce à deux articles publiés dans le quotidien généraliste *le Monde*, un grand nombre des enquêtés a également été recruté par le bouche-à-oreille. L'enquête Algopol a été portée par des collègues chercheurs auprès de leurs étudiants ou transmise par les répondants eux-mêmes à leurs amis via une fonctionnalité de l'application. Cette communication particulièrement de l'enquête a induit des biais de sélections auxquels qu'il convient de ne pas négliger.

En sus du fait que les lecteurs du *Monde* ne sont probablement pas représentatifs de la population française, le corpus compte également un nombre important de jeunes et d'étudiants, probablement en partie recrutés sur les bancs de l'université. Nos enquêtés sont également en général citadins et technophiles. Compte tenu de cela, un panel complémentaire a été commandé à l'institut de sondage CSA, qui s'est généreusement prêté au jeu du questionnement des nouvelles méthodes d'enquêtes en ligne. Le panel dit CSA est composé de 835 comptes Facebook, dont les propriétaires, toujours anonymes, sont représentatifs de l'ensemble des internautes français. Avec ces deux corpus, nous pouvons ainsi comparer les proportions de représentations des différentes populations entre un panel représentatif et un plus nombreux.

Le tableau 2.1 souligne les fortes disparités entre les deux panels, justifiant de fait l'utilisation de l'échantillon CSA qui permet d'éviter les interprétations trop biaisées. On note ainsi que le ratio entre le nombre d'hommes et de femmes est très déséquilibré dans le cas du corpus viral avec un rapport de 68/30 contre 57/42 pour le corpus CSA, bien plus homogène. Le corpus viral est également plus jeune en moyenne que la population représentative des internautes français puisque 45% de ses membres ont entre 18 et 24 ans tandis que seuls 10% des enquêtés du panel CSA appartiennent à cette tranche d'âge. À l'inverse 29% des répondants CSA ont plus de 45 ans contre 9% seulement dans le corpus viral.

En plus de ces fortes disparités de genre et d'âges, les répondants à l'enquête qui n'ont pas été recrutés au sein du panel représentatif sont en moyenne plus diplômés et ont un emploi plus qualifié, comme l'attestent les 32% de cadres ou travaillant dans une

	Corpus		Panel CSA		Panel « viral »	
Total	**15 408**		**768**		**14 640**	
Total Genre renseigné	**15 142**		**757**		**14 386**	
Femme	**4 648**	31%	**438**	58%	**4 210**	29%
Homme	**10 494**	69%	**319**	42%	**10 175**	71%
Total Âge renseigné	**11 741**		**471**		**11 270**	
moins de 18 ans	**328**	2.8%	**2**	0.4%	**326**	2.9%
18-24 ans	**5 081**	43%	**46**	10%	**5 035**	45%
25-34 ans	**3 524**	30%	**163**	35%	**3 361**	30%
35-44 ans	**1 668**	14%	**126**	27%	**1 542**	14%
45-59 ans	**961**	8.2%	**98**	21%	**863**	7.7%
plus de 60 ans	**179**	1.5%	**36**	7.6%	**143**	1.3%
Total Profession renseignée	**10 833**		**479**		**10 100**	
Lycéens	**299**	2.7%	**1**	0.2%	**298**	3.0%
Étudiants	**4 261**	40%	**17**	3.5%	**4 244**	42%
Chômeurs, inactifs	**416**	3.8%	**118**	25%	**279**	2.8%
Ouvriers, employés	**829**	7.7%	**138**	29%	**691**	6.8%
Prof. intermédiaires	**1 003**	9.2%	**87**	18%	**916**	9.1%
Cadres, prof. libérales, prof. intellectuelles	**3 537**	33%	**81**	17%	**3 456**	34%
Commerçants, chefs d'entreprises	**401**	3.7%	**13**	2.7%	**388**	3.8%
Retraités	**87**	0.8%	**24**	5.0%	**63**	0.6%

TABLE 2.1 – Les catégories socio-professionelles des enquêtés, selon leur panel d'origine.

profession libérale ou intellectuelle contre 14% au sein de l'échantillon représentatif, ou encore les 39% d'étudiant du corpus viral, contre 4%. Le panel CSA comporte 36% d'ouvriers ou employés ainsi que 19% de professions intermédiaires et 16% de chômeurs ou inactifs pour ce qui concerne les catégories professionnelles les plus représentées.

Ces constatations illustrent bien les difficultés inhérentes aux méthodes de recherche basées sur des données massives. Bien que celles-ci, et notamment celles issues du net, puissent être plus facilement accessibles aux chercheurs en sciences sociales que celles récupérées au travers d'entretiens ciblés sur des populations représentatives, leur grand nombre n'en fait pas pour autant un gage d'une quelconque représentativité, et les conclusions qu'il est possible d'en tirer se doivent d'être drapées de précautions [Bastard et al., 2013].

2.1.4 Les métadonnées

En plus de la représentativité des individus auxquels elles appartiennent, la collecte des données sur des plateformes complexes telles que Facebook pose également un problème d'interprétabilité. Si les données sont en effet riches et nombreuses, il n'en demeure pas moins que leur utilisation repose parfois sur des jeux d'équilibriste. Plusieurs champs descriptifs des données à notre disposition se sont ainsi retrouvés être redondants, incomplets ou inutilisables. De ce qu'on a pu apprendre, cela résulte du fait que chaque équipe de développement de Facebook utilise ses propres terminologies, chacune modifiant les données indépendamment et que peu d'opérations de normalisation sont effectuées. En outre, il arrive que Facebook mette en place de nouvelles fonctionnalités à la disposition de ses usagers, ou bien en supprime, amenant l'apparition ou la disparition de champs consacrés qui sont parfois difficiles à détecter dans l'amas de données, comme l'illustre la figure 2.2, dans laquelle les sauts de valeurs correspondent à des changements des métadonnées par les équipes de Facebook.

L'étude de la dynamique des réseaux sociaux est un exemple parmi les recherches que nous aurions souhaité entreprendre, puisque la date d'apparition des alters et des liens dans le réseau égocentré de l'individu peut par exemple guider la construction d'hypothèses sur la caractérisation de ses communautés de contacts (groupes d'amis du lycée, de l'université selon l'âge de l'enquêté, ou même changement d'emploi etc.). Mais l'entreprise est *a minima* difficile à exécuter sans passer par la mise en place d'une méthodologie hasardeuse. En effet, au lieu de fournir des informations telles que « l'enquêté est devenu ami avec Jean-Marc Duvoisin » assorties d'un horodatage précis, l'API (*Application Programming Interface*, logiciel permettant d'interroger la base de données d'un site) de Facebook agrège le plus souvent les nouveaux liens d'amitié d'ego sous la forme de messages tels que « l'enquêté est devenu ami avec 6

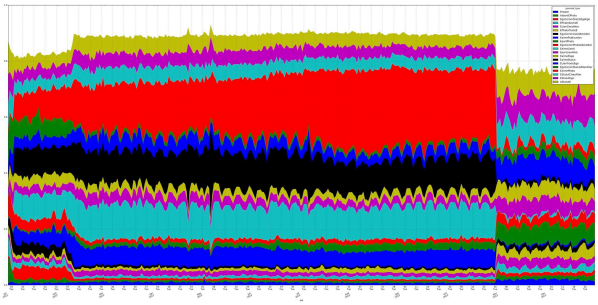


FIGURE 2.2 – Chaque plage de couleur représente un type d'activité élémentaire. On voit bien sur cette figure qu'au delà des variations normales journalières ou hebdomadaires, certains évènements changent en profondeur la répartition des activités élémentaires.

nouvelles personnes ». Il est de ce fait impossible de déterminer le moment d'apparition des sommets dans le réseau.

2.2 Méthodes d'analyse

Afin de construire, puis interpréter les catégories d'usage de Facebook qu'on veut différencier les unes des autres préalablement à une étude des réseaux, il nous a fallu mettre sur pied une batterie de méthodes à même de détecter les actions réalisées sur la plateforme. On introduit également dans cette section la *représentativité*, métrique d'abord utilisée dans ce chapitre pour mettre en valeur les différences de constitution des groupes puis, à partir du chapitre 4, pour la construction de typologies de réseaux.

2.2.1 Accéder aux activités élémentaires

En tenant compte des éléments évoqués plus haut, nous avons choisi de différencier les catégories d'usages de Facebook par les taux d'apparition des différentes *activités élémentaires*, accessibles via les métadonnées des statuts de nos enquêtés. Nous avons nommé activités élémentaires les différents types de métadonnées formant l'ensemble de la diversité des activités auxquelles nous avons accès pour chaque enquêté. C'est à force d'un travail acharné et de relectures fastidieuses qu'Irène Bastard et Baptiste Fontaine ont réussi à exhumer cette catégorisation de nos données.

En plus des inévitables « Autre » et « Erreur » indiquant une publication inclassable ou bien un export erroné des données, les activités élémentaires sont au nombre de 90

dont la majorité peut être classée selon 4 critères :

— l'auteur de l'action qui peut être soit l'enquêté lui-même soit un de ses amis
— le « lieu » de l'action (la page de l'enquêté, celle d'un de ses amis ou celle d'un groupe)
— le type d'action (partager, commenter, etc.)
— le contenu associé à l'action (une photo, un texte, un lien, etc.)

Quelques exemples parmi les 92 activités élémentaires :

— *Ego* partage un lien hypertexte sur sa page
— *Alter* publie une photo assortie d'un texte sur la page d'*ego*
— *Ego* publie une photo sur la page d'un groupe, assortie d'un texte et de la mention de certains de ses alters.

Rien ne permet de croire que la présence ou l'activité d'un usager de Facebook soit continue dans le temps. C'est peut-être encore plus souvent le cas pour la majorité des utilisateurs de la plateforme, loin des clichés de l'égocentrisme à tout va et de la quête acharnée de la reconnaissance d'un public qu'on leur attribue souvent. Afin de contrecarrer cet effet, nous n'avons traité pour chaque enquêté que la plus petite fenêtre temporelle contenant 80% de son activité.

2.2.2 La méthode du χ^2

Dans le reste du manuscrit, plusieurs croisements de populations seront présentés, comme par exemple, dans la section 2.4, le croisement des catégories d'usage avec les catégories d'âge ou professionnelles. Afin de vérifier si le fait d'appartenir à une catégorie (disons d'âge) particulière a une influence sur l'appartenance à une catégorie croisée (par exemple d'usage), on utilise souvent le test du χ^2, développé par Karl Pearson entre les... 19eme et 20eme siecles. C'est cette méthode que je vais employer. Une **matrice de contingence** est une matrice croisant deux découpages différents d'une population. Chaque case d'une telle matrice correspond au nombre d'individus qui appartiennent à la sous-population en colonne (par exemple les moins de 25 ans, dans le cas de catégories d'âge, et à la sous-population en ligne (par exemple les non-actifs) donnant ainsi le nombre d'individus correspondants à l'union de ces deux sous-catégories (soit les non-actifs de moins de 25 ans dans notre exemple).

La méthode du χ^2 dite d'indépendance, procède en comparant la matrice de contingence obtenue O à une matrice de contingence dite attendue E. Cette matrice attendue peut être fournie par le chercheur, dans le cas où il attend des résultats particuliers, mais correspond généralement (et c'est le cas pour nous) à une répartition équiprobable, qui traduit donc le fait que les deux catégorisations sont indépendantes.

Chaque case de la matrice attendue est définie par

$$E_{ij} = \frac{O_{i+} \times O_{+j}}{N}$$

où O_{i+} est le nombre d'individus qui appartiennent à la même classe i de la catégorie en lignes, O_{+j} celui du nombre d'individus qui partagent la même classe j de la catégorie en colonnes, et N est le nombre total d'individus.

Pour chaque case de la matrice observée, le processus fournit finalement un score $T_{i,j}$ d'écartement entre la matrice observée et la matrice attendue :

$$T_{i,j} = \frac{O_{i,j} - E_{i,j}}{\sqrt{E_{i,j}}}$$

Cette valeur est positive lorsque le nombre observé d'individus est supérieur à celui attendu, négative lorsqu'il est inférieur et nulle dans le cas où les deux valeurs concordent. Sa valeur absolue augmente quand l'éloignement entre l'attendu et l'observé augmente.

2.3 Activités élémentaires et catégories socio-professionnelles

Un premier niveau d'analyse des activités élémentaires, est d'observer la distribution des plus pertinents d'entre eux pour les catégories socio-professionnelles usuelles que notre enquête a permis d'obtenir pour chaque enquêté. Les cartes de chaleur, ou *heatmaps*, présentées ci-après décrivent les liens entre des catégories d'individus, par exemple les jeunes de moins de 25 ans, et des variables d'intérêts sélectionnées parmi les activités élémentaires.

Pour ce faire, on commence par calculer la moyenne de chaque variable pour l'ensemble des individus. Soit M_v la moyenne globale de la variable v. Par exemple dans le cas où v représente *ego joue*, M_v représente le nombre moyen de publications catégorisées par notre algorithme comme « *Ego joue* ». Pour chaque heatmap, on regroupe ensuite les individus selon les catégories correspondantes puis on calcule la moyenne des variables pour ces catégories, soit $M_v(C)$ la moyenne de v pour la catégorie C. En poursuivant avec « *Ego joue* », on obtient par exemple que $M_v(< 25)$ est le nombre moyen de publications de jeu parmi les moins de 25 ans. Finalement, la valeur d'une case de heatmap, à l'intersection de la ligne assignée à la variable v et de la colonne assignée à la catégorie C vaut :

$$\log \frac{M_v(C)}{M(v)}$$

Dans la première carte de chaleur, présentée en figure 2.3, la valeur de 0.12 à l'intersection entre la colonne décrivant les valeurs moyennes de nos variables pour les enquêtées et la première ligne *ego joue* signifie que les femmes ont $10^{0.12} = 1.32$ fois plus de publications liées aux jeux que les hommes. Pour chaque heatmap, une échelle de couleur allant du rouge pour les valeurs les plus élevées au bleu pour les plus faibles permet de rapidement identifier les tendances.

La répartition selon le genre des enquêtés est présentée en figure 2.3. On note de prime abord que les variations de valeur moyenne n'ont pas une amplitude très importante mais permettent néanmoins de déterminer quelques tendances. On relève ainsi que les femmes semblent en moyenne être plus actives pour ce qui est des actions liées à la conversation, comme le fait de commenter un statut sur son mur, ou bien de publier un statut sec, c'est-à-dire sans photo ni lien. C'est probablement en partie pour cette raison qu'encouragés à participer à la discussion, les amis des nos enquêtées sont plus enclins à publier des statuts sur le mur de celles-ci que dans le cas des hommes. Ces derniers semblent en moyenne être plus nombreux à se spécialiser dans le partage d'information avec un haut taux de publication de liens hypertextes de pages internet tant par nos enquêtés que par leurs amis sur leur mur.

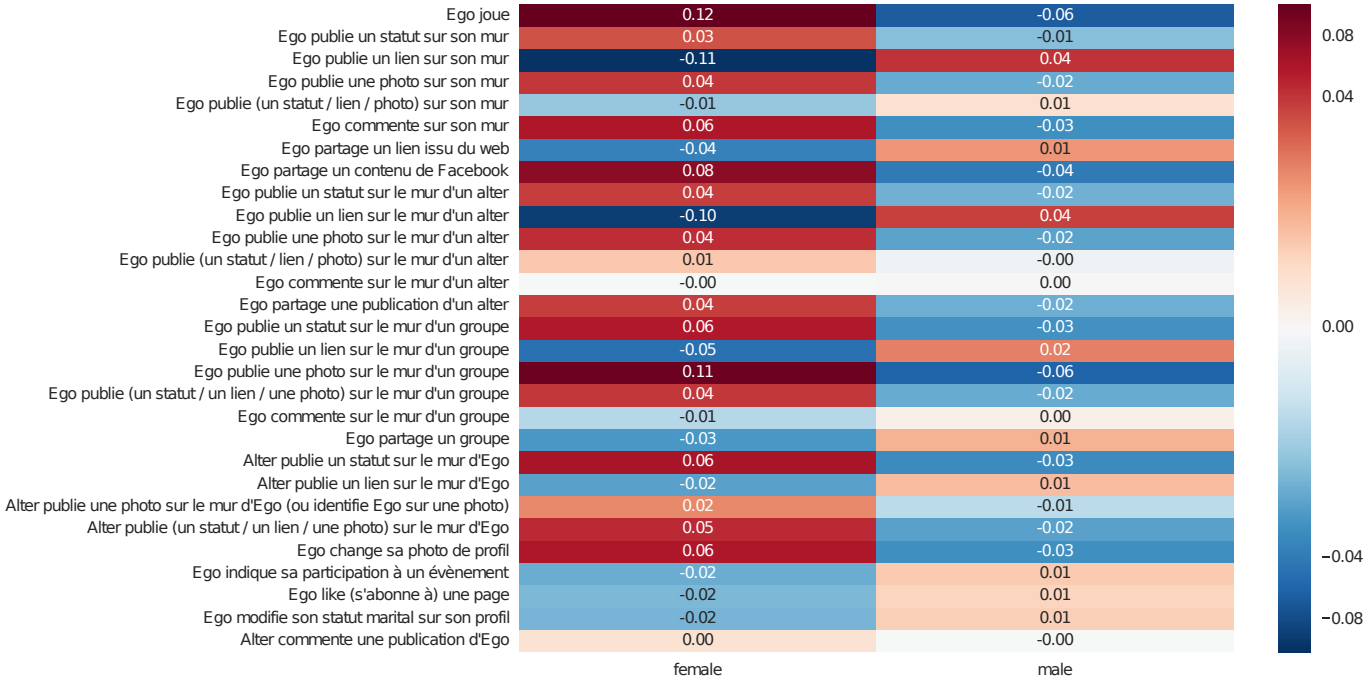


FIGURE 2.3 – Heatmap des activités élémentaires par genre en prenant les données CSA et virales

Bien qu'en lien moins direct avec la sociologie des réseaux, dans le sens où il n'est pas

possible de savoir ici si l'activité est participative ou individuelle, on retrouve le fait que les femmes jouent plus, de manière même plus marquée que pour les autres variables. Par ailleurs, elles sont généralement plus concernées par la publication de photos, aussi bien sur leur page que sur celle d'un groupe ou d'un de leurs alters. Elles changent également plus souvent leur photo de profil, ce qui pourrait tout aussi bien être la marque d'une plus grande recherche de visibilité qu'à l'inverse, d'une désacralisation de cet avatar, qui changerait alors au gré des humeurs.

La comparaison de nos variables par rapport à l'âge de nos enquêtés, illustrée par la figure 2.4, met en évidence une plus forte dépendance des usages de Facebook qu'avec le genre des répondants. On note d'ailleurs que la majorité des activités élémentaires présentent un niveau de corrélation ou de corrélation négative remarquable avec les catégories d'âges proposées ici, avec de fortes régularités puisque les lignes sont généralement formées de cases variant de manière monotone du bleu foncé au rouge foncé, ou inversement.

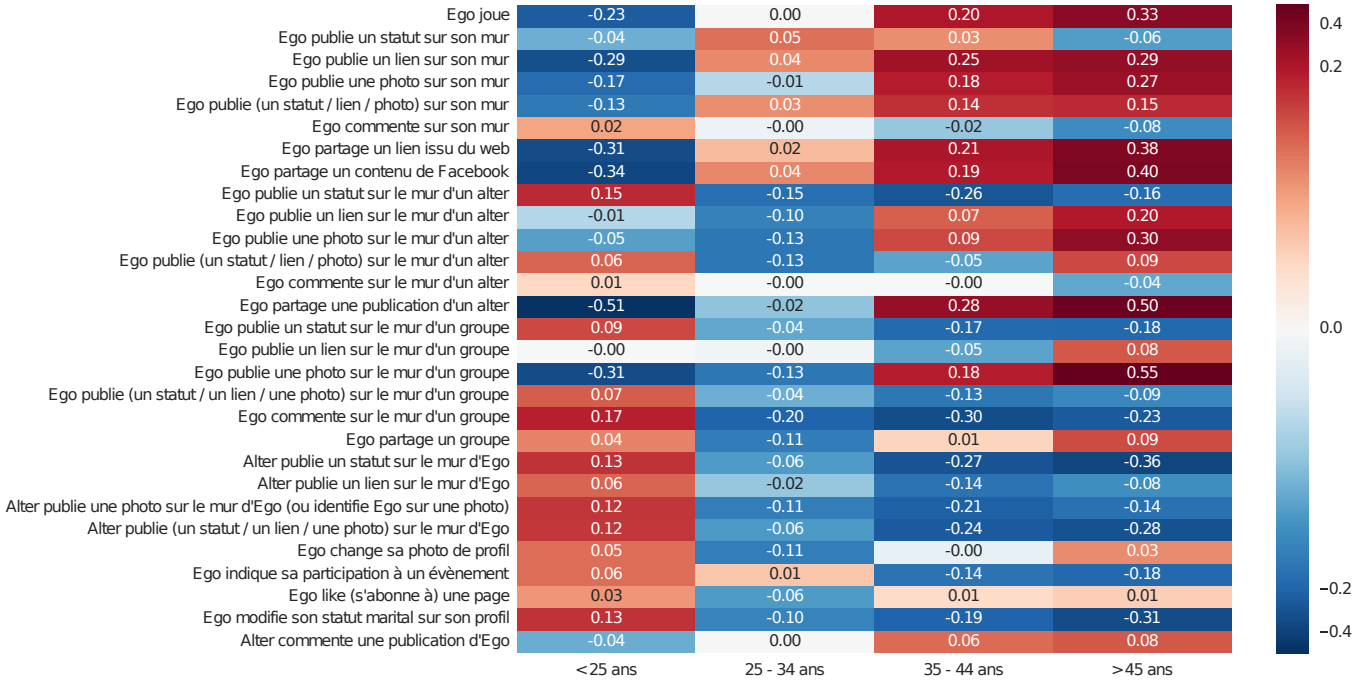


FIGURE 2.4 – Heatmap des activités élémentaires par tranches d'âges en prenant les données CSA et virales

Ainsi, le partage et la publication de liens hypertextes ou de contenus trouvés sur Facebook est ainsi le domaine privilégié des utilisateurs plus âgés, que ce soit sur leur propre page ou sur celle d'un de leurs alters ou d'un groupe. Les utilisateurs de plus

de 35 ans et *a fortiori* de plus de 45 ans sont ceux qui publient le plus de photos et également ceux qui jouent le plus.

À l'inverse, les plus jeunes semblent plus portés par les activités liées à la conversation en ligne, ce que suggèrent leur tendance à plus commenter les publications sur leur mur et surtout leur activité de publication et de commentaire sur les murs de leurs amis, ainsi que, réciproquement, de ces derniers sur la page des enquêtés. Cela semble indiquer que les jeunes utilisent plus volontiers Facebook comme un outil de communication avec leurs proches, « chez lesquels » ils se sentent légitimes de publier. Cette idée est également mise en lumière par le fait que ces derniers sont également plus enclins à déclarer participer à un évènement, information qu'on imagine être destinée à leurs amis susceptibles de les y accompagner.

Parmi toutes les catégories socio-professionnelles, celles qui décrivent l'activité professionnelle des enquêtés proposent l'amplitude la plus importante, comme le montre la figure 2.5. On note que cela vient également du fait que les catégories étudiants et surtout lycéens dépendent également de l'âge des répondants, et on remarque d'ailleurs que beaucoup d'indicateurs sont positifs pour ces deux catégories et négatifs pour les autres, ou à l'inverse négatifs pour eux et positifs pour le reste.

Encore une fois, le jeu est assez discriminant sur Facebook, opposant les classes plus populaires aux autres, et notamment aux étudiants et lycéens. Rien n'indique cependant qu'on puisse en conclure que le jeu vidéo est une activité représentatives des classes populaires plutôt que des autres, tout au plus que Facebook semble être la plateforme privilégiée par ces dernières, peut-être au détriment de consoles de salon ou autres.

Comme dans le cas de la catégorisation par l'âge, on retrouve que les étudiants et lycéens sont plus généralement tournés vers la publication de statuts que de photos ou de liens hypertextes, et n'hésitent pas à interagir directement sur le profil de leurs amis ou sur des groupes. Ces deux groupes ne doivent cependant pas être traités comme identiques en tous points malgré leurs similitudes. Les lycéens partagent beaucoup moins de groupes et indiquent nettement moins participer à des évènements, deux activités qu'on peut probablement rapprocher de marqueurs de la sociabilité des jeunes adultes.

Parmi les autres catégories, on note une plus grande propension des classes populaires, incluant les commerçants et chefs d'entreprise, à la conversation distribuée sur la page de leurs amis ou sur les groupes, ainsi que de leurs alters sur les autres pages, là où les professions intermédiaires, cadres, professions libérales et intellectuelles sont plus centrées sur leurs propres pages.

Toutes ces catégories socio-professionnelles montrent que les activités élémentaires décrites par nos activités élémentaires exhibent des différences pertinentes et peuvent

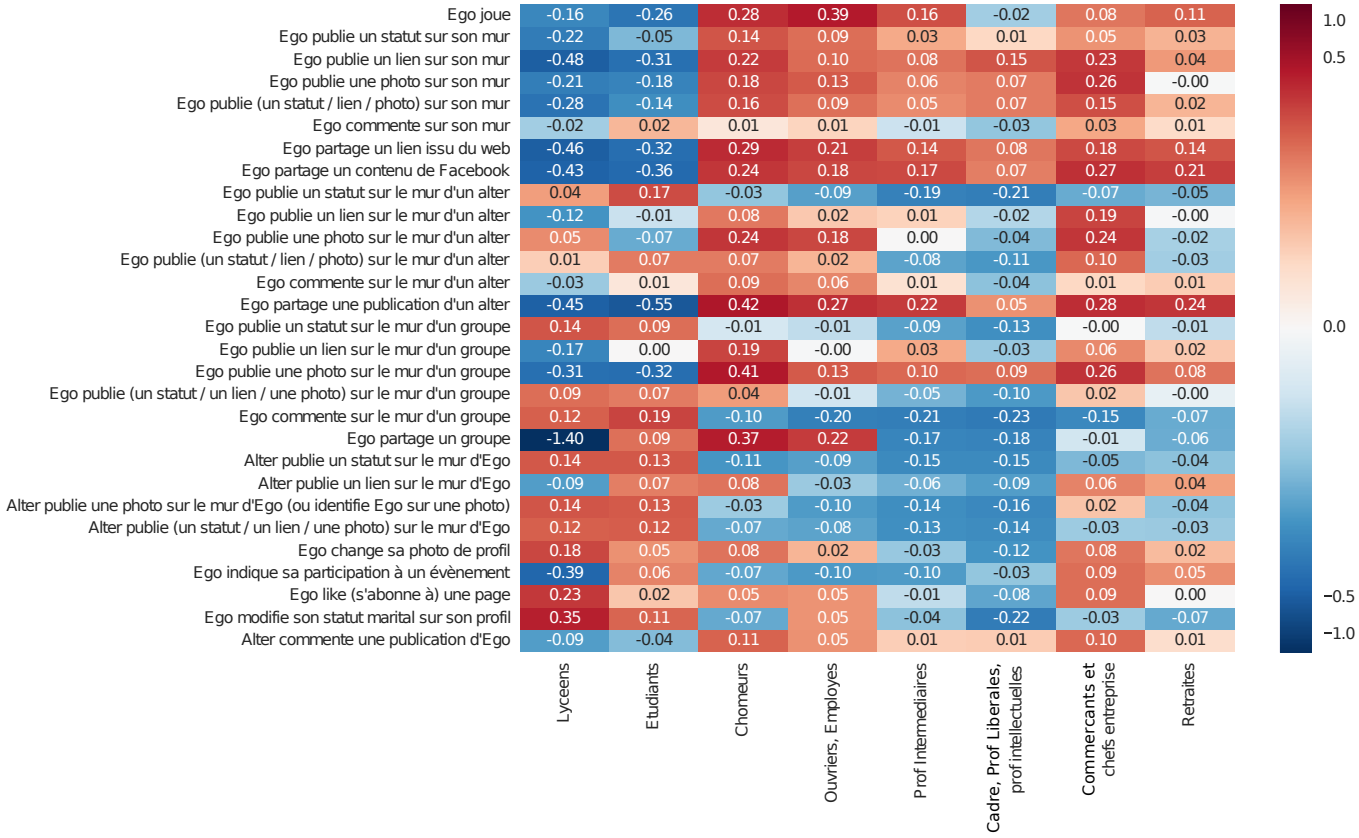


FIGURE 2.5 – Heatmap de la correspondance entre activités élémentaires et catégories professionnelles.

permettre de mettre au jour la variation des usages de Facebook. Il n'est cependant pas question, dans la suite, de considérer que la catégorie socio-professionnelle de nos enquêtés détermine la manière dont ils utilisent le site de réseau social.

2.4 Six profils d'usages de Facebook

Nous souhaitons construire une catégorisation effective des usages de Facebook afin de pouvoir associer chaque enquêté non pas à un groupe qui dépendrait des catégories sociales et professionnelles usuelles, mais aux activités élémentaires qu'il pratique le plus.

In fine, nous caractérisons chaque *ego* selon quatre critères :

— le volume total de son activité

- la répartition des activités d'*ego* parmi
  - la publication « *ego publie ...* »
  - le commentaire « *ego commente ...* »
  - le partage « *ego partage ...* »
- la répartition des lieux d'activité d'*ego* parmi
  - sa propre page « *ego ... sur son mur* »
  - les pages de ses contacts « *ego ... chez un alter* »
  - des pages de groupes Facebook « *ego ... dans un groupe* »
- la répartition des personnes qui publient sur son mur parmi
  - ego lui même « *ego ... sur son mur* »
  - l'un de ses amis « *Alter ... sur le mur d'ego* »

Une fois les enquêtés représentés par ces 9 indicateurs, nous avons pu bâtir une typologie des usages de Facebook en utilisant la méthode de clustering *kMeans*. Comme on l'a vu au cours de la section 1.6, la méthode des kMeans prend en paramètre le nombre de groupes dans lesquels on souhaite distribuer les variables. Nous avons fait le choix de les regrouper en six configurations d'activités, décrites par le tableau 2.2, que je vais présenter ici. Trois super-catégories chapeautent ces six groupes :

- les non-actifs
- ceux qui publient majoritairement sur leur page de profil
- ceux dont l'activité est répartie sur les différents « lieux » de publication offerts par la plate-forme.

La décomposition des activités élémentaires par configuration d'usage est décrite par la figure 2.6.

On peut commencer par noter qu'il y a une opposition systématique entre le corpus global et le corpus représentatif lorsqu'il s'agit de comparer la représentation dans les catégories d'usage selon que l'on soit un homme ou une femme. En plus de nous contraindre à analyser ces résultats avec prudence, cette opposition interroge également sur les spécificités de notre panel qui l'entraînent.

2.4.1 Publier chez soi

Principale catégorie en termes de nombres de personnes recrutées, aussi bien par le biais de l'enquête virale que dans l'échantillon CSA, le groupe des usagers dont l'activité se concentre majoritairement sur leur propre page se divise en trois profils. Au sein du panel CSA, les femmes sont plus présentes dans ces catégories, et notamment chez les égovisibles alors qu'on remarque l'inverse pour le panel viral. La figure 2.8 indique également que ces trois catégories correspondent à celles qui regroupent le plus d'enquêtés de plus de 35 ans.

		Non Actifs	Publier chez les autres		Publier chez soi		
Corpus			**Conv. de groupe**	**Conv. distribués**	**Égocentrés**	**Égovisibles**	**Partageurs**
Effectifs							
Corpus viral	**15 145**	**2 634** 17%	**1 273** 8%	**3 532** 23%	**3 610** 24%	**3 070** 20%	**1 026** 7%
Corpus CSA	**735**	**19** 3%	**30** 4%	**164** 22%	**302** 41%	**81** 11%	**139** 19%
Indicateurs constitutifs de la classification							
Activité totale (par jour)	**7.4**	3.0	7.6	6.6	4.5	28.2	11.7
Part *ego publie ...*	**60%**	88%	55%	57%	48%	64%	34%
Part *ego commente ...*	**33%**	1%	40%	40%	44%	25%	14%
Part *ego partage ...*	**7%**	11%	5%	4%	8%	11%	52%
Part *ego ... ... sur son mur*	**61%**	99%	31%	34%	67%	75%	80%
Part *ego ... ... chez un alter*	**9%**	1%	31%	59%	28%	20%	17%
Part *ego ... ... dans un groupe*	**30%**	0%	38%	7%	5%	5%	3%
Part *ego ... ... sur son mur*	**55%**	41%	54%	49%	60%	56%	71%
Part *Alter ... ... sur le mur d'ego*	**45%**	59%	46%	51%	40%	44%	29%
Nombre d'amis médian, calculé consécutivement à la classification							
	282	287	319	351	201	332	189
Nombre de commentaires médian reçu des alters, calculé consécutivement à la classification							
	155	84	140	160	116	372	101

TABLE 2.2 – Tableau récapitulatif des configuration d'usage de Facebook

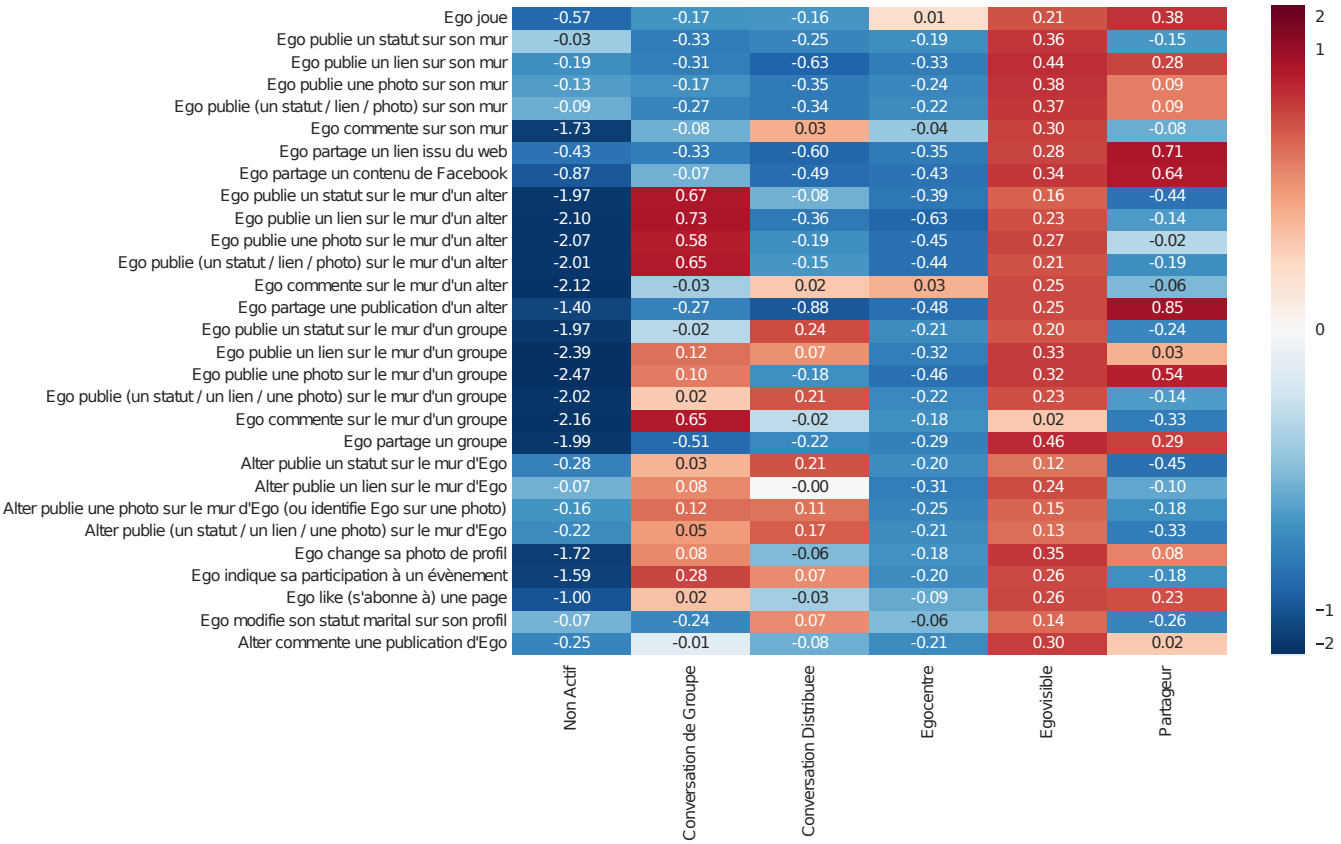


FIGURE 2.6 – Heatmap de la correspondance entre activités élémentaires catégories d'usage.

Les égocentrés

24% du panel viral - 41% du panel CSA

Les égocentrés publient et commentent majoritairement sur leur propre page mais demeurent, parmi les trois groupes ainsi caractérisés, celui qui a en moyenne le plus fort taux d'activité sur les pages de leurs amis avec pas moins de 28% de leurs publications qui y sont localisées. Avec un taux de publication et de commentaire sensiblement aussi importants l'un que l'autre, ces personnes n'hésitent pas à participer aux discussions qu'ils ne se contentent donc pas d'initier. Ce groupe correspond à 24% du corpus viral mais est de loin le plus important du panel représentatif avec 41% des enquêtés, indiquant qu'il s'agit probablement de l'usage le plus commun de Facebook. La relativement faible représentation de cette catégorie d'usage dans le panel viral suggère que les jeunes et les plus diplômés, qui y sont très représentés, ont un usage généralement différent de Facebook. En considérant son assez faible activité totale (0.65 en moyenne

contre 0.87 pour l'ensemble des enquêtés), on arrive à la conclusion que la majorité des gens publient donc assez peu et plutôt sur leur profil, ce qui suggère finalement le caractère peu démonstratif de la majorité des usagers.

Comme l'indique la figure 2.7, la part des femmes pour l'ensemble du corpus y est parmi les plus faible tandis qu'à l'inverse, elles sont les plus nombreuses au sein du panel CSA. Ces dernières y sont d'ailleurs sur-représentées au sein des trois classes qui publient majoritairement sur leur propre page. Concernant les catégories socio-professionnelles, récapitulées en figure 2.9, on note que les étudiants sont très sous-représentés, aussi bien dans le panel viral que représentatif, en accord avec la faible représentation des jeunes de moins de 24 ans, notamment au sein du panel représentatif (voir figure 2.8. Ces données suggèrent que la quête de visibilité est plutôt l'apanage d'adultes, généralement actifs.

Les égovisibles

20% du panel viral - 11% du panel CSA

Contrairement à l'usage précédent et plus à l'image des autres, ce groupe est plus représenté dans le panel viral avec 20% qu'au sein du panel CSA puisque représentant 11% de ses membres. À l'inverse des égocentrés dont on pourrait presque qualifier l'activité de timide, les égovisibles sont caractérisés par leur très grand nombre de publications journalières, dont une large majorité, 64%, le plus haut score si on excepte la catégorie des non-actifs, sont des publications sèches. Les égovisibles ont beaucoup d'amis, leurs publications reçoivent beaucoup de commentaires et ce sont même ceux qui changent le plus de photo de profil. Tout laisse à penser que ce sont eux qui nourrissent les représentations collectives de l'utilisateur de Facebook en quête de visibilité et de reconnaissance publique en ligne. La différence marquée de leur proportion d'apparitions selon le corpus (viral ou représentatif) montre cependant qu'ils sont moins nombreux que ce que laisse à penser l'omniprésence de la visibilité de leur activité en ligne. Il est délicat de dire si cette sur-représentation est liée aux catégories socio-professionnelles les plus représentées parmi nos enquêtés hors CSA ou bien si la recherche d'exposition n'a pas été dans certains cas le moteur de la participation à notre enquête.

Les égovisibles sont fortement sous-représentés parmi les jeunes de moins de 25 ans ainsi que chez les étudiants, ce qui suggère encore une fois que ces derniers ne sont pas les plus en recherche de la construction d'une réputation en ligne. À l'inverse, la catégorie est celle qui regroupe le plus de cadres, ou exerçant une profession libérale ou intellectuelle, dont on peut imaginer que les représentants se sentiraient plus facilement légitimes à exercer une pratique de Facebook menant à la reconnaissance d'une personnalité numérique.

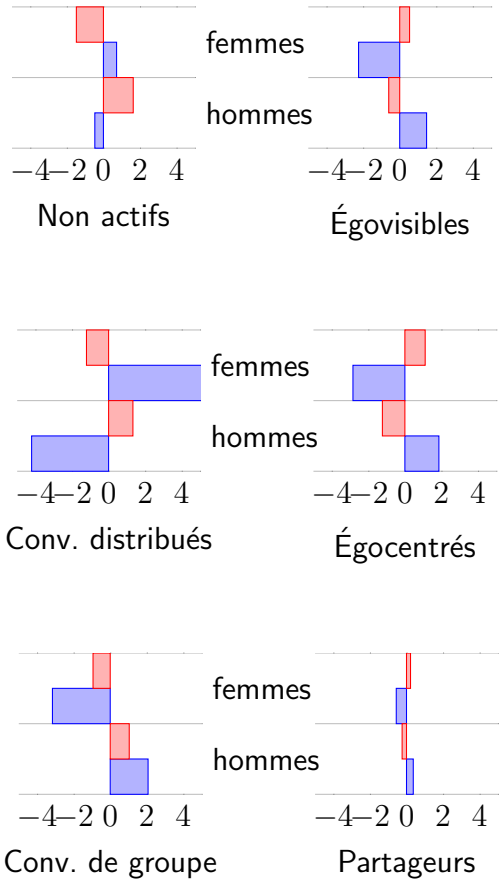


FIGURE 2.7 – Les valeurs des résidus de χ^2 appliqués à la répartition des femmes et des hommes au sein des catégories d'usage. Pour chaque catégorie d'usage et chaque genre, la barre rouge en haut correspond au panel représentatif, et la bleue du bas au panel complet.

Les partageurs

7% du panel viral - 19% du panel CSA
Les partageurs qui sont les moins représentés parmi nos enquêtés (7%) mais correspondent tout de même à 19% de l'échantillon représentatif se distinguent par un taux de partage très important d'artefacts depuis l'extérieur de Facebook (vidéos, articles, etc.) vers leur page personnelle qu'ils mettent ainsi à disposition d'un cercle restreint, étant le groupe avec le plus petit nombre médian d'amis. Commentant par ailleurs très peu et n'interagissant que rarement sur les pages de leurs amis, on peut se demander si le raccourcissement de la période qui regroupe 80% de leur activité n'est pas le témoin d'une lassitude de Facebook qui interviendrait plus rapidement que pour les autres catégories d'usage.

Les partageurs regroupent des membres du réseau social parmi les moins jeunes, avec une forte sur-représentativité des 45 ans et plus, et notamment des 60 ans et plus, et ce quel que soit le panel observé. Bien entendu, on observe les mêmes particularités concernant les retraités. À l'inverse, les jeunes sont sous-représentés et les moins de 25 ans sont même presque absents de la catégorie, marquant ainsi le fait que cette pratique n'est pas du tout recherchée par ces derniers, peu enclins à l'autopublication.

2.4.2 Publier partout

Deux configurations d'activité similaires dans leur fréquence d'usage mais bien différentes au niveau de la nature de ceux-ci regroupent les usagers de Facebook dont les activités ne sont pas concentrées sur leur propre page. Ces deux catégories regroupent le plus de jeunes de moins de 25 ans, bien que ces derniers ne soient pas les seuls qu'elles concernent.

Les conversants de Groupes

8% du panel viral - 4% du panel CSA
Facebook offre à ses usages la possibilité d'échanger sur des groupes d'utilisateurs. Ceux-ci peuvent regrouper des gens intéressés par des pratiques communes, des membres de mêmes groupes (classe par exemple) ou simplement des amis qui souhaitent partager un canal de discussion commun. Les conversants de groupes, qui sont assez peu nombreux, ont la particularité de concentrer 38% de leur activité au sein de ces groupes contre un maximum de 7% pour les autres catégories d'activité. On peut également imaginer qu'une partie des conversants de groupes utilisent Facebook,

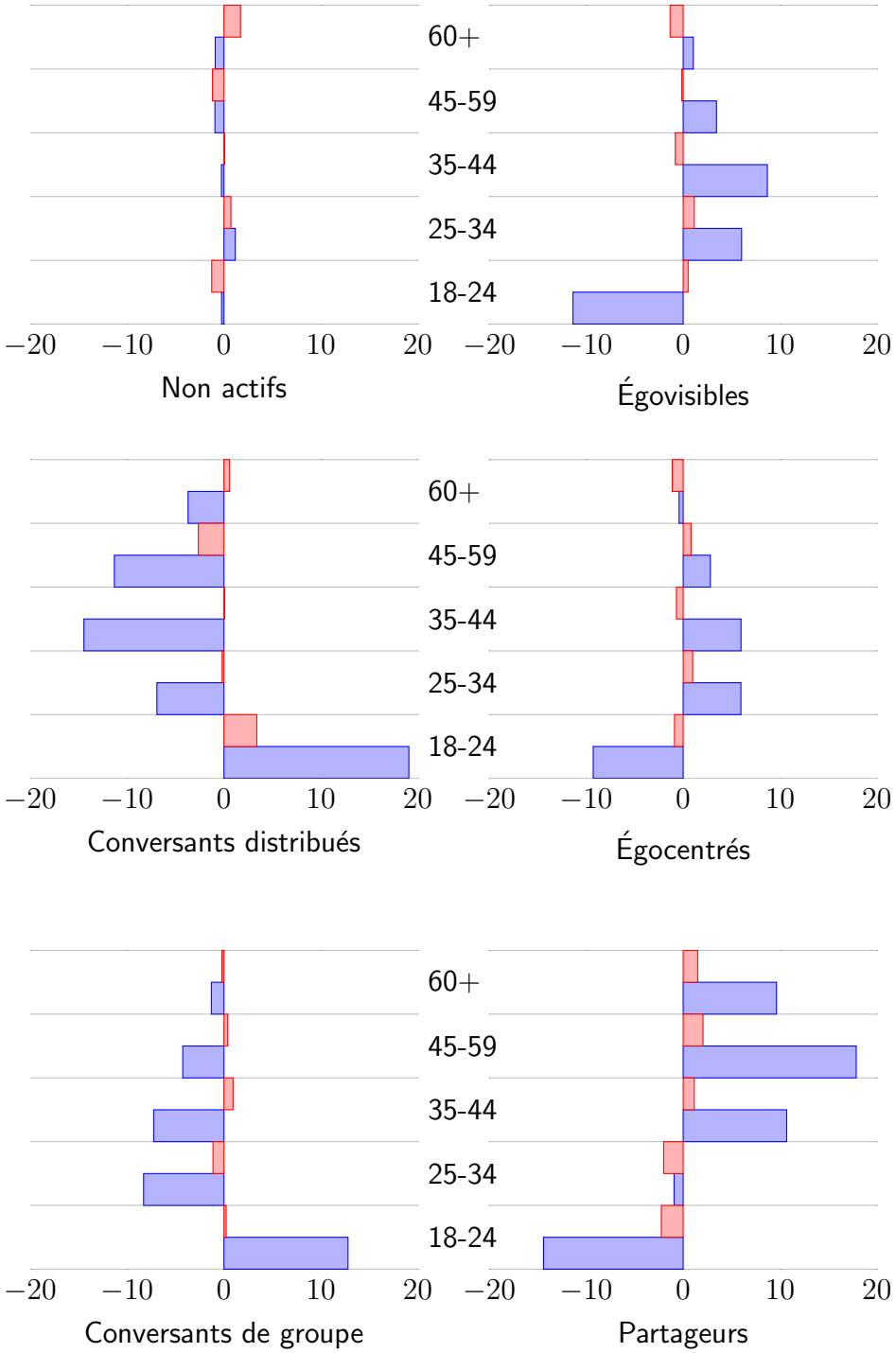


FIGURE 2.8 – Les valeurs des résidus de χ^2 appliqués à la répartition des différentes catégories d'âges au sein des catégories d'usage. Pour chaque catégorie d'usage et chaque catégorie d'âge, la barre rouge en haut correspond au panel représentatif, et la bleue du bas au panel complet.

entre autres, comme un outil collaboratif qui leur permet, par exemple, de communiquer avec des camarades de classe pour un exposé ou des collègues impliqués par le même projet qu'eux. Les conversants de groupes ont également une part plus importante que la moyenne de leur activité qui est située sur les pages de leurs amis, par ailleurs nombreux en moyenne.

Les étudiants sont d'ailleurs nettement sur-représentés parmi les conversants de groupes du panel global. On remarque des dissonances entre les deux panels au niveau des tranches d'âge 35-44 ans et 45-59 ans, ce qu'on peut sûrement expliquer par le côté technophile des enquêtés « viraux », probablement plus enclins à utiliser ce genre d'outils pour communiquer avec leurs collègues ou associés. Pour ce qui est des groupes plus thématiques que collaboratifs, on suppose plus rares ces discussions de passionnés, dont on imagine également qu'elles sont probablement distribuées assez largement au sein des différentes catégories d'âges ainsi que sociales.

Les conversants distribués

23% du panel viral - 22% du panel CSA

Les conversants distribués discutent aussi bien sur leur page personnelle que sur celles de leurs amis, avec une moyenne de 59% de leurs publications qui s'y trouvent. Ils ont un grand nombre d'amis (351 en médiane) qui participent d'ailleurs eux-même régulièrement aux discussions situées sur la page de l'enquêté. Leurs publications incluent rarement des photos ou des liens hypertextes ce qui suggère que leur pratique est majoritairement conversationnelle. Le lien entre ce groupe et la conversation est également souligné par le faible de taux de publications de photos ou de liens hypertextuels de ces enquêtés.

On remarque déjà que les femmes du panel viral sont très sur-représentées dans le groupe des conversations distribuées tandis que ce n'est pas forcément le cas dans le panel représentatif. Les conversants distribués sont encore plus jeunes que les conversants de groupes, avec la plus forte sur-représentation de 18-24 ans dans les deux panels. La plus forte part de 60 ans et plus qu'on y trouve au sein du corpus représentatif suggère également que l'utilisation de Facebook comme simple outil de discussion séduit différentes populations. On peut imaginer que l'utilisation du *chat* de Facebook, bien qu'on ne puisse pas la capturer, est également très utilisée par les membres de ce groupe au sein duquel les classes populaires sont en outre peu représentées.

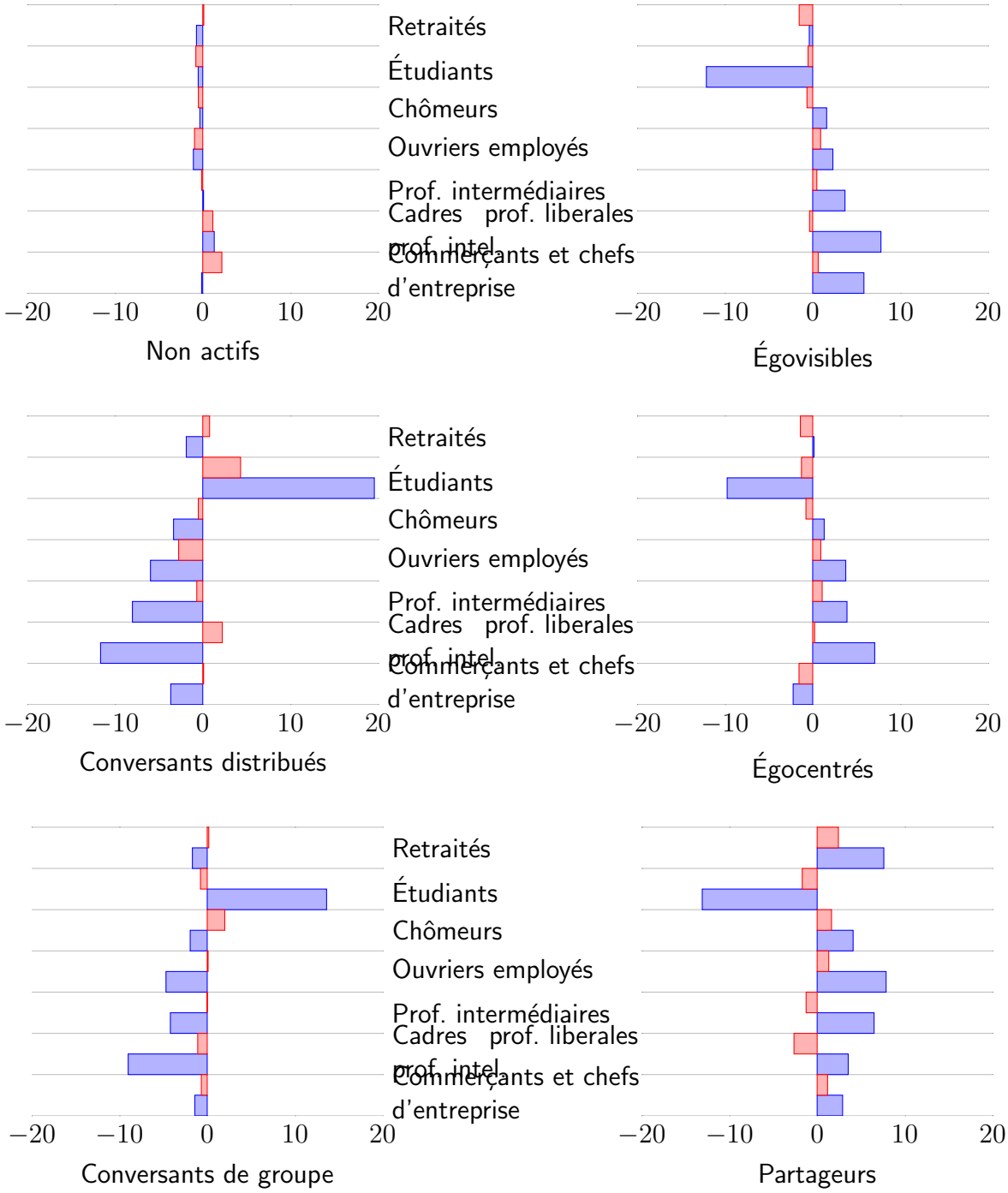


FIGURE 2.9 – Les valeurs des résidus de χ^2 appliqués à la répartition des différentes catégories professionnelles au sein des catégories d'usage. Pour chaque catégorie d'usage et chaque catégorie professionnelle, la barre rouge en haut correspond au panel représentatif, et la bleue du bas au panel complet.

2.4.3 Qui sont les non-actifs ?

17% du panel viral - 3% du panel CSA
Un dernier groupe rassemble les utilisateurs à l'activité publique la moins importante. Ses membres se contentent de répondre mollement aux souhaits de joyeux anniversaires que certains de leurs amis postent sur leur page. On peut d'ailleurs postuler que ces amis partagent des liens faibles avec eux, si on considère que leurs proches préféreront probablement d'autres canaux qu'un message public qui ne correspond pas aux habitudes de leur ami. On remarque que ces enquêtés ont un nombre d'amis légèrement supérieure à la moyenne de l'ensemble du corpus, ce qui indique clairement que cette non participation aux activités publiques de la plateforme ne s'accompagne pour autant pas d'un désintérêt complet de ce qui s'y passe. On est ainsi amené à considérer que ces utilisateurs passent peut-être toujours du temps sur Facebook, à regarder les publications de leurs amis ou à échanger avec eux via le chat.

Les non-actifs ont la particularité, au sein du panel complet, d'être répartis plutôt équitablement entre toutes les classes d'âges et professionnelles. Dans le panel représentatif à l'inverse, les 45-60 ans et les jeunes sont rarement non-actifs sur Facebook, tous comme les étudiants. La présence plus forte des jeunes adultes entre 25 et 34 pourrait être rapportée à la dynamique de ces comptes, dont on peut imaginer qu'il puissent avoir été plus actifs dans des périodes antérieure.

Au cours de ce chapitre, on a montré comment l'application Algopol nous a permis d'obtenir des données anonymes sur un grand panel d'utilisateurs de Facebook. On a également développé les méthodes d'analyses mises en place pour, dans un premier temps, mettre en avant la variété des usages qui sont faits de la plateforme et qu'on a décrits au travers de six catégories.

Les fortes différences concernant les pratiques des activités élémentaires ainsi que les variations de répartition des populations, selon des catégories usuelles de la sociologie, et en particulier les catégories d'âge, laissent à penser que malgré le côté désormais quasi-universel de Facebook, aucune pratique normative n'a réussi à s'imposer.

À l'inverse, chacun, en fonction de son âge, son origine culturelle ou bien d'autres facteurs que la recherche quantitative ne peut pas forcément trancher, tend à profiter de certaines fonctionnalités et à en délaisser d'autres. L'analyse des répartitions des différentes populations dans les catégories d'usage font ressortir une nette variabilité

selon l'âge des enquêtés.

Chapitre 3

Des usages intégrés dans des réseaux

Les catégories d'usages de Facebook que nous avons construites, présentées dans le chapitre précédent, ainsi que les données que le questionnaire soumis aux répondants de l'enquête nous a fournies, vont nous permettre d'explorer comment se structurent les interactions avec le réseau d'amis, et quelles sont les influences réciproques entre usages et interactions.

Nous allons pour cela entreprendre une exploration dans les réseaux personnels mis à notre disposition. Les outils que nous allons employer pour ce faire seront dans un premier temps les indicateurs classiques de la sociologie des réseaux, puis certains indicateurs originaux ayant vocation à capturer l'impact des publications sur le réseau formé par les amis de l'enquêté.

3.1 Des réseaux particuliers

Des réseaux personnels obtenus selon une même méthodologie peuvent avoir des similitudes de structure intéressantes à étudier. Cependant, il ne faut pas perdre de vue que ces similitudes sont justement très dépendantes de la méthodologie employée. En ce qui nous concerne, elles sont directement inhérentes à la manière donc Facebook suggère les ajouts de nouveaux amis à ses usagers.

Les premières enquêtes de la sociologie des réseaux étaient basées sur des *générateurs de noms*, c'est-à-dire sur un protocole d'interview tel que c'est le répondant qui fournit la liste de quelques-unes de ses relations parmi les plus proches ainsi que les liens entre eux. Il ressort de cela que les réseaux ainsi obtenus sont généralement très denses

[Héran, 1988, Rivière, 2000], ce qui est cohérent avec les propositions de Granovetter concernant les triades interdites [Granovetter, 1977], les individus ici étant tous plutôt proches d'*ego*. Des chercheurs ont par ailleurs noté que les liens entre les individus de ces réseaux peuvent parfois être remis en question puisque c'est *ego* qui les décrit et qu'il est possible que ce dernier se trompe en jugeant que deux de ses contacts entretiennent eux-même une forme de relation mutuelle [Milardo, 1988].

Les chercheurs ont essayé de construire ce qui a pris le nom de réseaux d'échanges afin de combler les vides méthodologiques décelés avec la méthode précitée. Le principe est de proposer à l'enquêté une liste d'interactions pour lesquelles ce dernier fournit un ensemble de personnes avec lesquelles il a l'habitude de réaliser l'échange en question. La méthode met généralement en avant des réseaux centrés autour du cœur familial. C'est par exemple le cas de l'enquête de Claude Fischer sur les communautés sociales dans les villes californiennes où 42% de l'ensemble des alters sont décrits comme étant des membres de la famille par *ego* [Fischer, 1982].

La méthode d'enquête que nous avons utilisée se rapproche plutôt de ce qu'on appelle des réseaux d'interaction. Les nœuds et les liens y sont créés via une mesure objective, dans notre cas l'amitié Facebook entre deux individus. L'enquête « Contacts » de l'INSEE en 1986, qui demandait aux répondants de noter et de qualifier l'ensemble des relations qu'ils avaient eues au cours de chaque journée [Héran, 1988], peut être considérée comme la première du genre en France, mais la méthode s'est répandue avec les enquêtes basées sur les contacts téléphoniques [Lambiotte et al., 2008, Stoica et al., 2013, Onnela et al., 2007]. Les réseaux d'interaction, qui agrègent indistinctement tous les contacts de l'enquêté donnent la part belle aux liens faibles [Rivière, 2000].

3.1.1 Comparaison au panel de Caen

Afin d'illustrer ces différences, nous pouvons comparer les valeurs d'indicateurs structurels classiques de la théorie des graphes ou de la sociologie des réseaux entre les réseaux issus de l'enquête Algopol et un groupe de réseaux, dit du panel de Caen, issu d'une méthodologie différente (principalement des générateurs de noms). Le panel de Caen a été construit sous la direction de Claire Bidart dans le cadre d'une enquête qualitative longitudinale débutée en 1995 auprès de jeunes vivant à Caen en Normandie et qui ont été ré-interrogés depuis à plusieurs reprises dans le but d'étudier l'évolution de leurs réseaux sociaux au fil de leur vie.

Si on compare nos réseaux à ceux de Caen (Table 3.1), on remarque déjà qu'ils sont beaucoup plus grands à la fois en nombre de sommets et en nombre de liens. Cette différence est due à leurs différentes méthodes de construction, qu'on a abordées en 1.4.3. Les réseaux de Caen sont en effet beaucoup plus resserrés autour des liens

	Caen		Facebook	
	mean	median	**mean**	median
$\|V\|$	**24**	22	**359**	264
$\|E\|$	**74**	57	**7386**	1965
diameter	**3.18**	3	**7.64**	8
density	**0.28**	0.26	**0.08**	0.06
transitivity	**0.72**	0.73	**0.49**	0.48
betweenness centralization	**0.26**	0.26	**0.20**	0.17
number of communities	**5.47**	5	**7.57**	7
modularity	**0.36**	0.37	**0.50**	0.53

TABLE 3.1 – Les moyennes des valeurs des indicateurs classiques de l'étude des réseaux pour les panels de Caen et de nos enquêtés.

proches de leurs enquêtés. Ce recentrage induit donc naturellement des réseaux à plus forte densité et à plus faible diamètre. La transitivité y est également plus importante, ce qui n'est pas surprenant puisque les réseaux de Caen sont composés des liens forts d'*ego*, qui sont plus reliés entre eux, comme l'a montré, encore une fois, Granovetter.

Les réseaux de notre panel ont par ailleurs une plus faible centralisation d'intermédiarité. Cette distinction, qui indique qu'ils ont moins systématiquement d'alter très central, peut raisonnablement être reliée au fait que ces alters, s'ils existent effectivement dans le réseau de sociabilité de nos enquêtés, ne sont pas pour autant forcément présents sur Facebook, et peuvent ainsi très bien être absents du réseau auquel nous avons accès. D'autre part, cela peut peut-être également s'expliquer par le fait que ces réseaux étant plus petits, il est plus « facile » pour n'importe quel alter de faire le pont entre les différentes communautés, qui sont d'ailleurs beaucoup moins nombreuses. Ces communautés sont également moins bien distinguables que dans le cas des réseaux Facebook, avec une modularité nettement plus faible.

Ces différences structurales entre les réseaux, selon leur origine, interrogent sur la pertinence de prendre les réseaux égocentrés basés sur les amis comme base de l'étude. D'autres méthodes de construction peuvent être monopolisées à partir des données à notre disposition.

3.1.2 Réseaux de commentateurs et likeurs

À la lumière de cette comparaison, on peut s'interroger sur des moyens de rendre nos réseaux moins dépendants de leurs modalités de construction. Une alternative à laquelle nous avons pensé est de considérer uniquement les sommets qui représentent des alters

ayant interagi sur au moins une publication de l'enquêté, via des commentaires ou un like.

Il est également possible de construire des réseaux où le lien entre deux sommets ne représente plus l'amitié mais le fait d'avoir commenté ou liké le même statut. Une des idées derrière cette approche est de se rapprocher encore de la sociabilité en ligne d'*ego*. Selon cette hypothèse, de tels réseaux seraient plus fidèles aux réelles interactions entre les amis de nos enquêtés. La table 3.2 présente les valeurs des métriques classiques qu'on a déjà vues pour ces réseaux d'interactions.

	Commentateurs		Likeurs	
	moyenne	médiane	**moyenne**	médiane
$\|V\|$	**116**	91	**127**	128
$\|E\|$	**853**	384	**5167**	1686
diamètre	**6.07**	6	**3.76**	4
densité	**0.12**	0.09	**0.24**	0.21
transitivité	**0.53**	0.53	**0.59**	0.59
centralisation d'intermédiarité	**0.21**	0.18	**0.1**	0.08
nombre de communautés	**13.9**	11	**9.91**	8
modularité	**0.46**	0.49	**0.2**	0.19

TABLE 3.2 – Les moyennes des valeurs des indicateurs classiques de l'étude des réseaux pour les réseaux Facebook dont on n'a gardé que les commentateurs

Ces réseaux pourraient effectivement sembler plus réalistes puisqu'ils comptent un nombre moins élevé de sommets que les réseaux d'amitiés qui paraissent trop fournis pour n'être constitués que de liens véritablement effectifs. Ces réseaux présentent cependant quelques défauts. Le réseau des likeurs est très dense, et a une transitivité également élevée amenée par le fait que certaines publications, celles relatives aux anniversaires ou aux passages de diplôme par exemple, attirent un grand nombre de likes depuis l'ensemble du réseau.

Les réseaux de co-commentaires ont des valeurs moyennes comme médianes bien plus proches de celles des réseaux d'amitiés et pourraient faire de bons candidats. Les catégories d'usages qu'on a construites plus tôt inspirent cependant la prudence. Il serait en effet dommage d'ajouter aux incertitudes liées à la non présence éventuelle sur Facebook de membres importants de nos réseaux celles relatives à leur non-activité également possible.

Dans la suite, je vais donc continuer à utiliser quelques métriques sur les réseaux d'amitié dont on ne conserve que les alters qui commentent les publications d'*ego* mais

la majorité de mon travail est basée sur les réseaux d'amitié complets. Ces derniers ont en effet l'avantage de ne pas dépendre des usages de Facebook des alters eux-mêmes.

3.2 Interroger les réseaux

Comment combiner une analyse des réseaux à celle des usages ? Voyons si les catégories d'usages que nous avons vues précédemment ou bien les catégories socio-professionnelles de nos enquêtés peuvent être reliées avec des indicateurs de la sociologie des réseaux.

3.2.1 Clés de lecture des réactions du réseau aux publications

La littérature nous fournit d'ores et déjà beaucoup d'indicateurs des caractéristiques des réseaux et leur pertinence est certaine. Il n'y a ainsi pas de raison de nous priver, pour conduire cette analyse, de la prise en compte de quelques mesures classiques, vues en section 1.3.1, pour étudier nos enquêtés : le nombre de clusters de son réseau égocentré, la modularité de ce découpage en clusters, ainsi que celle du réseau induit des alters ayant commenté au moins une publication d'*ego*, la densité de ces deux réseaux, la transitivité, le diamètre et, bien entendu, le nombre d'alters.

On a vu dans le chapitre précédent un premier niveau de lecture, basé sur la qualité des actions réalisées par l'enquêté, qui caractérise son ou ses usages de la plateforme à partir d'activités élémentaires. Un deuxième type d'approche est de prendre en compte les réactions qu'engendrent les publications de l'enquêté au sein des sphères sociales qui forment sont réseau.

La figure 3.1 présente ainsi deux publications ayant reçu un nombre relativement important de commentaires chacune et qui sont catégorisées comme représentant la même activité élémentaire, à savoir *ego publie un statut sur son mur*. Malgré cela, leurs réceptions respectives varient fortement avec, pour la publication de gauche un grand nombre de commentateurs ayant publié chacun un seul commentaire, tandis que celle de droite a plutôt donné lieu à une discussion entre un petit nombre de personnes. Dans cette section, je vais présenter des indicateurs conçus durant mon doctorat pour qualifier la manière dont une publication impacte le réseau égocentré. Ces indicateurs sont calculés soit à partir de l'ensemble des statuts de l'enquêté, soit pour chacun individuellement, auquel cas ils sont agrégés, en prenant la moyenne de leurs valeurs, afin d'obtenir un score directement relatif à *ego*.

Toutes les publications d'*ego* n'ont pas été prises en compte dans le calcul de ces

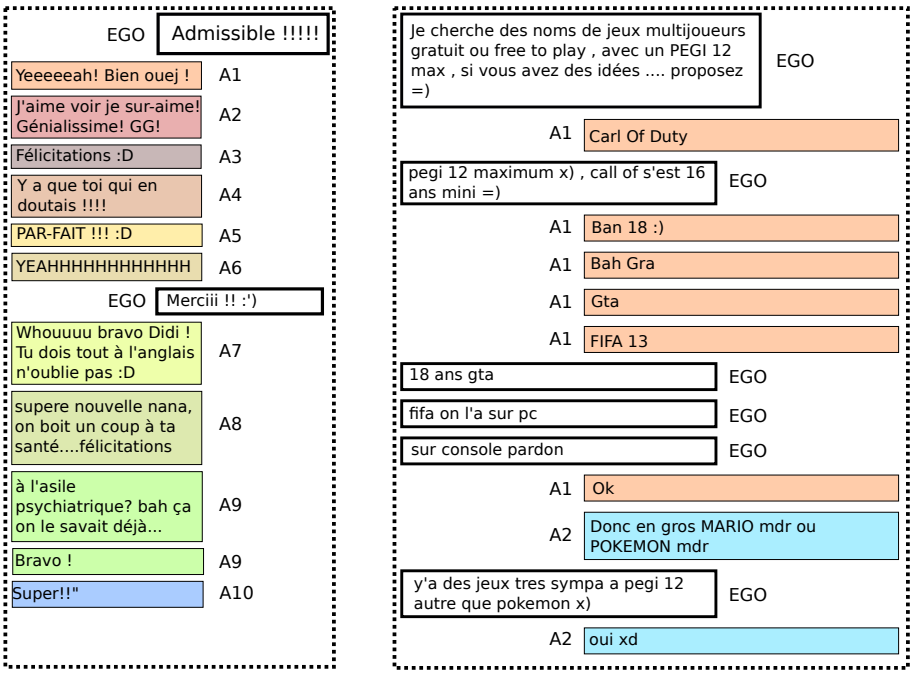


FIGURE 3.1 – Deux publications auxquelles le réseau des enquêtés a réagi différemment, avec beaucoup de commentateurs différents dans un cas et quelques-uns seulement dans l'autre.

Qui	Quoi	Où
Ego	publie un statut publie une photo et un texte publie une photo sans texte publie une vidéo sans texte publie un lien hypertexte et un texte publie un lien hypertexte sans texte publie une photo sur laquelle il identifie quelqu'un	sur son mur
	partage un tweet partage une photo, un album de photos ou une vidéo sans texte partage une photo en ajoutant un texte partage un lien hypertexte partage un statut sans ajouter de texte partage une page Facebook	
	change de photo de profil	*non localisé*

TABLE 3.3 – Table des activités élémentaires utilisées pour les indicateurs liés au réseau

indicateurs car toutes ne sont pas dirigées vers les alters. C'est pourquoi nous avons restreint les activités élémentaires concernées, afin de les réduire à celles qui correspondent bien à des publications d'*ego* (et pas d'un alter), sur sa page (et pas sur celle d'un groupe par exemple). Une publication d'un alter sur le mur de l'enquêté conduirait certainement à un biais concernant les commentateurs de ce statut puisqu'on peut imaginer que les amis communs de cet alter et d'*ego* seraient certainement plus enclins à y répondre et à le liker. La table 3.3 donne la liste de ces activités considérées.

Pour décrire de manière formelle les indicateurs, on utilisera trois ensembles : les statuts, les alters (qui sont les nœuds du réseau d'*ego*) et les commentaires. Pour un statut s donné, on utilisera les notations suivantes :

com(s) l'ensemble des commentaires émis par les alters sur s,

alters$_c(s)$ l'ensemble de ces alters,

alters$_c = \{a \mid \exists s,\ a \in \textbf{alters}_c(s)\}$,
l'ensemble des alters ayant commenté au moins un statut,

alters$_l(s)$ **ou likes**(s) les alters ayant liké s,

alters$_l = \{a \mid \exists s,\ a \in \textbf{alters}_l(s)\}$,
l'ensemble des alters ayant liké au moins un statut.

Notons qu'on choisit de se restreindre aux commentaires et likes provenant d'alters (ils peuvent aussi provenir d'utilisateurs qui ne sont pas amis avec *ego*, et également d'*ego*

lui-même) puisque les indicateurs utilisent le réseau égocentré, exclusivement composé par eux. Par ailleurs, deux notations équivalentes sont définies pour les likes par souci d'harmonie avec les commentaires, étant donné que chaque alter ne peut liker qu'une fois un statut, tandis qu'il peut poster plusieurs commentaires.

L'indicateur **pluralité des commentateurs** défini ci-dessous permet de différencier les réceptions des deux publications de la figure 3.1. La pluralité des commentateurs d'un statut s est le rapport entre le nombre de commentateurs d'un statut et le nombre de commentaires.

$$\textbf{Pluralite}(s) = \frac{|\textbf{alters}_c(s)|}{|\textbf{com}(s)|}$$

Elle vaut 1 dans le cas où chaque commentaire vient d'un commentateur différent et tend vers 0 dans le cas où tous viennent du monologue d'un unique commentateur. En pratique, cet indicateur est intéressant lorsque le statut a reçu au moins deux commentaires, son rôle étant de détecter si des conversations se sont engagées entre quelques amis d'*ego* ou bien si chacun est simplement venu apporter son commentaire. Dans l'exemple de la figure 3.1, le statut de droite n'a que 2 commentateurs, soit une pluralité de $2/8 = 0.25$. À l'inverse, celui de gauche, comme lancé à la cantonade, reçoit surtout des commentaires isolés et a un score de pluralité de $10/11 = 0.91$. Il est d'ailleurs amusant de noter que la seconde réponse du seul commentateur en ayant posté deux correspond à un changement de ton montrant l'ironie du premier.

On veut également savoir à quel point les publications de nos enquêtés touchent l'ensemble du réseau, comme on pourrait l'imaginer dans le cas d'un individu postant de nombreux messages à caractère politique par exemple, ou bien s'ils ont plutôt tendance à les séquencer en ciblant à chaque fois spécifiquement une des communautés de leurs contacts, comme le ferait une photo de groupe en vacances. Une série d'indicateurs joue ce rôle.

Le premier d'entre eux, le **taux d'inter-connaissance des commentateurs** indique à quel point les commentateurs des statuts d'*ego* se connaissent localement entre eux. Pour le calculer, on construit pour chaque statut le sous-réseau d'amitié induit par les commentateurs de ce statut puis on calcule la proportion de sommets non isolés dans ce réseau.

En notant $G_s = G|\textbf{alters}_c(s)$ le sous-graphe induit aux commentateurs d'un statut s, le taux d'interconnaissance est défini formellement comme suit :

$$\textbf{Interconnaissance}(s) = 1 - \frac{|\text{isolés}(G_s)|}{|V(G_s)|}$$

Cette valeur est comprise entre 0, lorsque tous les sommets du graphe induit par les commentateurs sont isolés et 1, lorsque ce dernier est connexe. On peut lier cette

métrique à l'analyse de la légitimité ressentie par les alters à commenter un statut qui ne leur serait pas destiné. Une publication à faible interconnaissance indiquant qu'elle ne demande pas de légitimité particulière (comme le fait d'appartenir à un groupe ciblé par elle) pour réagir.

La **distance entre les commentateurs** s'appuie elle, non pas sur un sous-graphe induit, mais sur l'intégralité du réseau égocentré. C'est la moyenne des distances entre chaque couple de sommets représentant les commentateurs du statut. Formellement :

$$\textbf{Distance}(s) = \frac{\sum_{a_1,a_2 \in \textbf{alters}_c(s)} d(a_1, a_2)}{|\textbf{alters}_c(s)| \times (|\textbf{alters}_c(s)| - 1)}$$

Notons qu'on prend ici comme distance entre deux alters appartenant à deux composantes connexes différentes la valeur $\text{diam}(G) + 1$, où G est le réseau égocentré considéré. Une faible distance entre les commentateurs d'un statut indique probablement que le statut visait directement un groupe d'alters spécifiques, par une photo de groupe ou une anecdote partagée par exemple. Au niveau agrégé c'est cette tendance à cibler une communauté d'alters proches entre eux, sans pour autant que cela ne soit toujours la même selon les statuts, qu'on veut capturer avec cette métrique.

Nous avons également sélectionné parmi des indicateurs généralement utilisés dans d'autres domaines pour les utiliser sur nos données. C'est le cas pour un indicateur que nous appelons la **concentration des commentateurs**, défini comme le coefficient de Gini du nombre de commentaires par commentateur, non pas cette fois sur un statut, mais pour l'ensemble des publications de l'enquêté. Le coefficient de Gini, du nom de son inventeur [Gini, 1921], est une mesure généralement utilisée en étude des sociétés qui indique à quel point les richesses sont réparties équitablement ou pas. Elle procède en comparant la courbe cumulative des richesses (le nombre de commentaires dans notre cas), qu'on nomme la courbe de Lorenz, à la diagonale, droite cumulative qui serait obtenue dans le cas d'une égalité parfaite. Pour cela on divise l'aire comprise entre ces deux courbes par l'aire sous la diagonale, selon la formule :

$$\textbf{Concentration} = \frac{A - A_L}{A}$$

où A_L est l'aire sous la courbe de Lorenz, soit l'aire bleue de la figure 3.2 et A l'aire sous la droite pointillée, correspondant à la somme des aires grise et bleue, soit la moitié de l'aire du carré.

La concentration des commentateurs varie ainsi entre 0 dans le cas où les commentateurs ont tous publié le même nombre de commentaires et 1 dans le cas où les commentaires viendraient tous du même alter. Notons que nous avons décidé de ne

pas considérer dans ce calcul les amis d'*ego* qui n'ont jamais commenté, afin d'éviter que les inactifs pèsent trop sur le résultat.

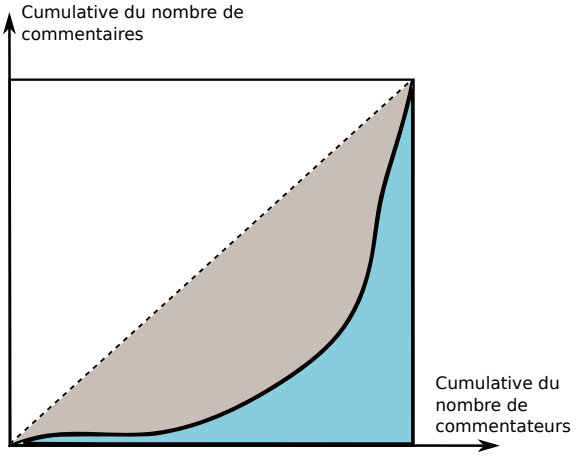


FIGURE 3.2 – Schéma du calcul du coefficient de Gini

La **vraisemblance des likeurs** et la **vraisemblance des commentateurs** partent des communautés d'alters obtenues grâce à l'algorithme de Louvain. Pour chaque communauté C, qui est en fait un sous-ensemble des alters du réseau égocentré $G = (V, E)$, et pour chaque statut s, on calcule le nombre de likes, qu'on note $n_l(C, s)$, et de commentaires, qu'on note $n_c(C, s)$, reçus par s qui proviennent des alters de C. Notons maintenant $C(a)$ la communauté à laquelle appartient un alter a donné. Ces nombres valent alors :

$$n_c(C, s) = |\{a \in \textbf{alters}_l(s) \,|\, C(a) = C\}|$$
$$n_l(C, s) = |\{a \in \textbf{alters}_c(s) \,|\, C(a) = C\}|$$

On obtient ainsi, pour chaque statut s, un vecteur

$$n_l(s) = (n_l(C_1, s), n_l(C_2, s), ...)$$

représentant le nombre de likes par communautés, et

$$n_c(s) = (n_c(C_1, s), n_c(C_2, s), ...),$$

son équivalent pour les commentaires. Leur taille est donc égale au nombre de communautés détectées au sein du réseau. On calcule par ailleurs la proportion totale des likes $P_l(C)$ et des commentaires $P_c(C)$ reçus par les statuts d'*ego* provenant de chaque communauté de son réseau. Cette proportion est définie pour une communauté C par la formule :

$$P_l(C) = \frac{\sum_{\text{statut } s} |\{a \in \textbf{alters}_l(s) \,|\, C(a) = C\}|}{\sum_{\text{statut } s} |\textbf{likes}(s)|}$$
$$P_c(C) = \frac{\sum_{\text{statut } s} |\{a \in \textbf{alters}_c(s) \,|\, C(a) = C\}|}{\sum_{\text{statut } s} |\textbf{com}(s)|}$$

On obtient encore deux vecteurs

$$P_l = (P_l(C_1), P_l(C_2), ...)$$

représentant cette fois les proportions de chaque communauté pour l'ensemble des likes et

$$P_c = (P_c(C_1), P_c(C_2), ...)$$

représentant les proportions de chaque communauté pour l'ensemble des commentaires. Une fois ces proportions capturées, on peut exprimer la vraisemblance d'un statut comme la probabilité de répartition de ses commentateurs ou likeurs selon la *loi multinomiale*, une loi probabiliste qui nous permet de comparer ces proportions à la proportion globale sur l'ensemble des statuts, comme si elles étaient des variables aléatoires.

Formellement, pour le cas des likes, à adapter pour les commentaires, si n est le nombre de likes et m le nombre de communautés détectées dans le réseau personnel, la formule finale est la suivante :

$$\textbf{Vraisemblance likes}(s) = \frac{n!}{\prod_{i\in[1,m]} (n_l(C_i, s)!)} \prod_{i\in[1,m]} (P_l(C_i)^{n_l(C_i,s)})$$

Un statut a donc une faible vraisemblance des commentaires si les communautés d'où proviennent ses commentaires sont peu habituelles. La figure 3.3 présente deux statuts d'un même enquêté dont l'un a une forte vraisemblance et l'autre une faible. Une fois les valeurs de ses statuts agrégées, un individu a une faible vraisemblance si les réponses à ses publications fluctuent fortement d'un statut à l'autre tandis qu'elle est forte dans le cas où il possède un noyau dur de commentateurs ou likeurs qui interagissent régulièrement à travers ses publications.

On peut noter qu'il est ardu de rapprocher une forte ou faible valeur de vraisemblance d'un type de publication particulier en ce sens qu'elle est très dépendante de l'ensemble des statuts de l'enquêté. Si ce dernier publie généralement des posts à vocation humoristique, on imagine qu'une publication d'ordre politique trouverait une audience

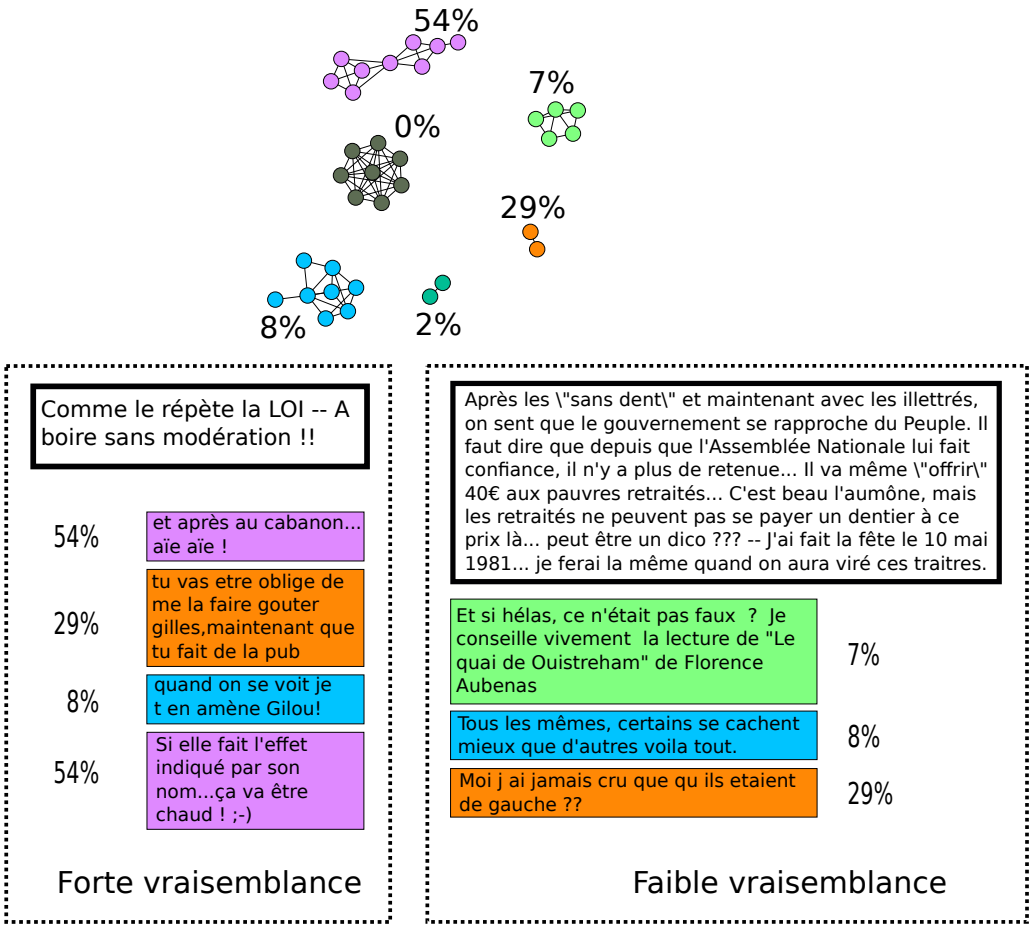


FIGURE 3.3 – Les réactions à deux statuts d'un même enquêté, projetées sur les communautés de son réseau. Celui de gauche a une forte vraisemblance, les commentaires provenant des communautés habituelles tandis que celui de droite, un statut lié à l'actualité politique du moment, soulève des commentaires issus de communautés plus inhabituelles.

inhabituelle et aurait ainsi une faible vraisemblance mais il en va de même pour une publication ayant pour but d'amuser, proposée par un individu postant majoritairement des avis politiques.

Finalement, les **likeurs exclusifs** représentent la part des alters qui ont liké au moins une publication mais n'en ont jamais commenté une seule. Nous considérons que le commentaire demande un investissement plus important de la part des alters et on imagine que certains usages feront apparaître des différences dans ce sens. La métrique est définie comme suit :

$$\textbf{Likeurs exclusifs}(s) = \frac{|\textbf{alters}_l \setminus \textbf{alters}_c|}{|\textbf{alters}_l|}$$

3.2.2 Des réseaux qui confortent la variabilité des usages

Voyons comment ces indicateurs se distribuent selon les différentes catégories, socio-professionnelles ou d'usage auxquelles on peut les confronter.

Pour ce faire on utilise de nouvelles matrices dont les valeurs sont définies de la même façon que celles de la section 2.3 .

Confrontation aux catégories socio-professionnelles

Au regard de la figure 3.4, nous retrouvons dès à présent un premier résultat classique de la sociologie des réseaux en comparant l'âge des enquêtés et les indicateurs de leurs réseaux, à savoir que les jeunes ont des réseaux denses [Bidart and Lavenu, 2005, Pasquier, 2005, Kalmijn, 2012]. Cela s'explique par le plus grand nombre de cercles sociaux que l'on est naturellement amené à fréquenter au fil du temps, puisqu'avec chaque nouveau cercle social, c'est un groupe d'alters qui s'ajoute potentiellement au réseau, et que ce dernier a peu de chances d'avoir beaucoup de connexions avec la famille ou les groupes de connaissances faites par *ego* au cours de ses études. Outre la baisse de la densité et de la transitivité observée avec l'augmentation en âge, on note également que la modularité, synonyme de découpage du réseau en groupes bien définis, augmente lors du passage de la catégorie des moins de 25 ans à celle des 25-34 ans, selon un phénomène déjà bien documenté [Bidart et al., 2011]. La densité des réseaux est souvent anti-corrélée avec leurs tailles. En effet, lorsqu'un nouvel alter le rejoint, il ne sera généralement relié qu'avec une portion de ceux déjà présents, faisant donc mécaniquement baisser la densité. Or ici, les jeunes ont également les réseaux avec le plus d'alters, bien que de peu, et on fait donc probablement face à une

tendance forte. Il convient cependant de tenir compte du fait que les jeunes sont sur-représentés sur la plateforme, ce qui implique certainement que d'autres méthodes de production de réseaux sociaux n'aboutiraient sans doute pas à la même conclusion. Les jeunes semblent avoir un noyau dur peu étendu, comme le suggèrent la faible distance et la forte concentration, assez fortement interconnecté de commentateurs et likeurs, réagissant rapidement à leurs publications. Cela amène les autres amis de nos jeunes enquêtés à privilégier le like au commentaire, nécessitant sûrement moins de légitimité et impliquant une vraisemblance des likes plus faible que celle des commentaires et un taux de likeurs exclusifs importants.

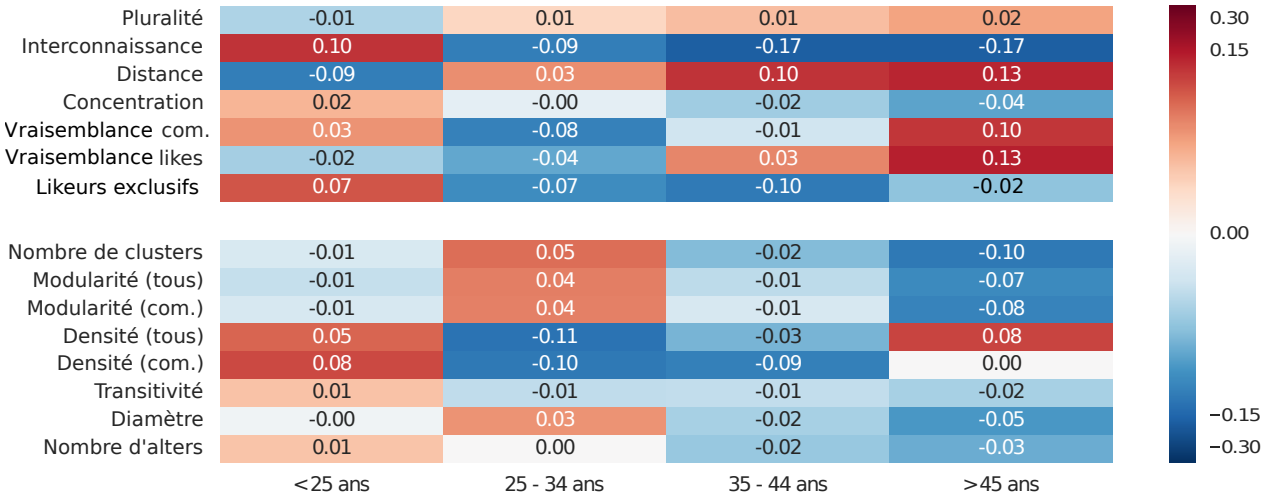


FIGURE 3.4 – Le lien entre les indicateurs de réseau avec les catégories d'âges.

À l'inverse, le fait que les personnes de plus de 45 ans aient également un réseau d'amis dense pourrait être lié à une sous-représentation de cette catégorie d'âge sur la plateforme ou bien à un recentrage, par ses représentants, de leur sociabilité autour de leurs contacts les plus proches et de leurs liens forts. Ce recentrage se traduit également par le faible nombre de communautés que contiennent leurs réseaux ainsi que par leur faible diamètre. La faible transitivité suggère que les réseaux ne sont pas constitués de groupes bien dessinés (ce qu'indique la faible modularité) mais plutôt de groupuscules faiblement reliés entre eux. Ces derniers commentent et likent de manière erratique les publications d'*ego*, amenant à des conversations auxquelles participent des alters qui ne se connaissent pas entre eux.

Le temps moyen de réaction aux statuts, très dépendant de l'âge des répondants, ne peut pas s'expliquer uniquement par une taille du réseau qui diminue avec l'âge des enquêtés et il semble qu'en moyenne, les jeunes ont une activité beaucoup plus intense sur Facebook que les autres catégories d'âge. Notons cependant que ces réactions

dépendent bien de l'âge des membres du réseau et non pas de celui d'*ego* lui même, âges que nous ne connaissons pas. L'indicateur suggère donc, de manière attendue, que l'on a tendance à avoir plus d'amis d'âges proches du nôtre. Il serait par ailleurs sûrement pertinent de regarder comment ces temps de réaction moyens évoluent en fonction de moments de la journée, ce qui pourrait éventuellement participer à une estimation de l'âge des différents membres du réseau de nos enquêtés.

Notons que les indicateurs de structure des réseaux convergent vers l'idée que la période entre 25 et 34 ans semble correspondre, par rapport aux autres catégories d'âge proposées ici, à un pic de la sociabilité. Ces jeunes enquêtés ont en effet des réseaux avec un plus grand nombre de communautés, par ailleurs bien distinguables, et un diamètre important. Ce sont donc ceux qui fréquentent le plus de cercles sociaux différents et la décroissance de ces valeurs qu'on observe après 35 ans peut découler de plusieurs facteurs chronophages comme un investissement plus important dans la carrière professionnelle ou encore le phénomène bien connu de la naissance du premier enfant [Manceron et al., 2002].

La pluralité des commentateurs croissante en avançant dans les catégories d'âge témoigne d'une différenciation avec un usage de discussion en ligne entre connaissances plus intensément porté par les jeunes, tandis que les publications des plus âgés sont plus expressives et tournées vers l'ensemble de leur réseau, et également marquées par une distance plus importante entre les commentateurs. On retrouve ici les tendances qu'on a pu déceler via la répartition des classes d'âge au sein des catégories d'usage construites au chapitre précédent.

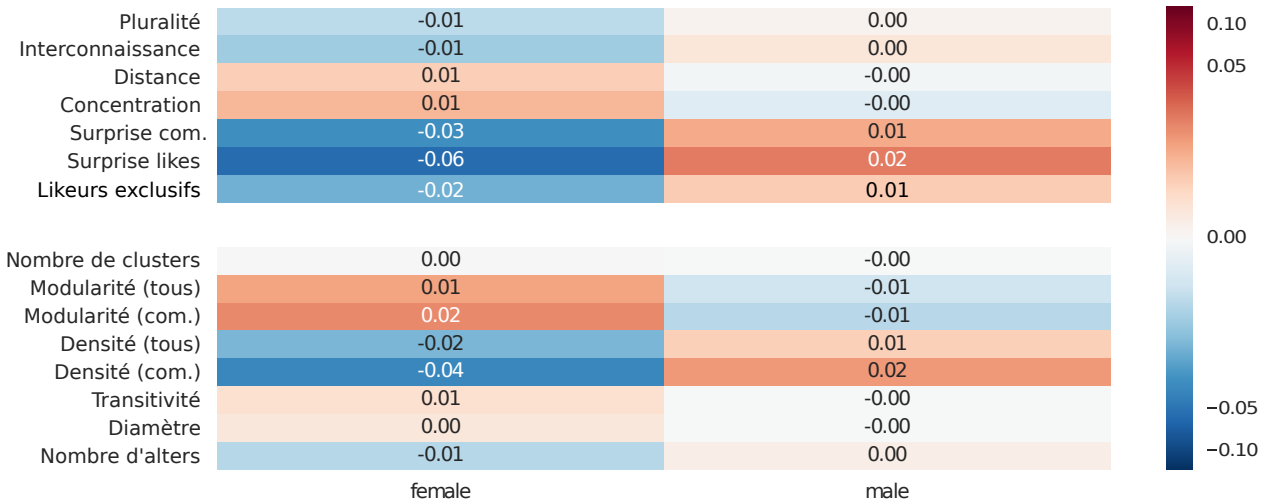


FIGURE 3.5 – Le lien entre les indicateurs de réseau selon le genre des enquêtés.

C'est également la conversation qui différencie les usages de la plateforme entre femmes et hommes, même si les variations entre les deux catégories demeurent du domaine de la nuance, comme le montre la figure 3.5. Les hommes ont en effet, en moyenne, une part moins importante de leur réseau qui participe aux conversations sur leur profil, correspondant également à un nombre moins élevé de communautés commentantes mais aussi une densité du réseau de leurs amis commentateurs plus importante et une vraisemblance des commentateurs plus importante qui traduit une audience moins diversifiée, comme resserrée autour d'un noyau d'amis.

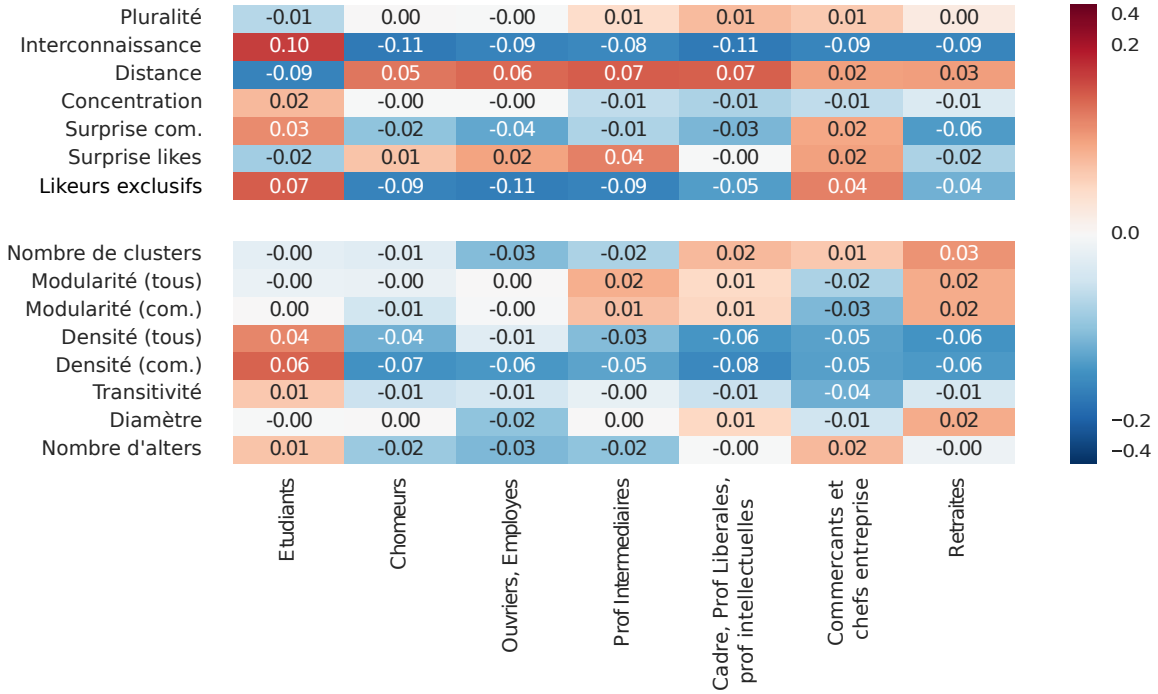


FIGURE 3.6 – Le lien entre les indicateurs de réseau selon les catégories professionnelles.

Les catégories professionnelles, dont les liens avec les indicateurs sont présentés en figure 3.6, marquent une différenciation forte entre les étudiants, aux comportements probablement plus homogènes, et les autres classes. Cette spécification provient du fait que les valeurs de corrélations de nos indicateurs pour les étudiants épousent celles de la catégorie des jeunes, dont on a déjà noté les particularités. On voit que le découpage par types de profession n'est pas aussi intéressant que celui par catégories d'âges, de loin le plus pertinent parmi les catégories socio-professionnelles.

Les catégories d'usages confortées par l'analyse des réseaux

Le croisement entre les catégories d'usages définies dans le chapitre précédent et les variables de réseau permet de comprendre l'articulation entre usage de la plateforme et interactions avec le réseau de ses contacts. Les nombres d'amis ainsi que de publications, qui diffèrent selon les catégories, font apparaître des variations importantes de ces indicateurs, qui sont présentées dans la figure 3.7.

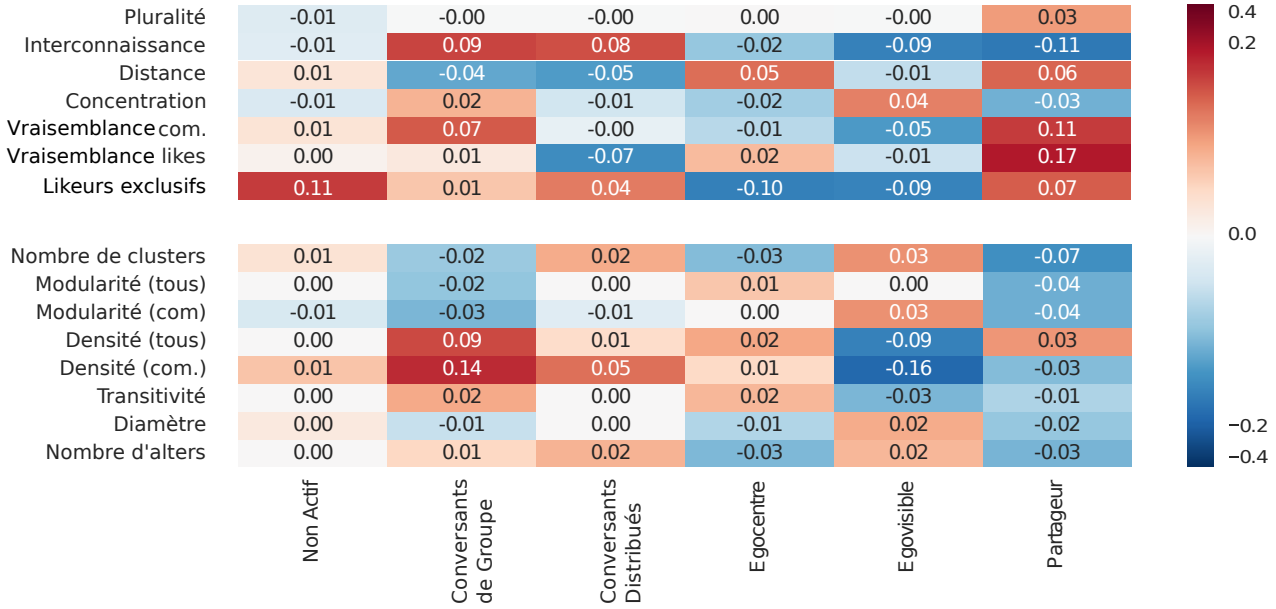


FIGURE 3.7 – Les valeurs moyennes des indicateurs de réseau selon la catégorie d'usage.

Les égocentrés dont on soupçonne qu'ils représentent l'usage typique peu expansif de Facebook, ont peu de clusters ou communautés dans leur réseau de commentateurs, qui ont également un faibles diamètres, ce qui correspond parfaitement à leur taux de publication peu élevé. On note qu'ils ont également un faible nombre de likeurs de leurs statuts puisque leurs amis qui ne commentent pas leurs publications ne les likent pas beaucoup plus, une preuve supplémentaire du peu d'attrait de ce groupe pour la recherche de la visibilité de soi. Cette catégorie d'usage est par ailleurs celle dont la distance moyenne des commentateurs est la plus élevée, ce qui indique que leurs publications sont peu ciblées par rapport aux communautés de leur réseau, d'autant plus que la faible concentration de leurs commentaires montre une participation plutôt homogène de leurs alters.

Le groupe des égovisibles est celui dont le nombre de publications, ainsi que celui d'amis, est le plus important du panel. La forte participation de ces individus entraîne

des réactions pouvant provenir des différentes zones de leurs réseaux, comme l'indiquent les faibles vraisemblances, aussi bien de leurs commentaires que de leurs likes. La concentration des commentaires élevée de ces enquêtés dénote cependant que cet auditoire réactif ne représente qu'une partie réduite de leur réseau. Si elle est réduite, elle demeure néanmoins répartie au sein des différentes communautés d'amis puisque la distance entre les commentateurs est forte et que leur interconnaissance est, elle, relativement faible.

À usage spécifique, interactions spécifiques avec le réseau : les publications des partageurs font généralement intervenir les mêmes contacts, d'un noyau particulièrement resserré, ce que montrent les vraisemblances fortes des commentateurs comme des likeurs. Les partageurs ont peu d'amis associés à leur compte Facebook et un réseau réduit et centré autour de leurs proches, ce que traduit un faible diamètre et une densité forte. La transitivité peu élevée tend cependant à montrer que ces réseaux ne sont pas construits autour d'un noyau d'interconnaissances mais bien par différentes sphères sociales qui vont réagir sporadiquement, le plus souvent en se contentant de liker et sans entrer dans une discussion, comme le montre la pluralité élevée des commentateurs, quand l'objet partagé les touche spécifiquement.

Les conversants de groupes qui, on l'a vu, sont plutôt jeunes, ont un nombre d'amis relativement important, concentrés dans peu de sphères sociales, induisant des réseaux denses et de faible diamètre. La densité encore plus importante du réseau constitué par leurs amis commentateurs, ainsi que la forte interconnaissance entre ceux-ci, traduisent un auditoire resserré autour des proches. Cela suggère que la majorité des groupes Facebook que fréquentent nos enquêtés sont des groupes d'interconnaissances plutôt que des groupes d'intérêts communs. La jeunesse relative des conversants de groupes, détaillée au chapitre précédent, encourage à imaginer que nos enquêtés catégorisés ici fréquentent des groupes formés autour de classes universitaires par exemple.

Avec un profil assez similaire, les conversants distribués ont un réseau très fourni mais avec une densité et une transitivité fortes, associées, comme vu précédemment à la jeunesse des membres qui forment cette catégorie. À l'instar des conversants de groupes, on note un ciblage important de chacune de leurs publications qui engendrent une interconnaissance forte des répondants et une distance faible. La faible vraisemblance de leurs likeurs peut également être associée à cela si ces derniers se contentent plus aisément à occasionnellement liker des statuts qui ne les ciblent pas particulièrement.

Les non-actifs sont non actifs. Ils ont donc peu de commentateurs, leurs amis se contenant de liker leurs quelques publications, et un auditoire, pour utiliser un terme quelque peu disproportionné ici, peu réactif. Si on peut extraire quelque chose d'intéressant des réseaux de ce groupe d'individus, c'est bien qu'aucun indicateur structurel n'est particulièrement important ou faible pour eux, ce qui montre qu'il n'y a pas une forme

de réseau social sur Facebook qui soit liée au fait de ne pas utiliser les outils de communication publics de la plate-forme.

Les configurations dégagées de la structure des activités des utilisateurs invitent à garder la trace de l'opposition entre deux lignées de pratiques qui se déploient à la suite des différents médias qui ont jalonné l'histoire du web [Cardon and Prieur, 2016]. Les conversants, de groupes et distribués, entretiennent ainsi des pratiques d'échanges interpersonnels qu'on pourrait rapprocher de celles des messageries instantanées tandis que les partageurs, égocentrés et égovisibles ont des usages plus proches de l'auto-publication des blogs ou des sites personnels. Bien que Facebook ambitionne de fusionner ces deux catégories de pratiques dans un nouveau contexte, ces dernières y sont toujours distinctement perceptibles.

3.2.3 Vers une catégorisation des publications

Jusqu'à présent, les indicateurs de la réponse du réseau aux statuts n'ont été étudiés que sous une forme agrégée afin de qualifier l'enquêté qui les a publiés. Une suite de ce travail, entreprise par Dominique Cardon, Jean-Philippe Cointet et Paige Camerino, vise à qualifier les statuts eux-mêmes.

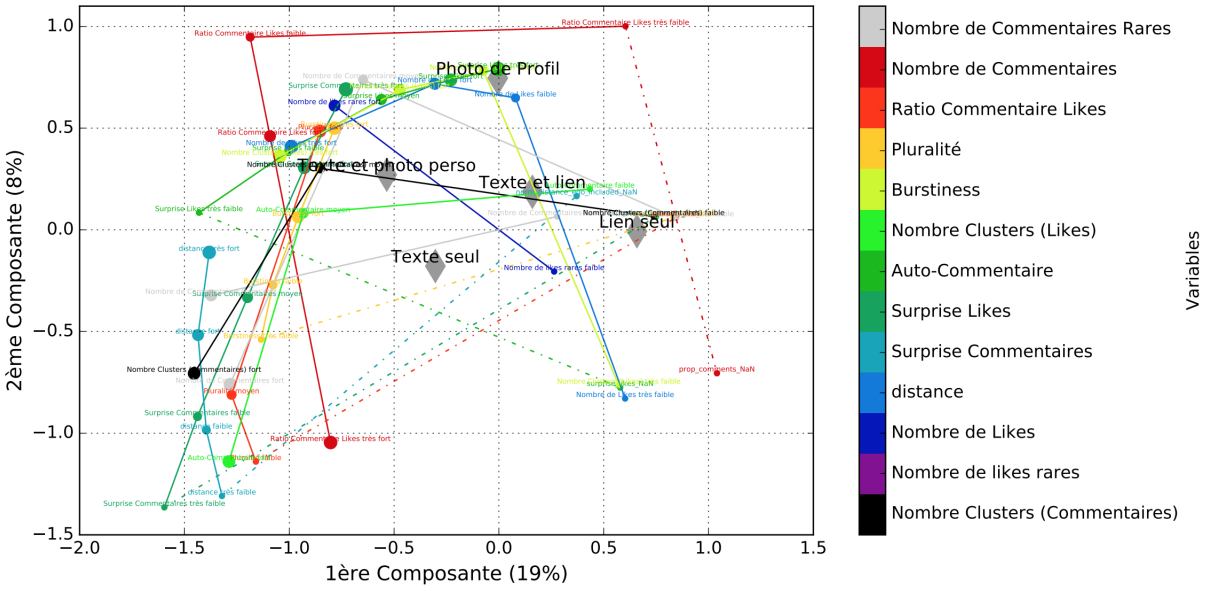


FIGURE 3.8 – Les deux premiers axes d'une analyse factorielle appliquée aux indicateurs de statut.

La figure 3.8 dénote le potentiel de cette recherche puisqu'on détecte que certaines

régions agglomèrent les statuts ayant par exemple un fort nombre de commentaires mais peu de pluralité. Le premier axe, horizontal, oppose ainsi les statuts à caractère public à ceux plus contextuels tandis que le second différencie les contenus avec une photo et ceux qui n'en ont pas.

Les points correspondent aux positionnements moyens, non pas des statuts, mais des hautes ou faibles valeurs des indicateurs. Par exemple, le point correspondant aux nombres élevés de commentaires, qui est le gros point rouge en bas à gauche de la figure, se trouve placé dans la zone des contenus publics sans photos.

Encore au stade embryonnaire, ce projet vise à étudier comment les publications de différentes natures (sujets politiques, photos de vacances, etc.) sont reçues par les alters et les engagent dans la discussion.

3.3 La nature du lien social

Depuis les travaux de Mark Granovetter sur la force des liens, évoqués dans la section 1.4.2, les sociologues des réseaux n'ont eu de cesse de tenter de qualifier ces liens, constituants des réseaux sociaux. Le cas des réseaux personnels offre à cet égard une clé de lecture légèrement différente : il ne s'agit pas ici de s'intéresser au lien entre deux alters, toute étude à ce sujet étant par nature biaisée par la relation qu'entretient *ego* avec eux. En effet, supposons par exemple que deux alters du réseau de l'un de nos enquêtés soient frère et sœur, mais que ce soient les deux seuls représentants de leur famille dans le réseau, par exemple dans le cas où ils étaient dans la même classe que l'enquêté à un moment donné de sa scolarité. Aucune conclusion pertinente ne pourrait être amenée par la lecture de leur position dans le réseau, où il manquerait le grand nombre de liens qu'ils partagent mais qui sont invisibles à *ego*. Non, il s'agit bien ici de caractériser seulement le lien qui relie *ego* à ses alters à travers leur position dans son réseau.

La qualification de certains alters que nous avons demandé à nos enquêtés de faire va ainsi nous permettre d'explorer quelques pistes de recherches concernant le lien entre force relationnelle et position au sein de la structure du réseau.

3.3.1 Les alters qualifiés

En suivant le questionnaire de l'enquête Algopol, les répondants devaient qualifier au moins cinq des amis de leur réseau : jusqu'à quatre (possiblement moins dans le cas de doublons) d'entre eux étaient les deux amis ayant le plus commenté sur le

mur de l'enquêté et les deux ayant le plus liké. Pour arriver à cinq, s'ajoutaient des alters parmi ceux ayant le plus d'amis communs avec le répondant. Ce dernier était également libre, de qualifier d'autres de ses amis quand il le souhaitait, via sa page dédiée de l'application. Nous avons finalement eu accès à un ensemble de 61 793 amis qualifiés.

— 2 alters ayant le plus commenté
— 2 alters ayant le plus liké
— 1 à 3 alters aux plus hauts degrés

Qualifier un ami revient à décrire la nature de la relation qu'on entretient avec lui selon un ou plusieurs choix : une connaissance, un ami, un membre de la famille ou bien un collègue de travail ; à lui donner un score de proximité affective, à quantifier la fréquence des contacts en face à face et des contacts médiés (téléphones, messagerie instantanée et autres) ainsi qu'à décrire son ancienneté. Un champ de texte libre était également mis à disposition des enquêtés qui pouvaient ainsi, s'ils le souhaitaient, décrire plus précisément une relation.

3.3.2 Des mesures structurelles et d'interactions difficilement conciliables

Dans un premier temps, on s'intéresse aux relations qu'entretiennent certains indicateurs de réseau et d'usage de ces alters qualifiés. Comme toute méthode, on a vu que l'extraction de réseaux via Facebook entraînait un certain nombre de biais qu'on souhaite pouvoir explorer ici.

En analysant la matrice des corrélations, présentée en figure 3.9, sur l'ensemble des qualifiés, on remarque immédiatement que ces variables sont toutes plutôt corrélées entre elles, à l'exception du degré qui est, comme on le suggérait déjà en section 1.2, anti-corrélé avec certaines autres métriques. Certains groupes de variables montrent au contraire des corrélations fortes en leur sein. Le degré et la centralité d'intermédiarité, les deux variables de structure du réseau, sont ainsi corrélées entre elles. Idem pour les variables de fréquence d'interactions : nombre de likes et de commentaires ainsi que pour les variables de fréquence de contacts, à distance et face à face. On voit néanmoins dès à présent qu'il est plus difficile d'établir des liens entre ces groupes.

Le degré est donc un peu particulier ici car il n'est corrélé qu'à l'autre variable de structure des réseaux et avec la fréquence des contacts en face à face. On peut en partie expliquer cela par un nouveau biais de recrutement. Les alters qualifiés ont en effet été sélectionnés parmi ceux ayant le plus interagi avec *ego* ou bien ceux ayant le plus d'amis communs avec lui. Il est donc clair que dans ce lot d'alters, ceux avec le plus fort

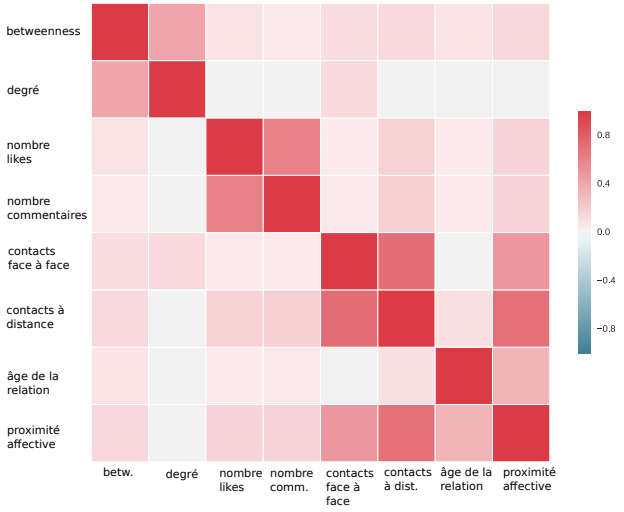


FIGURE 3.9 – Matrice des corrélations entre quelques indicateurs sur les amis qualifiés

degré ne sont généralement pas ceux ayant le plus d'interaction avec l'enquêté et vice versa. Ces mesures sont donc naturellement anti-corrélées. La fréquence des contacts en face à face augmente cependant, elle, légèrement avec le degré des sommets et les alters qualifiés pour leur degré important ont donc moins de chance d'être des amis réellement proches de nos enquêtés que des gens qu'ils fréquentent régulièrement. On peut alors imaginer que ce sont peut-être des camarades de classe ou des collègues de travail populaires et avec lesquels ils ont alors beaucoup de contacts communs sur Facebook.

Le groupe des variables d'usage de Facebook, c'est-à-dire le nombre de likes et le nombre de commentaires, dénote également une corrélation forte, ainsi que des corrélations similaires avec les autres indicateurs présentés. Les alters qui publient le plus likent également le plus parmi les alters qualifiés. On retrouve encore une tendance forte au fait que ces alters ne sont pas nécessairement les plus proches affectivement de l'enquêté puisque ces variables sont peu corrélées avec celles qui représentent la force de la relation, telle que décrite par *ego*. La légère tendance de ces variables à augmenter avec le nombre de contact à distance suggère que les gens qui sont à distance utilisent plus Facebook pour interagir entre eux que s'ils se voyaient régulièrement. Ces deux indicateurs sont néanmoins ceux qui, mis à part les indicateurs de structure du réseau, sont les moins corrélés avec la proximité affective.

Les qualifiés qui ont été déclarés par l'enquêté comme ayant beaucoup de contacts

à distance avec lui sont généralement plus proches affectivement que ceux qu'il voit régulièrement et le connaissent également depuis plus longtemps. On en déduit que les gens avec qui nos enquêtés communiquent malgré la distance sont en moyenne affectivement plus proches d'eux. La fréquence des contacts semble d'ailleurs plus importante que l'âge de la relation dans la déclaration de proximité affective.

On remarque donc qu'il est assez périlleux d'entreprendre une réelle analyse des liens entre usages, interactions et affection à partir d'un échantillon si restreint du réseau de chacun de nos enquêtés. D'autant plus que les interactions publiques entre *ego* et ses alters semblent plus gouvernées par la propension qu'a chacun d'entre eux à commenter ou liker à tout va.

3.3.3 La position des conjoints dans le réseau dépend de la nature de leur relation

Le terme d'*alter-amour*, que j'emprunte à Claire Bidart, désigne l'alter représentant le conjoint au sein d'un réseau égocentré. C'est sans surprise l'un des rôles que les chercheurs ont le plus cherché à détecter. On désigne généralement comme *alter-ego* un alter qui est relié avec une grande partie , *a fortiori* une plus grande partie que n'importe quel autre alter, des autres sommets du réseau égocentré auquel il appartient. Ces sommets doivent probablement être à l'interface entre les communautés du réseau égocentré, les sphères sociales fréquentées par nos enquêtés. En fait on peut imaginer que l'alter-*ego* et *ego* ont *grosso modo* le même réseau personnel.

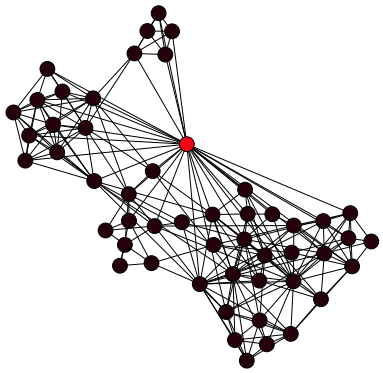


FIGURE 3.10 – Un réseau égocentré avec, en rouge, un alter-égo

Dans le cas où un réseau comprend un tel alter-*ego*, on suppose généralement que cet individu est en fait l'alter-amour. On peut en effet imaginer que lorsque deux individus partagent des relations qui sont issues de plusieurs sphères sociales différentes, la probabilité qu'une relation de couple existe entre eux est forte. Le travail de Lars

Backstrom et Jon Kleinberg [Backstrom and Kleinberg, 2014] s'est de ce point de vue distingué en proposant une mesure permettant de détecter les alter-amours qu'ils ont testée sur 1,3 million de réseaux égocentrés issus de Facebook et dont les propriétaires se sont déclarés comme marié, fiancé ou en couple avec l'un de leurs alters. Les chercheurs ont utilisé la mesure de *plongement* ou *embeddedness* d'un lien proposée par Marsden et Campbell [Marsden and Campbell, 1984] qui consiste à compter le nombre d'amis communs que partagent les deux individus, ce qui revient à calculer le degré des sommets, dans le cas d'un réseau égocentré. Ils ont trouvé que l'alter avec le plus haut score est dans 24,7% des cas l'alter-amour, ce qui selon eux montre le potentiel de la méthode structurale, sans pour autant être satisfaisant en l'état. Ils ont donc ensuite proposé un nouvel indicateur, la *dispersion*, qui leur permet de deviner qui est l'alter-amour dans près de 60% des cas lorsque celui-ci est l'époux ou l'épouse d'*ego*.

La dispersion d'un sommet dans un réseau égocentré capture le fait que les amis communs d'*ego* et de cet alter sont dispersés, et se connaissent eux-mêmes peu entre eux. La dispersion des voisins est conceptuellement plus forte que le simple fait d'en avoir un grand nombre et rejoint, en un sens, la proposition de Brooks et al. qu'on a vu en section 1.4.3 [Brooks et al., 2014]. La dispersion des voisins d'un alter v est formellement définie par la formule :

$$disp(v) = \sum_{s,t \in C_v} d_v(s,t)$$

où C_v est le sous-graphe induit par les voisins de v dans le réseau égocentré étudié, soit en quelque sorte le réseau égocentré induit de v. $d_v(s,t)$ est une distance valant 1 si s et t ne sont ni connectés ni n'ont de voisin commun dans C_v et 0 dans le cas contraire. La figure 3.11 décompose ces étapes.

J'ai comparé la mesure de Bascktrom et Kleinberg avec la centralité d'intermédiarité, plus usuelle, sur les réseaux de notre corpus. Puisque notre questionnaire ne permettait pas aux enquêtés de renseigner la présence éventuelle d'un alter-amour dans leur réseau, j'ai décidé d'utiliser le champ qu'il était possible aux enquêtés de remplir librement afin de les retrouver lorsqu'ils y étaient mentionnés. 2 799 alters ont ainsi reçu une qualification libre, parmi lesquels 522 alters-amour qualifiés d'amoure-ux-se, de compagn-on-e ou encore de LUI, par exemple.

À partir de ces qualifications d'alters-amour, j'ai créé une liste synthétique, fournie en figure 3.12, des différentes familles de relations à notre disposition afin de pouvoir travailler sur des données agrégées à partir desquelles il est possible de produire des analyses statistiques. Si l'on perd ainsi quelque peu en précision, il semble que différentes dénominations sont parfois proches les unes des autres, pouvant par exemple être liées à la casse ou à un espace supplémentaire placé en fin de ligne. On ne considère

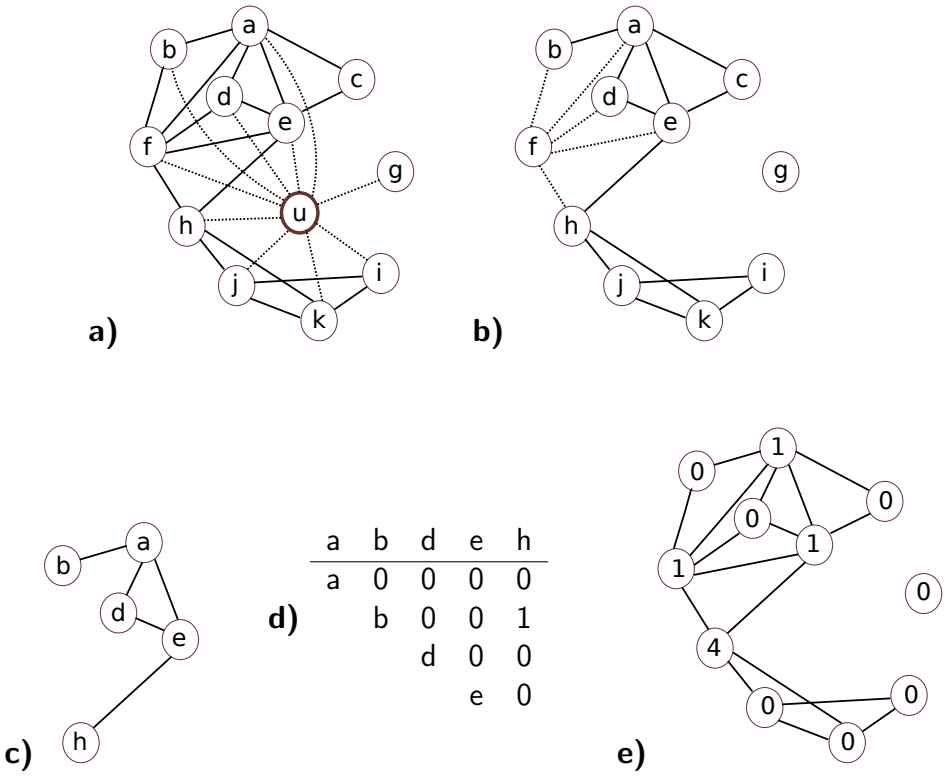


FIGURE 3.11 – **a)** Le réseau égocentré de u et u lui même, **b)** Le réseau égocentré de u, **c)** C_f, **d)** la table des $d_f(s,t)$ et **e)** les valeurs de dispersion de chaque sommet du réseau. On note que les sommets qui font le plus de ponts entre les autres alters ont une dispersion plus élevée. La métrique suggère que l'alter-amour de ce réseau est *h*.

par ailleurs que les types de relations qui ont au moins 10 représentants, finalement répertoriées par la figure 3.13.

J'ai calculé les centralités d'intermédiarité de ces alters-amour. Parmi ceux-ci, 48.7% sont effectivement l'alter le plus central du réseau égocentré. En analysant les résultats à l'aune des différentes qualifications obtenues, on remarque que les alters-amour qualifiés comme mariés, c'est-à-dire les « mari », « femme » et « épouse » sont ceux dont la part d'alters-amour qui sont les plus centraux est la plus importante à l'inverse des « copain », « petit ami » ou « petite amie » (cf. Figure 1).

La figure 3.14, où les types d'alters-amours sont triés selon la part ayant la plus importante centralité d'intermédiarité montre que les deux métriques sont similaires, bien que certaines nuances ressortent. La dispersion retrouve ainsi avec plus de précision les concubins et les maris mais moins efficacement les compagnes ou copines. On note d'ailleurs qu'elle semble plus efficace pour retrouver les alters-amours hommes tandis que l'intermédiarité détecte les alters-amours femmes plus souvent, à l'exception des conjoints et conjointes. La dispersion est une mesure de centralité locale tandis que l'intermédiarité dépend de l'ensemble du réseau, et donc également des sommets éloignés du réseau. On en déduit qu'un concubin aura plus tendance à être localement central que de représenter un pont entre les différentes communautés d'alters d'*ego*.

La figure 3.15 effectue un zoom sur la centralité d'intermédiarité des alters-amours. Ainsi, on retrouve à sa gauche les relations de couples *a priori* les moins fortes, copains, petits et petites ami-e-s, qu'on imagine moins structurantes par rapport au réseau que les maris, femmes et épouses qu'on retrouve à droite, parmi les valeurs les plus élevées. On note cependant de manière un peu surprenante qu'autour de 30% seulement des 14 fiancées sont dans la position la plus centrale du réseau de leur futur époux. Si les fiançailles représentent un engagement important dont on pourrait s'attendre à retrouver la trace dans le réseau égocentré, on peut également avancer l'explication que ces dernières s'effectuent généralement à un âge encore relativement peu avancé de la vie, et que d'autres acteurs (frère ou sœur, meilleur-e ami-e d'enfance, etc.) puisse encore occuper la position centrale du réseau au sens de nombreuses métriques.

Pour finir, on peut également remarquer que mis à part, encore, les conjoints et conjointes, les alters-amours femmes sont toujours plus centrales que leurs équivalents hommes (par exemple Mari et Femme). Dans tous les cas, il est difficile de juger si la position centrale de l'alter-amour vient de l'arrivée de membres du réseau de celle-ci parmi les alters de l'enquêté ou bien si elle résulte d'un investissement plus important avec le temps et de la force de la qualification envers les alters de l'enquêté (dans ce cas, on pourrait alors conclure que les femmes ont plus souvent tendance à s'approprier les amis de leur conjoint).

3.3.4 Cas des réseaux contenant un alter-ego

Cette section est un avant-goût de la seconde partie du manuscrit qui se concentre sur l'analyse des réseaux de Facebook. Elle postule qu'à l'image des usages différenciés de la plateforme qui induisent des rapports aux amis différents, la variabilité des structures relationnelles justifie que le rapport aux alters — et notamment à l'alter central — change également.

On a vu que l'alter-amour avait plus de chances d'avoir une centralité importante, et ce d'autant plus que la relation avec l'enquêté est déclarée comme étant une relation forte par ce dernier. On peut néanmoins considérer que la centralisation du réseau est également à prendre en compte puisque dans une structure sociale ne faisant pas émerger d'individu particulièrement central, ce dernier a certainement moins de chances d'être le fameux alter-amour.

De nombreux réseaux comportent ainsi un alter très central. Celui-ci est un candidat naturel au rôle d'alter-amour mais peut également très bien être un ou une ami(e) très proche, un frère, une soeur. Un approfondissement de l'étude du lien entre relation amoureuse et positionnement structural dans le réseau consiste alors à faire un nouveau zoom sur ces réseaux, qui comportent un alter central.

Quels réseaux correspondent à ce critère ? Pour les choisir, on s'intéresse au ratio de centralité d'intermédiarité entre les deux alters les plus centraux d'ego. Ce ratio semble être un indicateur pertinent, à condition de pouvoir déterminer un seuil approprié au-delà duquel on considérerait que le réseau comporte un tel alter très central. Cette sous-section présente des premiers travaux de recherche de ce seuil empirique.

Pour étudier la relation entre ce ratio et l'âge, la situation amoureuse et le genre des enquêtés, on représente les valeurs dans un tableau croisé, présenté en figure 3.16 indiquant, pour chaque combinaison de ces caractéristiques sociales, le nombre d'enquêtés concernés, et la médiane du ratio pour ces enquêtés. Dans ce tableau, couple s'entend hors mariage (comme le montrent les effectifs, qui peuvent être inférieurs aux effectifs des mariés). Notons que chaque condition supplémentaire (âge, genre ou situation amoureuse) fait diminuer l'effectif total, puisque tous les enquêtés n'ont pas renseigné toutes les rubriques.

Il faudrait bien entendu approfondir ce premier aperçu par une analyse statistique rigoureuse, mais le tableau obtenu suggère une influence notable de la situation amoureuse, comme on pouvait s'y attendre, avec une gradation, y compris dans une même tranche d'âge, entre célibataire, en couple et marié(e).

Le genre ne semble pas avoir d'impact, mais l'âge, lui, en a manifestement un, avec des valeurs plus fortes sur la tranche intermédiaire des 31–50 ans. Comme l'âge peut être

corrélé à la taille du réseau, il est intéressant de vérifier si les tendances subsistent en se restreignant à une certaine taille de réseaux. Le deuxième tableau, présenté en figure 3.17 montre les mesures pour les réseaux ayant entre 50 et 100 nœuds. Bien entendu les effectifs plus réduits poussent à la prudence, mais sur les catégories agrégées, les observations concernant la situation amoureuse d'une part (première ligne) et l'âge d'autre part (première colonne), paraissent assez conformes à celles du tableau précédent sans restrictions de taille de réseau.

Les valeurs elles-mêmes suggèrent un seuil heuristique à 2 très aisé à utiliser : on considérerait ainsi un réseau comme comportant un alter-ego si la centralité (d'intermédiarité) de l'alter le plus central est au moins deux fois plus élevée que celle de tous les autres alters. Cette valeur correspond peu ou prou à la valeur médiane sur l'ensemble du corpus, la médiane des enquêtés célibataires étant en dessous, celle des enquêtés en couple, dans ou hors mariage, se trouvant au-dessus.

Il serait alors intéressant de reproduire le travail de la section précédente en se restreignant au sous-corpus des comptes ayant un alter très central. Par ailleurs, une autre étude consisterait à reproduire les analyses de la section précédente, non plus simplement sur les alter-amour qualifiés mais sur l'ensemble des alters dont nos enquêtés ont précisé leur lien, en se concentrant sur les alter-ego (au sens, désormais, d'alter dont la centralité est au moins deux fois plus forte que celle des autres alters). Comme déjà évoqué, une analyse statistique plus précise des valeurs de centralité permettrait également de mieux comprendre le lien entre centralisation du réseau et situation amoureuse.

Au long de ce chapitre, on a donc exploré la manière dont se tissent les relations entre les usages de Facebook et les réseaux égocentrés en ligne de nos enquêtés, bien particuliers de par leur origine numérique, et comment eux aussi étaient dépendant de certaines de leurs caractéristiques socio-démographiques et socio-professionnelles. On a vu que les interactions entre les utilisateurs de Facebook et leurs alters permettaient de déceler, au travers des publications d'*ego*, deux principales lignées d'utilisation, la discussion et l'auto-publication.

On a également vu qu'il est malaisé de lier la position des sommets aux interactions sur la plateforme qui dépendent en effet beaucoup de l'usage propre qu'en fait chaque alter. Cependant on a quand même été en mesure de retrouver les conjoints de nos enquêtés, avec des divergences de résultats selon le type de relation, ce qui montre le

rôle grandissant du conjoint au sein du réseau, la vie allant.

Bien entendu, il est impossible de considérer que tous les réseaux égocentrés se ressemblent. Ils n'ont pas le même nombre d'alters et ceux-ci ne sont jamais reliés de la même manière. Il en résulte des réseaux aux formes différentes qu'on doit considérer différemment. Ce sont ces différentes formes qu'on va étudier dans la suite.

Amoureuse	Amoureuse, amoureuse, amoureuse , mon amoureuse
Amoureux	Amoureux, amoureux, mon amoureux, monamoureuxxxxx, moureux
Copine	C'est ma copine, Copine/Couple, Ma copine, Ma copine., Ma copine..., ma copine
Copain	Mon Copain, Mon copain, Mon copain., boyfriend, mon copain, mon copain , mon copain actuel
Femme	C'est ma femme, FEMME, Femme, Ma femme, Ma femme !, femme, ma femme, ma femme , ma femme :), marié
Mari	C'est mon mari, MON MARI, Mari, Mon Mari, Mon mari, mari, mon mari, mon mari
Meuf	C'est ma meuf !, Ma meuf, ma gonzesse, ma meuf, meuf
Mec	Mon mec, mon keum
Concubine	Concubine, Ma concubine, concubine
Concubin	CONCUBIN, Concubin, Mon concubin, concubin, concubin , mon concubin
Compagne	Compagne, Ma Compagne, Ma compagne, compagne, ma compagne, ma compagne !
Compagnon	Compagnon, Mon compagnon, c'est mon compagnon, Mon compagnon depuis 10 ans, avec qui je vis depuis 9 ans, compagnon, mon compagnon, mon compagnon de vie
Conjointe	Conjointe, Conjointe , Ma conjointe, conjointe, ma conjointe
Conjoint	Conjoint, conjoind, conjoint, mon conjoin, mon conjoint
Épouse	Epouse, Mon épouse, epouse, est mon épouse, mon épouse, épouse
Époux	époux
Fiancée	Fiancée, Ma fiancée, Ma future femme :), fiancee, fiancée, ma fiancee, ma fiancée
Fiancé	Fiancé, fiancé, mon fiancé
Petite amie	Ma petite amie, Ma petite amie., Ma petite-amie, PETITE AMIE, Petite Amie, Petite amie, Petite-Amie, ma petite amie, petit amie, petite amie, petite amie ! , petite-amie
Petit ami	Mon petit ami, Mon petit-ami, PETIT AMI, Petit Ami, Petit ami, Petit-ami, mon petit ami, mon petit ami , petit ami, petit ami, petit-ami
Petite copine	Petite copine, petite copine, petitz copine
Petit copain	Mon petit copain, Petit copain, mon petit copain, petit copain, petit copin
Mon amour	MON AMOUR , Mon amour, amour, love, mon amour
Couple	Couple, couple, en couple
Relation amoureuse	Relation amoureuse, relation amoureuse
L'homme de ma vie	l'homme de ma vie, l'homme de ma vie, l'amour de ma vie
Partenaire	ma partenaire (PACS), mon partenaire

FIGURE 3.12 – La liste des qualifications libres à droite, et en gras la classe que nous leur avons attribués

Copine	37	Copain	15
Femme	52	Mari	16
		Concubin	12
Compagne	54	Compagnon	29
Conjointe	11	Conjoint	59
Épouse	18		
Fiancée	14		
Petite amie	65	Petit ami	57

FIGURE 3.13 – Table du nombre de représentant selon le type de qualification d'alters-amours

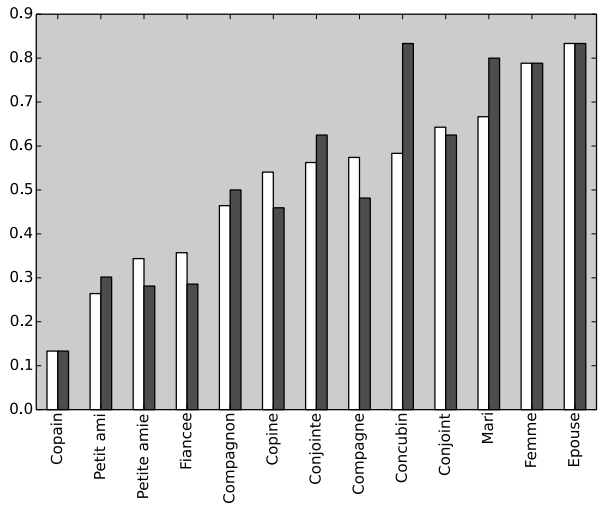


FIGURE 3.14 – Ces deux tableaux permettent de comparer la part d'alters-amours qui sont les plus centraux selon deux centralités : la centralité d'intermédiarité de (barre de gauche en blanc) et la dispersion de Kleinberg (barre de droite en gris).

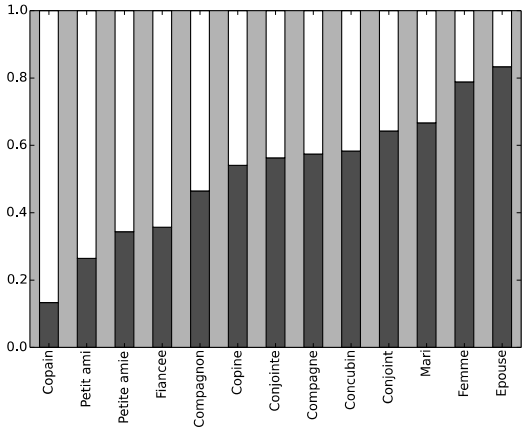


FIGURE 3.15 – La centralité d'intermédiarité des différents types d'alters-amour. En gris ceux qui ont la plus haute centralité du réseau et en blanc ceux qui n'ont pas la plus forte centralité du réseau.

		tous		célib		couple		marié	
		nb	med	nb	med	nb	med	nb	med
tous	**tous**			2756	1,8	2151	2,1	1332	2,4
	hommes	8846	1,9	2127	1,8	1448	2,1	957	2,4
	femmes	3593	1,9	610	1,8	679	2,1	354	2,3
> 50	**tous**	457	1,8	45	1,5	45	1,9	110	1,8
	homme	329	1,8	31	1,5	33	1,8	89	1,8
	femme	117	1,8	14	1,6	12	2,1	20	2,0
31–50	**tous**	2657	2,2	338	1,9	459	2,4	595	2,9
	homme	1964	2,2	280	1,8	347	2,3	450	2,9
	femme	618	2,1	54	2,0	103	2,5	130	2,8
20–30	**tous**	5442	1,9	1248	1,8	994	2,1	219	1,9
	homme	3738	1,9	989	1,8	630	2	140	2,0
	femme	1615	1,9	251	2	354	2,1	77	1,9
< 20	**tous**	1180	1,8	495	1,8	146	2,0	23	2,2
	homme	817	1,8	367	1,8	90	2,1	10	2,0
	femme	352	1,8	126	1,7	55	1,9	12	2,2

FIGURE 3.16 – Rapport de centralité médian entre les deux alters les plus centraux des réseaux selon différentes catégories (et effectif de chaque catégorie). Exemples de lecture : parmi les 140 hommes mariés entre 20 et 30 ans, la médiane du rapport de centralité vaut 2 ; parmi les 457 personnes de plus de 50 ans, la médiane est de $1,8$.

		tous		célib		couple		marié	
		nb	med	nb	med	nb	med	nb	med
tous	**tous**			861	1,8	695	2,1	365	2,5
	hommes	2564	2,0	643	1,8	454	2,1	247	2,8
	femmes	1185	2,0	212	1,8	235	2,1	112	2,1
> 50	**tous**	96	1,6	12	1,3	9	2,0	23	1,7
	homme	66	1,7	8	1,3	8	2,5	19	1,7
	femme	29	1,5	4	1,2	1	2,0	4	1,7
31–50	**tous**	729	2,2	94	1,8	138	2,2	168	3,3
	homme	525	2,2	71	1,6	103	1,7	128	3,4
	femme	178	2,4	20	2,5	33	2,8	34	3,1
20–30	**tous**	1706	1,9	404	1,8	302	2,2	61	2,2
	homme	1136	1,9	311	1,7	178	2,3	32	2,4
	femme	548	2,0	91	2	120	2,1	29	2,0
< 20	**tous**	375	1,7	151	1,6	53	1,9	5	1,5
	homme	258	1,7	110	1,6	34	2,1	3	2,1
	femme	113	1,7	40	1,6	19	1,7	2	1,3

FIGURE 3.17 – Rapport de centralité médian entre les deux alters les plus centraux pour les réseaux comptant entre 50 et 100 sommets, selon différentes catégories (et effectif de chaque catégorie).

Deuxième partie

Réseaux

Chapitre 4

Graphlets

Au cours du Festival **Futur en Seine 2014** l'équipe d'Algopol (dont Irène Bastard, Dominique Cardon, Stéphane Raux, Christophe Prieur et moi-même) avons réalisé une action de médiation pour tenter de faire connaître au plus grand nombre l'application d'enquête, dans l'idée de recruter de nouveaux participants, mais également afin de réaliser des entretiens en face à face et de confronter des visiteurs à leur réseau personnel. Pour donner un accès plus sympathique et plus ludique au concept de structure des réseaux aux intéressés, une clé de lecture avait été proposée par Rose Dumesny, qui est désormais doctorante en design à Orange Labs, partenaire du projet de recherche. La proposition consistait en une collection de petites illustrations de légumes dont les formes évoquaient des spécificités de la structure des réseaux.

Si les visiteurs de notre stand semblaient très satisfaits de repartir avec une petite carte de visite affublée du légume totem de leur réseau égocentré, nous nous sommes bien gardés de leur dire que dans certains cas, sa spatialisation n'était pas vraiment représentative de leur structure (on l'a vu au premier chapitre de cette thèse). Les réseaux en forme de patate (comme celui à droite de la figure 4.1) étaient par exemple trop gros pour être spatialisés en un temps acceptable par l'algorithme. Celui-ci finissait alors par les afficher de manière plus ou moins aléatoire, d'où justement ce résultat patatoïde, même dans les cas où la densité des liens aurait pu permettre (ce qui n'est pas toujours le cas) d'obtenir une visualisation propice à l'interprétation.

Dans la suite de notre enquête sur les réseaux sociaux, je vais présenter un découpage de notre panel d'enquêtés selon la forme de leur réseau. Il serait comme on l'a vu beaucoup trop long, fastidieux et imprécis de tenter de le faire empiriquement à partir du dessin via un algorithme de spatialisation en raison de la taille de certains réseaux, et surtout de leur nombre. On va donc, pour cela, faire appel à un indicateur, généralement plutôt utilisé en bio-informatique, qu'on appelle les *graphlets*. Confortés par la littérature

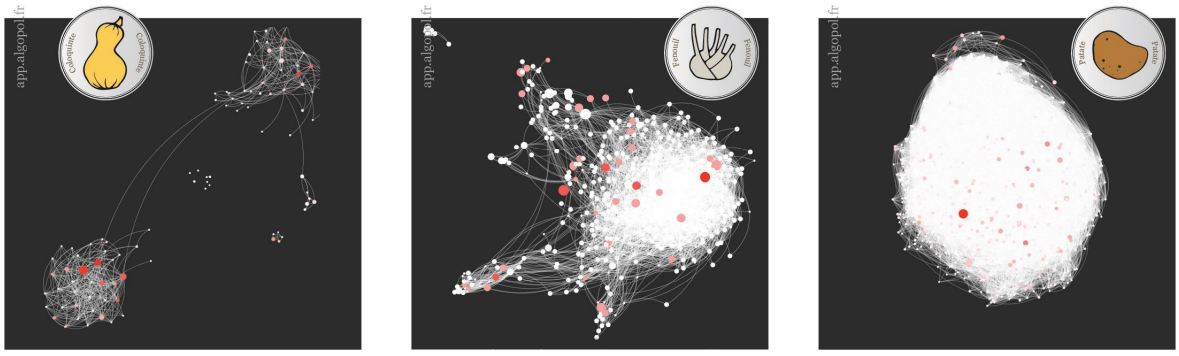


FIGURE 4.1 – Trois exemples de réseaux et leur légume associé

[Janssen et al., 2012], nous pensons en effet qu'il se révéler très efficace pour lire les différents modes de sociabilité et pour interpréter les rôles des alters dont nos réseaux sont composés.

Cette partie de la thèse correspond parfaitement à des recherches où se mêlent questions théoriques sur les réseaux et graphes et investigations concernant un outil d'analyse particulier, qui se nourrit d'une étude de cas, les réseaux sociaux d'internet, tout en apportant une contribution à la compréhension de ces objets eux-mêmes, comme on va le voir.

Je vais maintenant présenter cet outil, en commençant par faire remarquer qu'on peut également en trouver des proches cousins, sous le nom de *motifs* ou *patterns*, ou encore avec de subtiles variations dans leur définition, fluctuante selon le domaine et les chercheurs.

4.1 Définition

Les *graphlets* d'un graphe G sont l'ensemble de ses sous-graphes induits connexes. En d'autres termes, chaque ensemble de sommets du réseau et les arêtes qui les relient entre eux forment un graphlet, à condition que pour chaque couple de sommets, il existe au moins un chemin dans G les reliant et ne passant que par des sommets de ce sous-ensemble.

En pratique, et pour des raisons de combinatoire et de temps de calcul que je vais aborder dans la suite, on ne considère que les graphlets d'un nombre maximal de sommets, généralement 4, 5 et parfois 6 selon le nombre et la taille des réseaux étudiés.

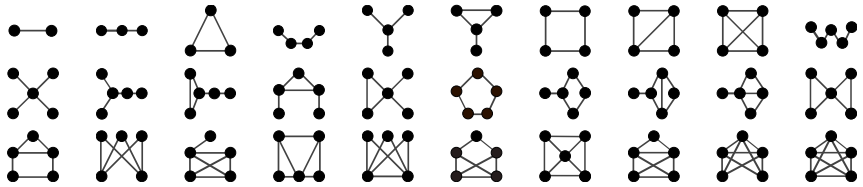


FIGURE 4.2 – Les 30 graphlets jusqu'à la taille 5.

Par la suite, on se limitera nous aux graphlets ayant un nombre de sommets allant jusqu'à 5. Ces graphlets sont présentés dans la figure 4.2.

Il existe donc un graphlet de taille 2 , deux graphlets de taille 3 , six graphlets de taille 4 et vingt-et-un graphlets de taille 5. Notre idée est de caractériser les réseaux personnels que nous avons à notre disposition par le nombre de chacun de leurs graphlets. Puisque ces petits réseaux ont des formes très variées , les différences en termes de nombre de graphlets entre deux réseaux doivent apporter des renseignements sur leur structure.

On parle parfois de *k-graphlets* pour dénommer ceux avec k sommets ou moins. J'emploierai, par la suite, ce terme pour désigner les graphlets ayant exactement k sommets de manière, ce qui est plus cohérent avec l'utilisation qu'on en fera. Je préfère en effet ne pas mélanger dans un même calcul ou dans la construction d'un même indicateur l'énumération de graphlets ayant un nombre de sommets différent. Ce choix découle du fait que les uns sont alors inclus dans les autres. Par exemple l'étoile à trois branches est trois fois induite dans l'étoile à quatre branches , comme l'illustre la figure 4.3, amenant des différences combinatoires dont il n'est pas certain qu'elles puissent être maîtrisées.

La figure 4.4 propose un court exemple de l'énumération des graphlets jusqu'à la taille 3 d'un petit réseau. Notons bien que toutes les combinaisons de sous-graphes connexes de trois sommets maximum sont visitées durant le parcours. En particulier, le chemin de taille 3 n'est visité que 2 fois.

La figure 4.5 compare l'énumération des graphlets de taille 4 entre le réseau du métro parisien et un des réseaux personnels de notre panel. Les deux réseaux ont le même nombre d'arêtes, mais on ne peut pas dire que leur similarité aille plus loin. Le réseau de transport a en effet plus de sommets et est beaucoup moins dense. Cette différence se traduit par un nombre de graphlets beaucoup plus petit. Comme les sommets y sont en moyenne plus éloignés les uns des autres, il y a en effet beaucoup moins de combinaisons de quatre sommets telles que le sous-graphe induit par ses sommets soit connexe, et donc moins de graphlets. Par ailleurs, comme ce réseau a également une transitivité très faible, il est surtout composé de chemins , qui sont seulement 10 fois moins nombreux que dans le réseau personnel. La différence du nombre d'étoiles

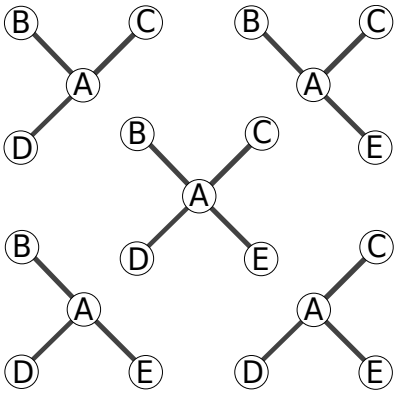


FIGURE 4.3 – Au centre : une étoile à quatre branches. Elle est entourée par chacune des étoiles à trois branches qui lui sont induites, à partir du sommet central (A) et des combinaisons possibles de trois sommets parmi les quatre autres (B, C, D, E)

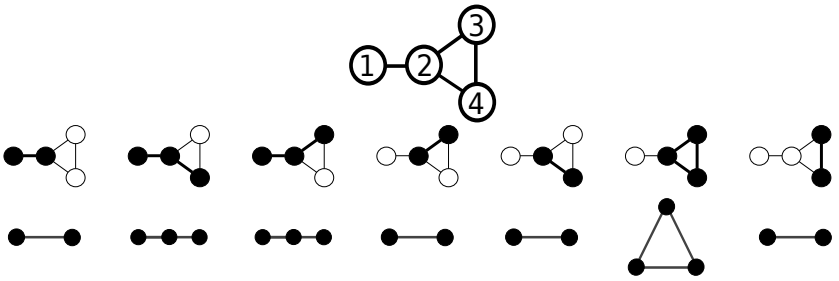


FIGURE 4.4 – Un exemple d'énumération des graphlets (jusqu'à la taille 3) du petit graphe, en haut. La ligne du milieu correspond aux sommets visités à chaque étape, le graphlet correspondant étant dessiné en dessous.

est également d'un facteur proche de 10 mais on voit que les graphlets qui induisent des triangles sont absents, ou relativement absents par rapport au réseau social.

Grâce à l'analyse des graphlets de nos réseaux sociaux, nous souhaitons faire émerger les différents modes de socialisation avec les formes des réseaux. On aimerait pouvoir détecter efficacement les réseaux centrés autour d'un conjoint ou bien autour d'une communauté, par exemple familiale ou de promotion d'école. On va par la même occasion, proposer et explorer une méthode d'interprétation de l'énumération obtenue qui met en avant les différences structurales entre ces réseaux, qui sont somme toute, assez semblables entre eux. On imagine que celle-ci peut être adaptée à d'autres contextes de recherche qui font usage de l'outil graphlets, donnant ainsi des perspectives plus

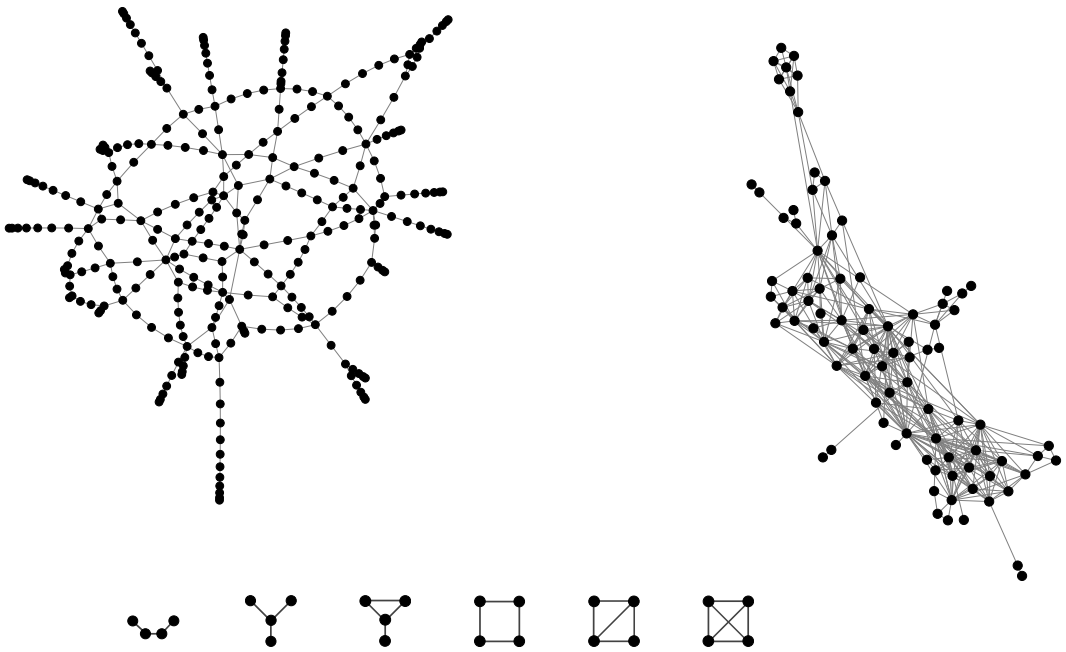


FIGURE 4.5 – Deux réseaux de même nombre d'arêtes, à gauche le réseau du métro parisien et à droite un de nos réseaux égocentrés.

larges que l'analyse des réseaux sociaux à ce chapitre et aux suivants.

4.2 Les positions au sein des graphlets

On va ici présenter un élément subséquent à la définition des graphlets, qu'on appelle les *positions* mais qu'on trouve également dans la littérature sous le nom d'*orbites*. Ce sont les places qu'occupent les alters dans les graphlets dans lesquels ils sont inclus. Lors de la présentation des concepts de la sociologie des réseaux (cf. 1.4.2), j'avais évoqué les travaux d'Harrison White concernant l'équivalence structurale. Cette dernière est dérivée d'un concept d'algèbre nommé *automorphisme* directement lié aux positions.

4.2.1 Notions d'algèbre

Une branche de l'algèbre est dédiée à l'étude des morphismes, une abstraction des fonctions qui s'appliquent aux objets mathématiques, comme les graphes, pour calculer par exemple leur diamètre. Un tel morphisme, représentant la fonction, qu'on appelle ici $diam$, de calcul du diamètre, serait alors décrit ainsi :

$$\begin{array}{rcl} diam : \mathcal{G} & \rightarrow & \mathbb{N} \\ G & \mapsto & \text{diamètre de } G \end{array}$$

Soit : La fonction $diam$ associe à un objet G , appartenant à l'ensemble des graphes $\mathcal{G}$, son diamètre qui appartient lui à l'ensemble des nombres entiers $\mathbb{N}$. L'ensemble des graphes, qu'on a ici pris comme exemple est l'ensemble de départ, tandis que l'ensemble des entiers est l'ensemble d'arrivée du morphisme.

Il existe plusieurs sortes de morphismes. Ainsi les *endomorphismes* sont caractérisés par le fait que leur ensemble d'arrivée est le même que l'ensemble de départ. C'est-à-dire que si le morphisme part de l'ensemble des graphes, son résultat doit lui aussi être un graphe, qui appartient donc bien à l'ensemble des graphes. Par exemple la fonction *anti_ego*, qui à un graphe rajoute un sommet et le lie à tous les nœuds précédemment présents, définit un endomorphisme de graphes :

$$\begin{array}{rcl} \textit{anti_ego} : \mathcal{G} & \rightarrow & \mathcal{G} \\ G = (V, E) & \mapsto & (V \cup \{v\}, E \cup \bigcup\limits_{u \in V} (v, u)) \end{array}$$

Le morphisme le plus important est probablement l'*isomorphisme* qui associe à son élément d'entrée une sortie ayant la même structure. Dans le cas où ces deux éléments sont le même objet mathématique, on parle alors d'*automorphisme*.

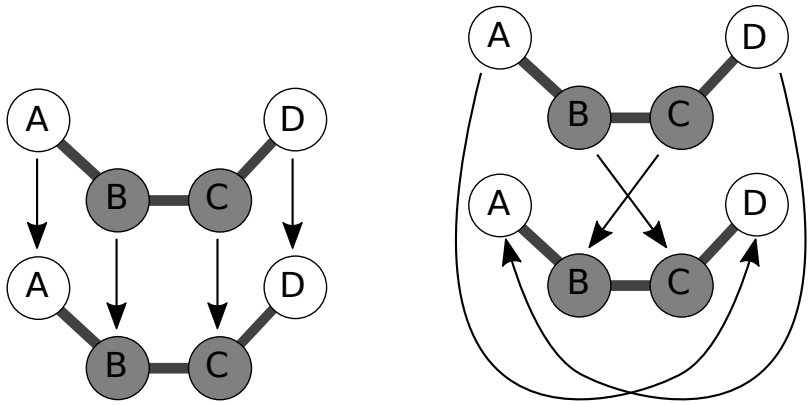


FIGURE 4.6 – Les 2 seuls automorphismes du chemin de taille 4 . Ce graphe possède donc deux classes d'équivalence qui sont (A, D) et (B, C), ici matérialisées par les couleurs des sommets.

Un automorphisme ϕ d'un graphe $G = (V, E)$ est une permutation de l'ensemble de ses sommets telle que pour tous sommets $u, v \in V$, une arête relie u à v si et seulement si une arête relie $\phi(u)$ à $\phi(v)$. On note généralement l'ensemble des automorphismes d'un graphe G vers lui même $Aut(G)$.

On dit que deux sommets u et v d'un graphe G sont équivalents si et seulement si il existe un automorphisme de G vers lui-même qui envoie u sur v. Pour chaque sommet v, sa classe d'équivalence, que je note $C_e(v)$ est l'ensemble des sommets qui lui sont équivalents. Dans le cas du chemin de taille 4, illustré par la figure 4.6, les classes d'équivalence sont les suivantes : $C_e(A) = \{A, D\}$, $C_e(B) = \{B, C\}$, $C_e(C) = \{B, C\}$ et $C_e(D) = \{A, D\}$. Le chemin de taille 4 possède donc exactement deux classes d'équivalence.

Dans le cas de l'équivalence structurale de White vue en section 1.4.2, deux sommets étaient équivalents s'ils avaient les mêmes voisins. Ici, l'équivalence est moins contraignante puisqu'elle concerne les sommets qui ont pour voisins des sommets équivalents entre eux.

4.2.2 Positions et intérêts de la notion

On appelle *positions*, ou *orbites*, les classes d'équivalence des graphlets. Elles ont été introduites pour la première fois par Natasa Pržulj ([Pržulj, 2007]). Dans le chemin de la figure 4.6, il y a donc deux positions, tout comme dans l'étoile , par exemple qui contient une position centrale, occupée par un sommet et une position périphérique, qui est partagée par quatre sommets. L'ensemble des positions des graphlets jusqu'à

la taille 5 est présenté en figure 4.7

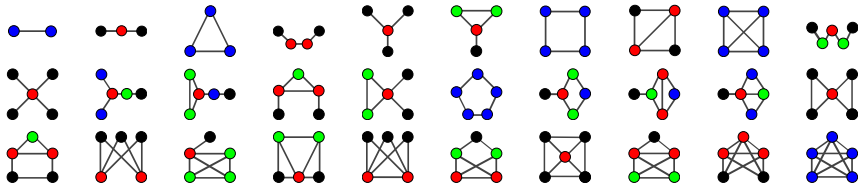


FIGURE 4.7 – Les positions au sein des 30 graphlets jusqu'à la taille 5.

On va voir plus loin de quelle manière ils opèrent, mais certains algorithmes font plus que compter les graphlets : ils comptent également, pour chaque sommet du réseau, les positions auxquelles ceux-ci appartiennent dans les graphlets énumérés. Ces indications pourraient être riches concernant l'interprétation du rôle des alters au sein des réseaux personnels. Par exemple : on peut imaginer qu'un individu très souvent présent dans la position centrale de l'étoile a un rôle important auprès d'*ego*. En effet, cette personne connaîtrait alors beaucoup des amis de l'enquêté sans que ceux-ci se connaissent entre eux (en extrapolant le fait qu'ils ne soient pas amis sur Facebook, bien sûr). De tels alters sont des gens rares et proches de l'enquêté, qui leur a présenté des amis de ses différentes sphères sociales. Au delà de cet exemple, la diversité des positions de chaque sommet doit permettre de trouver des profils pouvant donner des qualifications de proximité avec l'enquêté contenant encore plus d'informations qu'un score dans une métrique de centralité par exemple.

Il y a une seule position dans l'unique graphlet à deux sommets , 3 positions dans les deux graphlets de taille 3 , 11 pour les graphlets de taille 4 et 58 pour ceux de taille 5. Il y a donc en tout 73 positions différentes dans les graphlets à 5 sommets ou moins.

4.3 Limitations des graphlets

Arrêtons-nous un instant sur ce que les graphlets disent et ce que ne peuvent pas dire, ou du moins pas directement. En fait, ils capturent localement la structure d'un réseau. Par accumulation et agrégation de ces clichés localisés, on souhaite aboutir à une représentation globale du graphe étudié mais il est possible que cette représentation soit imparfaite.

La taille maximale des graphlets peut poser un problème que le réseau du métro parisien illustre bien. Celui-ci est en effet composé de beaucoup de cycles de grande taille, par exemple celui qui passe par Montparnasse et Place d'Italie, au sud de la figure, si

j'ose dire, est composé de 19 arêtes. Une énumération des motifs de taille 5 sur ce sous-réseau ferait ressortir 19 chemins de taille 5 . Mais alors, comment faire la différence avec un grand arbre, un réseau sans cycle lui-même composé de beaucoup de chemins ? Encore une fois il est certain qu'une bonne connaissance générale des réseaux étudiés et donc des formes pouvant être rencontrées aide à la lecture des résultats de l'énumération des graphlets.

Sans être une limitation à proprement parler, puisque n'étant pas l'objectif de la mesure, on peut noter quelques liens entre graphlets et connexité du réseau. La définition que j'utilise des graphlets, en ne prenant conjointement que ceux d'une taille donnée, contrairement à plusieurs métriques de la littérature que je vais présenter en section 4.4, empêche notamment de détecter les sommets isolés. Et même, pour une taille k donnée, elle ne permet pas de recenser les composantes connexes de taille $k-1$ puisqu'aucun graphlet de taille k n'y serait inscrit. Il est cependant nettement plus rapide d'énumérer les composantes connexes et leurs tailles que les graphlets, ce qui peut donc être fait par ailleurs.

Considérant tout ceci, on conclut déjà que les formes de sociabilité qu'on va trouver reposent surtout sur l'analyse de la composante connexe principale, et que les sphères sociales isolées seront peu lisibles. Dans le cas de l'analyse de réseaux sociaux, il est difficile de considérer ces problèmes comme réellement handicapants puisque ceux-ci ont peut de cycles non fermés, et les sphères sociales isolées sont généralement très petites et peu nombreuses.

4.4 Historique de la notion

On trouve dans la littérature beaucoup d'occurrences d'emplois des graphlets ou d'outils proches. Je vais faire une courte présentation de l'évolution de ces méthodes, qui traversent plusieurs champs disciplinaires.

On peut remonter aux années 60 et aux travaux de l'américain James Davis, ancien doctorant à Harvard et alors chercheur pour l'université de Chicago, pour trouver les racines de l'étude des motifs. Il réinterprète alors 56 questions sociologiques en problèmes liés aux réseaux et montre l'importance de l'étude des structures locales de ceux-ci pour y répondre [Davis, 1963]. Il a ensuite utilisé, avec Samuel Leinhardt, les triades pour vérifier la thèse de George Homans [Homans, 1950] qui postule qu'un groupe de personnes génère naturellement une structure regroupant des cliques ainsi qu'un système hiérarchique [Davis and Leinhardt, 1967].

Les triades, présentées en figure 4.8, sont des objets qui ont été proposés par Georg Simmel et qui consistent en des réseaux, généralement orientés, de trois sommets qui

représentent l'ensemble des façons, pour trois individus, d'être connectés ensemble. Elles ont été abondamment utilisées, et on a notamment déjà évoqué le cas de la triade interdite, où deux personnes étant chacune en relation forte avec une même troisième ne se connaissent pas entre elles [Granovetter, 1977]. Le travail de Davis et Leinhardt correspond à la première occurrence de leur dénombrement et ces chercheurs les utilisent ensuite pour comparer leur distribution à celle de modèles aléatoires. Ils ne précisent cependant pas quelle méthode ils utilisent pour dénombrer ces triades mais on peut supposer qu'ils travaillaient à base de parcours matriciel puisque les réseaux étaient représentés sous la forme de sociogrammes, tels que les a développé par Moréno [Moreno et al., 1934].

Paul Holland et Leinhardt ont continué à explorer les possibilités du modèle, tout en déplorant déjà sa limite combinatoire. Ils notent en effet qu'il y a 218 tétrades et 9 608 pentades (les équivalents respectifs des triades pour 4 et 5 sommets), ce qui semble rédhibitoire en termes de calculs. Au long de plusieurs articles publiés successivement [Holland and Leinhardt, 1971, Holland and Leinhardt, 1974, Holland and Leinhardt, 1977], ils ont développé des méthodes pour calculer et analyser la distribution du nombre d'apparitions des triades dans un réseau social, en se basant toujours sur des graphes aléatoires, .

C'est finalement une publication dans Nature dont le premier signataire est le biologiste Ron Milo [Milo et al., 2002] qui a donné sa visibilité à la méthode, et ce au delà des frontières de la sociologie et de la psychologie. La publication fait suite à une étude préalable au cours de laquelle Milo et son équipe ont énuméré les triades induites d'un réseau construit à partir de l'activité des *facteurs de transcription*. Ces facteurs de transcription sont en fait des protéines impliquées dans la lecture et l'interprétation des gênes. Ils décrivent les triades qui apparaissent de manière significativement plus importante dans ce réseau que dans des graphes aléatoires dont les propriétés globales (distribution de degré, densité, nombre de sommets, etc.) sont similaires. Ils nomment alors ces sous-graphes proportionnellement sur-représentés les *motifs* et proposent par la suite des méthodes pour les calculer [Shen-Orr et al., 2002]. Dans cette première publication, donc, ils généralisent l'approche à différents types de réseaux et expriment son intérêt pour l'ensemble des domaines employant la modélisation par les réseaux.

Plus tard, la même équipe a mis au jour des « super-familles » de réseaux présentant des caractéristiques communes. Pour ce faire ils ont comparé la distribution des triades dans ces réseaux, non plus seulement les plus représentées, mais cette fois toutes [Milo et al., 2004]. Deux réseaux sociaux ont ainsi été regroupés avec des réseaux issus de liens hypertextes dans une super-famille où la triade 3 0 0 (dans laquelle tout le monde se connaît) est sur-représentée et où la triade interdite, qui n'a aucune raison *a priori* de l'être dans un réseau qui ne soit pas un réseau social, est sous-représentée (voir la figure 4.9). On imagine donc que les réseaux de liens hypertextes se comportent de

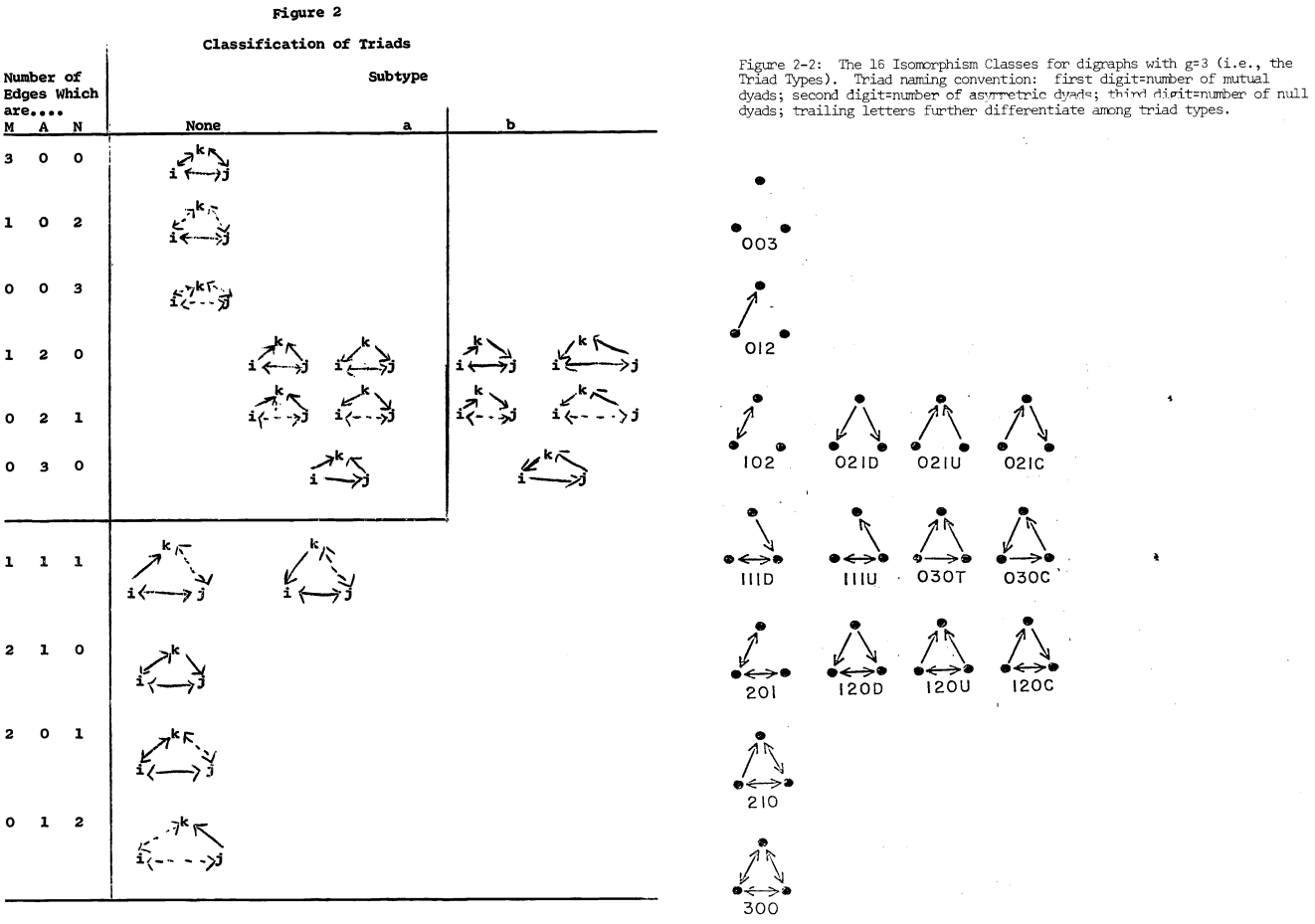


FIGURE 4.8 – Les triades décrites par Davis et Leinhard en 1967 à gauche et par Holland et Leinhardt en 1974 à droite. Les triades sont nommées par leur nombre de liens M pour « mutuel » $\leftrightarrow$, A pour « asymétrique » $\leftarrow$ ou $\rightarrow$ et N pour « mutuellement nul » dessiné en pointillés puis plus du tout. Les lettres indiquent l'orientation des liens asymétriques. Davis et Leinhard montrent que les cinq triades du bas sont peu présentes dans leurs réseaux et on retrouve d'ailleurs parmi elles la triade interdite de Granovetter, notée ici 2 0 1. La notation de Holland et Leinhardt deviendra standard pour les triades.

manières sensiblement similaires aux réseaux d'interactions entre individus, avec une transitivité forte. Ces travaux vont précipiter les chercheurs à massivement utiliser les graphlets, triades, ou autres méthodes associées, dans des prolongements qu'on verra plus loin et qui ont inspiré les travaux que j'ai réalisé.

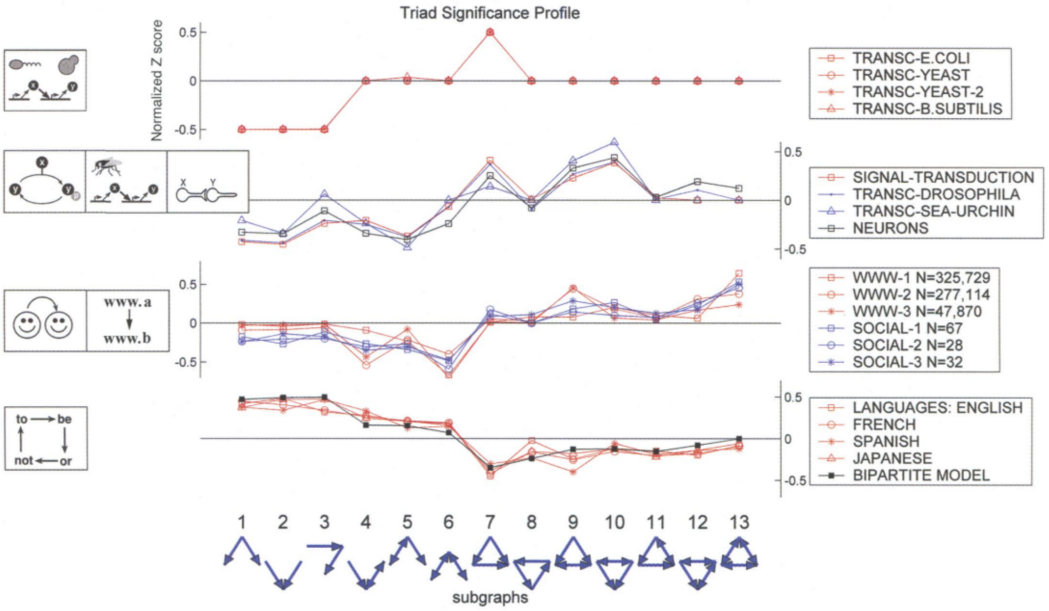


FIGURE 4.9 – Les super familles de réseaux issues de [Milo et al., 2004]

4.5 Considérations algorithmiques du calcul des graphlets

La méthode de calcul employée pour énumérer les graphlets est extrêmement importante puisque comme l'ont souligné Holland et Leinhardt, la combinatoire des graphlets est particulièrement exigeante. C'est d'autant plus vrai que nous avons un panel de 16 000 graphes dont certains ont jusqu'à 5 000 sommets, ce qui représente un nombre d'opérations considérable. Pour illustrer cette difficulté, la figure 4.10 présente d'ailleurs le temps d'énumération, en secondes, par rapport au nombre de sommets et de liens des réseaux sur lesquelles nous avons pu effectuer le calcul. Notons que l'implémentation des algorithmes a été réalisée par mes soins en python et sans parallélisme, et qu'il est certain que de meilleures performances peuvent être atteintes. On remarque cependant que le temps de calcul semble plus nettement dépendant du nombre de liens du réseau plutôt que de celui des sommets. La présence de réseaux contenant de nombreuses arêtes parmi ceux ayant nécessité le moins de temps de calcul montre la difficulté à l'analyser.

L'explosion du temps de calcul est en fait combinatoire. On a en effet vu dans la figure 4.4 que toutes les combinaisons de sommets connexes (jusqu'à une certaine taille) étaient visitées. Dans le petit graphe de l'exemple, ce nombre de combinaisons de-

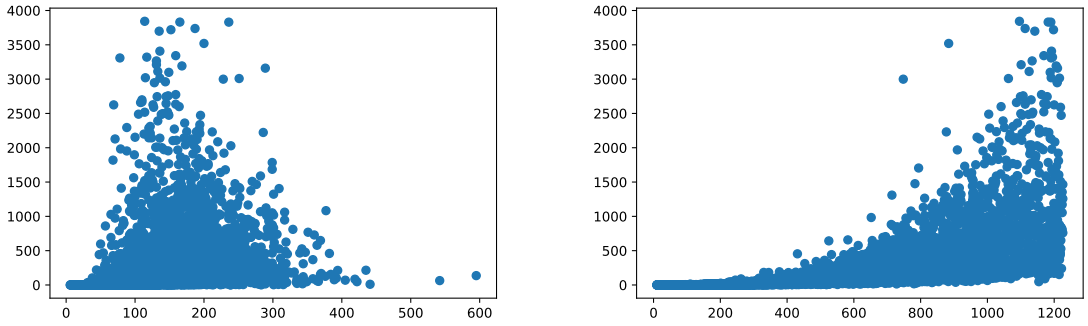


FIGURE 4.10 – Le temps du calcul de l'énumération totale des graphlets, en secondes. À gauche en fonction du nombre de sommets, et à droite du nombre d'arêtes.

meure peu élevé, mais c'est loin d'être le cas lorsque la quantité de sommets du réseau augmente. On peut ainsi utiliser nos réseaux égocentrés pour observer la croissance du nombre effectif de graphlets selon la taille des réseaux considérés. C'est ce que représente la figure 4.11, pour chaque taille de graphlets entre 3 et 7, contenus dans les graphes de notre panel, en fonction du nombre de liens entre les alters. Il est à noter que l'échelle en ordonnée est trompeuse, et que le nombre de graphlets augmente exponentiellement avec leur taille puisque $e^{15} \simeq 3.000.000$ tandis que $e^{20} \simeq 500.000.000$, et que les variations sont donc très importantes.

Le nombre de sous-graphes induits à visiter est déterminant, puisque l'algorithme doit identifier chacun d'entre eux à un graphlet lors du parcours afin d'incrémenter la valeur représentant le nombre de fois où ce graphlet a été rencontré. L'identification du sous-graphe induit consiste à trouver, parmi la liste des graphlets étudiés, celui qui lui est identique, ou isomorphe (on a vu cette notion en section 4.2.1). En général, l'isomorphisme est une question longue à traiter, mais comme la liste des graphlets à traiter est courte et connue d'avance, des optimisations existent. On peut par ailleurs noter que les différents graphlets ne sont pas tous liés de la même manière avec le temps de calcul, comme le montre la figure 4.12, ce qui doit certainement participer à la difficulté de proposer une complexité algorithmique.

Je vais maintenant présenter quelques algorithmes d'énumération des graphlets dans un graphe. À quelques exceptions près, ceux-ci sont donc constitués d'un parcours de tous les sous-graphes induits au réseau, et de la détermination du graphlet correspondant à chaque sous-graphe courant. Il en existe différentes modalités, selon les types de graphes et de graphlets à traiter (par exemple, des versions existent pour les graphes pondérés, orientés, etc.). On peut également rapidement mentionner l'existence d'al-

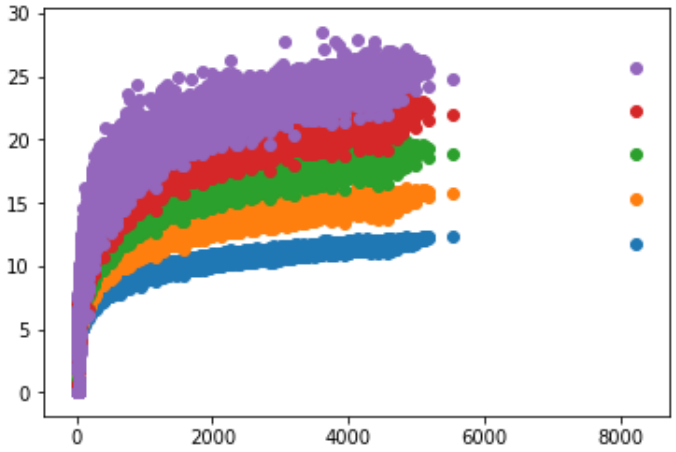


FIGURE 4.11 – En abscisse le nombre d'arêtes des graphes et en ordonnées le logarithme du nombre de graphlets induits. Chaque couleur, ou trace distinctive représente un une taille maximale de graphlets.

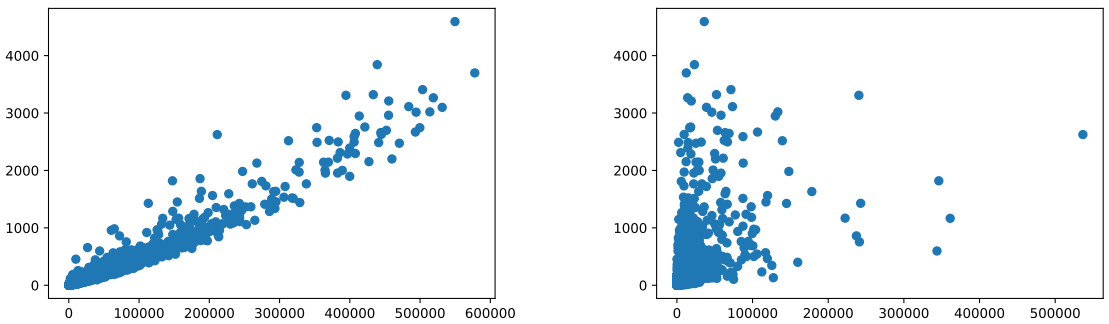


FIGURE 4.12 – Le temps du calcul de l'énumération total des graphlets, en secondes, en fonction du nombre d'apparition d'un graphlet particulier. À gauche le cerf-volant et à droite la clique

gorithmes retournant des valeurs approchées du nombre de graphlets dans le réseau, dans le but de faire face au temps de calcul exponentiel en sacrifiant une partie de la précision des résultats (on peut par exemple citer [Pržulj et al., 2006]).

4.5.1 Parcours du réseau

Dans leur publication de 2002 basée sur le réseau de régulation de la transcription des gènes [Shen-Orr et al., 2002], Shai Shen-Orr et ses collègues décrivent de manière succincte leur méthode d'énumération des graphlets basée sur l'analyse matricielle. Je rappelle que dans le cas usuel d'une représentation d'un réseau sous la forme d'une matrice, celle-ci est carrée et que chaque sommet est représenté par une ligne et une colonne. Chaque case $M_{i,j}$ vaut alors 1 si les sommets de la ligne i et de la colonne j au croisement desquelles elle se situe sont reliés entre eux et 0 sinon. Pour visiter chaque sous-graphe connexe ils recherchent récursivement les cases valant 1 puis parcourent la ligne et la colonne de cette case en cherchant d'autres 1 qui permettent donc de retrouver les sommets adjacents au sous-graphe courant.

En 2004, la même équipe - ou du moins une équipe composée autour du même noyau de chercheurs - propose un algorithme d'échantillonnage permettant l'évaluation du nombre de chaque sous-graphe d'une taille donnée dans un réseau [Kashtan et al., 2004]. L'objectif de la méthode d'échantillonage est de contrebalancer l'augmentation importante du temps de calcul avec la taille du réseau étudié. Seul un certain nombre de sous-graphes vont donc être visités au fil d'un parcours aléatoire des arêtes. Pour visiter un nouveau sous-graphe, l'algorithme part d'une arête tirée au hasard et qui lui sert de base. Il lui ajoute ensuite progressivement des liens et des sommets, jusqu'à ce que le nombre de sommets souhaité soit atteint. Chaque nouveau lien est tiré au sort parmi ceux adjacents au sous-réseau courant et qui ne relient pas deux sommets en faisant déjà partie. Ce nouveau lien ajoute ainsi un nouveau sommet au sous-réseau.

Formellement, et en partant toujours d'un graphe $G = (V, E)$, l'algorithme qui construit un sous-graphe aléatoire de taille k est présenté dans la figure 4.13. Notons bien qu'au moment de retourner le sous-réseau de taille k, l'algorithme reconstruit la liste des arêtes du sous-graphe à partir de G. En effet, certaines arêtes de ce sous-graphe ont pu ne pas être visitées dans les étapes précédentes mais doivent néanmoins y être incluses.

Par ailleurs le but du processus est de calculer ce que les auteurs appellent la *concentration* d'un sous-graphe, c'est-à-dire le ratio d'apparitions de ce sous-graphe par rapport à l'ensemble des sous-graphes du réseau étudié $C_i = \frac{N_i}{\sum_i N_i}$ où i est l'index d'un des différents sous-graphes possibles de taille k. C'est ensuite la comparaison de la concentration d'un sous-graphe du réseau étudié avec sa concentration moyenne dans des

réseaux aléatoires qui va permettre de déterminer quels sont les motifs du réseau, ses sous-graphes sur-représentés. Dans cette optique, et sans rentrer dans les détails, je voudrais mentionner le fait que les auteurs tiennent bien compte du biais de parcours de leur algorithme qui aura tendance à favoriser le tirage de certains sous-graphes par rapport à d'autres.

Malgré cela, certains reproches lui sont faits. Sebastian Wernike note ainsi que tous les biais ne sont pas résolus comme l'illustre la figure 4.14. Il propose donc un autre algorithme de parcours des sous-graphes [Wernicke, 2006]. Sa méthode d'échantillonage des graphlets, que je présente juste après, se base sur un algorithme de parcours exhaustif des sous-graphes qu'il nomme *ESU* pour *Enumerate SUbgraphs*.

Pour décrire l'algorithme de Wernicke, on va avoir besoin de la définition du *voisinage exclusif* d'un sommet. On a déjà vu que le voisinage d'un sommet était l'ensemble de ses voisins. Le voisinage exclusif d'un sommet w d'un graphe $G' = (V', E')$ est défini par $N_{exclu}(w, V') = \{u \in N(v) | u \notin V' \cup N(V')\}$.

Dans l'algorithme de Wernicke, décrit par la figure 4.16, chaque sommet est identifié par un numéro unique et sert de racine à l'énumération de l'ensemble des sous-graphes connexes le contenant. Cette énumération est effectuée par la fonction ExtendSubgraph qui construit récursivement tous les sous-graphes possibles n'ayant que des sommets d'indice supérieur à celui de la racine. Wernicke propose dans son article une illustration des appels successifs à la fonction ExtendSubgraph pour l'énumération des sous-graphes de taille 3 d'un petit graphe 4.15. En pratique Wernicke n'utilise pas son algorithme pour énumérer tous les sous-graphes mais ajoute un degré d'aléatoire au moment de chaque appel à la fonction ExtendSubgraph. Afin d'obtenir un échantillonage du nombre de graphlets, il arrête le parcours à chaque étape avec une certaine probabilité dépendant de la taille du sous-graphe courant.

Avec la méthode qu'il emploie, quel que soit le groupe de k sommets formant un sous-graphe connexe, Wernicke s'assure, et le prouve formellement dans sa publication, qu'il les visite bien tous une et unique fois durant le processus. Par ailleurs, il précise dans sa revue de littérature que trouver les motifs d'un réseau consiste en trois sous-tâches qui sont : 1- trouver les sous-graphes dans le réseau étudié, 2- les regrouper en classes de sous-graphes identiques (isomorphes, donc, comme on l'a vu) et 3- déterminer lesquels apparaissent proportionnellement plus souvent que dans le cas de graphes aléatoires. La publication propose une méthode efficace pour le premier point et développe également une nouvelle approche concernant l'échantillonage des graphlets. Voyons donc maintenant comment on peut s'y prendre pour accélérer le processus d'identification des sous-graphes en graphlets.

Algorithme 1 : Kashtan_Random

Données : $G = (V, E)$, k
Résultat : ClustersParK

`/* Initialisation */`

ClustersParK $= \{\}$
$E_S = \{\}$ est l'ensemble, au départ vide, des arêtes du sous-réseau
$V_S = \{\}$ est l'ensemble, au départ vide, des sommets du sous-réseau
L est la liste, au départ vide, des arêtes adjacentes au sous-réseau

`/* Parcours */`

(1)
Tirer au hasard une arête $e_1 = (v_1, v'_1) \in E$
$E_S = \{e_1\}$,$V_S = \{v_1, v'_1\}$

(2)
pour *chaque arête* $e_2 = (v_2, v'_2) \in E$ **faire**
  si *(*$v_2 \in E_S$ *et* $v'_2 \notin E_S$*) ou (*$v'_2 \in E_S$ *et* $v_2 \notin E_S$*)* **alors**
    ajouter e_2 à L

(3)
si L *est vide* **alors**
  si V_S *contient* n *sommets* **alors**
    $E_S = \{\}$
    pour *chaque arête* $e = (u, v) \in E$ **faire**
      si $u \in V_S$ *et* $v \in V_S$ **alors**
        $E_S = E_S \cup \{e\}$
      retourner $G_S = (V_S, E_S)$
  sinon
    retourner « *Pas de sous-graphe possible de taille* k »

(4)
Retirer au hasard une arête $e_3 = (v_3, v'_3)$ de L
$E_S = E_S \cup \{e_3\}$
$V_S = V_S \cup \{v_3, v'_3\}$

(5)
Recommencer à l'étape (2) si $\|V_S\| < k$

FIGURE 4.13 – Algorithme de parcours d'un sous-graphe aléatoire de taille k décrit dans [Kashtan et al., 2004]

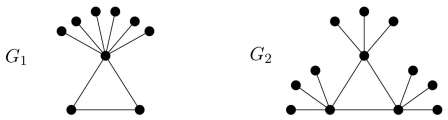


FIGURE 4.14 – Dans ces graphes, dont j'emprunte le dessin à la publication de Sebastian Wernicke, le triangle est considéré comme sur-représenté par l'algorithme de Kashtan et al. alors qu'il n'apparaît qu'une seule fois.

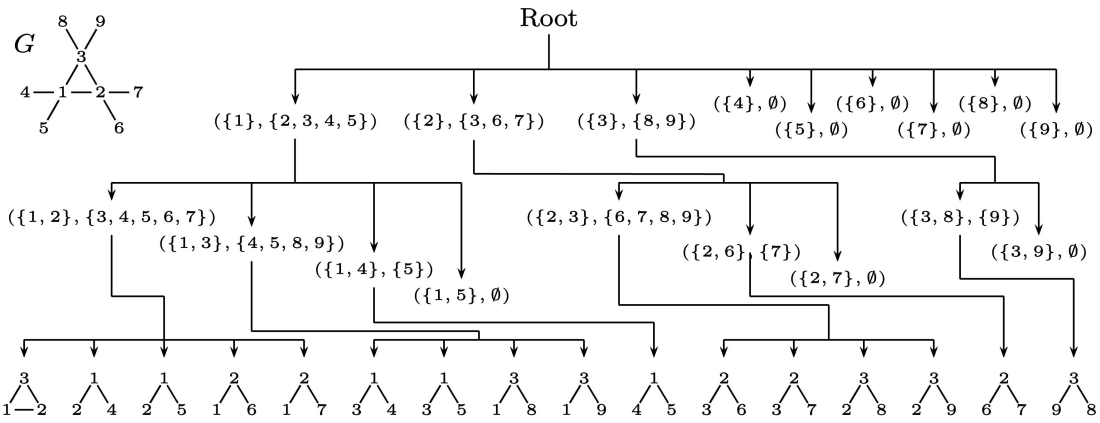


FIGURE 4.15 – Les appels successifs à ExtendSubgraph pour l'énumération au sein du graphe à gauche, image tirée de [Wernicke, 2006].

4.5.2 Détermination du graphlet

En 2009, Stoica et Prieur adaptent l'algorithme de Wernicke [Stoica and Prieur, 2009] en lui ajoutant une routine d'identification des graphlets tout en supprimant le caractère aléatoire du processus. Cette routine leur permet d'énumérer les positions de chaque sommet dans les sous-graphes. Ils proposent ce faisant de l'utiliser afin de caractériser le rôle des alters au sein du réseau égocentré. Comme on l'a mentionné, l'isomorphisme de graphe est un processus long mais l'univers des graphes connexes de taille 5 est suffisamment restreint pour élaborer quelques raccourcis.

Pour ce faire, ils remarquent donc que deux graphlets de taille 4 ou moins sont isomorphes si et seulement s'ils ont la même distribution de degrés et que deux sommets d'un de ces graphlets sont dans la même position s'ils ont le même degré. Ce n'est néanmoins plus vrai à partir de la taille 5, comme le montre l'exemple de la figure 4.17. Les deux chercheurs proposent alors le concept de *voisins-degré)* pour passer à la taille

Algorithme 2 : EnumerateSubgraphs

Données : $G = (V, E)$, k
Résultat : ClustersParK
pour *chaque sommet* $v \in V$ **faire**
  $V_{Extension} = \{u \in N(\{v)\}$ / id$(u) >$ id(v) $\}$
  ExtendSubgraph($\{v\}$, $V_{Extension}$, v, k)

Algorithme 3 : ExtendSubgraph

Données : $V_{Subgraph}, V_{Extension}, v, k$
si V_S *contient* k *sommets* **alors**
  retourner $G|V_{Subgraph}$
tant que $V_{Extension} \neq \emptyset$ **faire**
  Retirer un sommet w arbitrairement choisi depuis $V_{Extension}$
  $V'_{Extension} = V_{Extension} \cup \{N_{exclu}(w, V_{Subgraph})|$ id$(u)>$id$(v)\}$
  ExtendSubgraph($V_{Subgraph} \cup \{w\}$, $V'_{Extension}$, v, k)

FIGURE 4.16 – Algorithme de parcours de tous les sous-graphes de taille k décrit dans [Wernicke, 2006]

5. Le voisins-degré d'un sommet v se caractérise par la formule :

$$vd(v) = d(v) + \sum_{u \in N(v)} (d(u)).$$

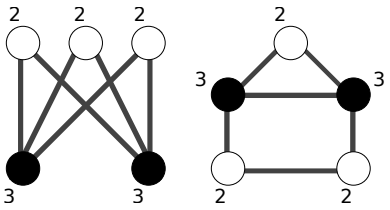


FIGURE 4.17 – Deux graphlets de taille cinq avec deux sommets de degré 3 (en noir) et trois de degré 2 (en blanc) chacun. La distribution des degrés ne permet donc plus de déterminer le graphlet correspondant à un sous-graphe.

Le voisins-degré d'un sommet est donc la somme de son degré et de celui de ses voisins. La figure 4.18 donne les voisins-degré des deux mêmes graphlets que dans l'exemple précédent. Dans le cas des graphlets jusqu'à la taille 5, deux sommets d'un même graphlet ont le même voisins-degré si et seulement s'ils sont dans la même position. Encore une fois, cela n'est plus le cas lorsqu'on passe à la taille 6 comme illustré par

la figure 4.19. La distribution des voisins-degrés est, comme la distribution des degrés, unique pour chaque graphlet jusqu'à la taille 5. Pour chaque sous-graphe visité par le processus de Wernicke, ils calculent donc leur voisins-degré et déduisent ainsi le graphlet qui lui est isomorphe et également les positions de chacun de ses sommets. Décrite tel quel, l'algorithme trouve le graphlet isomorphe à un sous-graphe en temps $O(|E|)$ puisque deux parcours de l'ensemble des arêtes du sous-graphe permettent ainsi de trouver le voisins-degré de chacun de ses sommets.

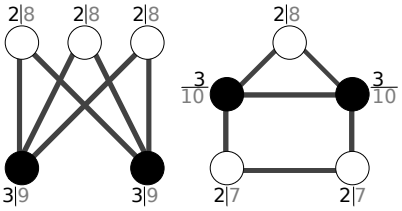


FIGURE 4.18 – Les valeurs des voisins-degré de chaque sommet ont été ajoutées. On voit que les deux réseaux n'ont plus la même distribution de voisins-degré et que les deux sommets blancs du réseau de droite qui ne sont pas dans la même position que celui du haut n'ont pas le même voisins-degré que lui.

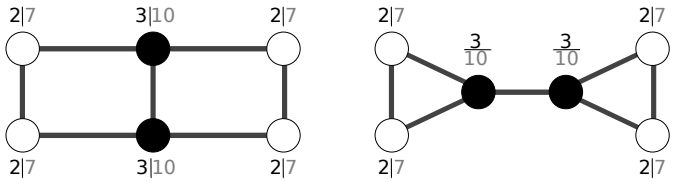


FIGURE 4.19 – Deux graphlets de taille six avec deux sommets de degré 3 et de voisins-degré 10 et quatre de degré 2 et de voisins-degré 7 chacun. La distribution des voisins-degré ne permet donc plus de déterminer le graphlet correspondant à un sous-graphe de taille six.

Stoica et Prieur utilisent cette méthode pour décrire un grand réseau de relations téléphoniques. Ils calculent le réseau égocentré de chaque sommet de ce réseau complet et énumèrent les graphlets, jusqu'à la taille 5, de ces petits réseaux afin de définir ceux qui sont caractéristiques du grand réseau. Ils testent trois méthodes : 1- le nombre d'occurrences du sous-graphe est supérieur à un certain seuil, 2- le nombre de réseaux égocentrés qui contiennent le sous-graphe est supérieur à un certain seuil et 3- Le taux d'apparition du sous-graphe dans ces réseaux est supérieur à celui qu'il a dans des graphes aléatoires de même taille. Cette dernière méthode est d'ailleurs celle proposée par Milo et ses collègues [Milo et al., 2002] qu'on a évoquée plus tôt.

Au début de mon travail de thèse, j'ai moi-même proposé une adaptation de cette méthode de reconnaissance des sous-graphes apportant une amélioration marginale du temps de calcul. En remarquant que la majorité des graphlets de taille 5 avaient une distribution des degrés unique à l'exception de deux graphlets avec des distributions de degrés (2,2,2,3,3) et deux avec pour distribution (1,2,2,2,3) . Pour le premier cas de conflit, il suffit alors de vérifier si les deux sommets de degré 3 sont connectés entre eux, ce qui est vrai dans un cas mais pas dans l'autre . Dans le deuxième cas, il faut voir si le sommet de degré 1 est voisin de celui de degré 3 ou pas . Cette méthode permet donc de déterminer le graphlet en un seul parcours de ses arêtes pour la majorité des cas.

4.5.3 Autres approches

Si l'intuition nous dirige vers un processus basé sur le couple parcours-identification tel que décrit plus haut, certains chercheurs se sont penchés sur l'élaboration de méthodes différentes pour énumérer les graphlets d'un réseau.

Ainsi, Tomaž Hočevar et Janez Demšar de l'université de Ljubljana proposent de compter les orbites de chaque sommet à partir d'une liste d'équations qui décrivent les relations entre le nombre d'apparitions de ces différentes orbites [Hočevar and Demšar, 2014]. Pour décrire les liens entre les orbites, ils partent du principe que chaque graphlet de taille 4 peut être décrit comme un des deux graphlets de taille 3 auquel un sommet a été ajouté. Pour un sommet donné du réseau, le nombre de ses orbites de graphlets de taille 4 peut donc directement être déduit de ses orbites de graphlets de taille 3 ainsi que de la connaissance des liens entre ses voisins et les sommets qui lui sont à distance 2. À partir de ces observations, les deux chercheurs construisent l'ensemble des équations qui lient entre elles les orbites des graphlets de taille 4 et celles qui lient les orbites de taille 5 entre elles. Ils peuvent finalement résoudre l'ensemble de ces équations, c'est-à-dire énumérer les orbites dans lesquelles apparaît chaque sommet, en comptant son nombre d'apparitions dans une seule des orbites.

4.6 Méthodes d'interprétation

L'énumération des graphlets effectuée par l'intermédiaire d'une des méthodes proposées ci-dessus ou une autre de la littérature, reste à utiliser ce résultat. Classiquement, le nombre de fois où chaque graphlet apparaît dans des réseaux est mobilisé pour caractériser des graphes les uns par rapport aux autres ou pour trouver le modèle de graphes aléatoires le plus approprié à un type de données particulier. Quoi qu'il en soit,

la simple énumération n'est généralement pas suffisante puisque très dépendante de la taille des réseaux par exemple. La dernière étape du processus usuel de l'étude de graphes par leurs graphlets consiste donc à transformer le résultat des étapes précédentes en un indicateur plus convenable. De nombreuses distances basées sur cette énumération des graphlets ont ainsi été construites par nos collègues.

Une telle distance, disons d, est alors une fonction prenant en entrée les énumérations des graphlets de deux graphes et retournant un nombre positif (ou nul, dans le cas exclusif où les deux énumérations sont strictement identiques). Comme toutes les distances, elle est symétrique, c'est-à-dire que la distance entre les énumérations de deux graphes G_1 et G_2 est la même que la distance inverse entre les énumérations de G_2 et G_1. Finalement, elle respecte l'inégalité triangulaire, qui stipule que pour trois énumérations de réseaux $\acute{e}_1$, $\acute{e}_2$ et $\acute{e}_3$, $d(\acute{e}_1, \acute{e}_2)+d(\acute{e}_2, \acute{e}_3) \leqslant d(\acute{e}_1, \acute{e}_3)$, ce qui correspond à l'idée que passer par un point intermédiaire ne peut pas raccourcir le chemin.

Avant de présenter la métrique que je propose pour caractériser les réseaux égocentrés de Facebook, voyons donc certaines d'entre celles les plus marquantes de la littérature et qui ont inspiré ce travail.

4.6.1 Relative graphlet frequency

La première que je vais aborder date de 2004 et a été mise au point par Natasa Pržulj et ses collègues de l'université de Toronto et de l'Ontario Cancer Institute. Ils proposent de mesurer la distance entre deux réseaux d'après ce qu'ils nomment la *relative graphlet frequency distance* [Pržulj et al., 2004]. Le terme de « graphlet » est d'ailleurs pour la première fois proposé lors de cette publication, et ce afin d'éviter toute confusion avec le terme de « motif » qu'utilise Milo pour décrire les sous-graphes induits sur-représentés (voir 4.4). À partir d'ici, on ne s'intéresse plus à ces sous-graphes sur-représentés, mais bien à tous.

La distance se base donc sur la *fréquence relative des graphlets*

$$N_i(G)/T(G)$$

où $N_i(G)$ est le nombre de graphlets de type i (avec $i \in \{1, ..., 29\}$) énumérés dans un graphe G et $T(G) = \sum_{i=1}^{29} N_i(G)$ est le nombre total de graphlets énumérés. Cette mesure est assez semblable à la *Concentration* utilisée par les équipes de Milo à la différence près qu'elle ne tient pas compte de la taille des différents graphlets et les ajoute indifféremment dans le nombre total de graphlets énumérés. On remarque par ailleurs que la mesure tient compte de 29 graphlets, en excluant le graphlet arête •—• .

Finalement, la *relative graphlet frequency distance* entre deux graphes G et H est défini comme suit :

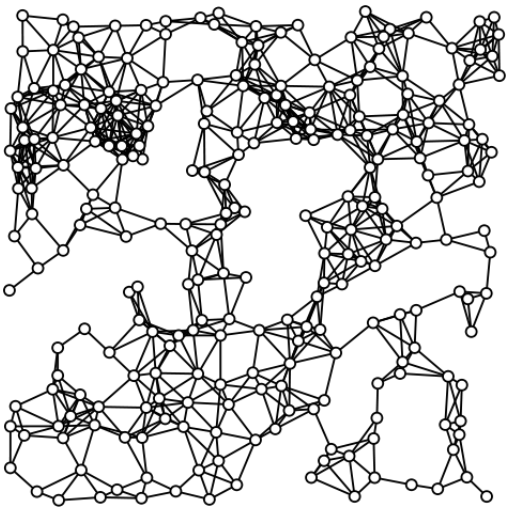


FIGURE 4.20 – Un graphe aléatoire géométrique, tiré de la page Wikipédia dédiée. Les points ont été placés aléatoirement puis ont été reliés si leur distance est inférieure à un certain seuil.

$$D(G,H) = \sum_{i=1}^{29} |F_iG - F_i(H)|,$$

où $F_i(G) = -log(N_i(G)/T(G))$. Le logarithme est ici employé par Pržulj et son équipe afin de lisser les proportions de graphlets. Ils remarquent en effet que ceux-ci apparaissent de manière très hétérogène les uns par rapport aux autres au sein des réseaux d'interactions entre protéines sur lesquels ils travaillent. L'utilisation du $-$ permet d'obtenir une valeur de $F_i(G)$ toujours positive puisque le logarithme d'une proportion (comprise par nature entre 0 et 1) est négatif.

Leurs travaux ont permis de conclure que ces deux réseaux sont plus proches, selon leur distance, des réseaux géométriques aléatoires que de réseaux construits selon d'autres modèles de génération de graphes aléatoires. Un réseau géométrique aléatoire, tel qu'illustré par la figure 4.20, est un réseau tel que ses sommets sont placés aléatoirement dans un espace et dont les arêtes relient les sommets qui sont proches entre eux, selon une distance et une limite arbitrairement choisies.

4.6.2 Graphlet degree distribution agreement

En 2007, Pržulj, qui travaille désormais à l'University of California, publie une nouvelle mesure des similarités entre les réseaux qui utilise pour la première fois les positions automorphiques au sein des graphlets et qu'elle nomme alors les *orbites* [Pržulj, 2007]. Elle les utilise pour définir une élégante généralisation de la distribution de degrés

qu'elle appelle la *Graphlet degree distribution* : Le degré d'un sommet étant en effet son nombre de voisins, il est aussi égal au nombre d'arêtes qui lui sont adjacentes, c'est-à-dire au nombre de fois où il apparaît à l'unique position ou orbite du graphlet arête •—• . À partir de là, il est ainsi possible d'étendre le concept au nombre de fois où un sommet apparaît dans la position périphérique du chemin à trois sommets •••, dans sa position centrale, etc. Ces distributions de graphlet-degrés d'un graphe (dans une version francisée du terme) correspondent ainsi au nombre de fois où chaque sommet apparaît dans chacune des 73 positions des 30 graphlets de taille 5 ou moins.

Pour chaque orbite j, elle note $d_G^j(k)$ le nombre de sommets du graphe G qui apparaissent exactement k fois dans j. $d_G^j = (d_G^j(0), d_G^j(1), ...)$ est donc la distribution de graphlet-degrés (GDD) de l'orbite j dans G. Afin de réduire l'influence des hauts degrés sur les GDD, celle-ci est normalisée une première fois selon la formule

$$S_G^j(k) = \frac{d_G^j(k)}{k},$$

donnant un vecteur normalisé $S_G^j = (S_G^j(1), S_G^j(2), ...)$.

Les composantes de ce vecteur sont ensuite sommées : $T_G^j = \sum_{k=1}^{\infty} S_G^j(k)$.

Et la distribution normalisée des graphlets-degrés est donnée par la formule :

$$N_G^j(k) = \frac{S_G^j(k)}{T_G^j}.$$

La distance, comprise entre 0 et 1, de deux réseaux G et H pour une orbite j donnée est ainsi définie par :

$$D^j(G,H) = \left(\sum_{k=1}^{\infty} [N_G^j(k) - N_H^j(k)]^2 \right)^{1/2}$$

La mesure est en fait une mesure de similarité et pas de distance puisque Pržulj calcule ce qu'elle nomme l'agreement, l'accord, entre deux réseaux :

$$A^j(G,H) = 1 - D^j(G,H).$$

Finalement, le score de proximité entre deux réseaux, toutes orbites prises en compte est la moyenne, au sens arithmétique ou géométrique, de l'ensemble des accords relatifs aux 73 positions des graphlets de taille 2 à 5 :

$$A_{arith}(G,H) = \frac{1}{73} \sum_{j=0}^{72} A^j(G,H),$$

$$A_{geo}(G,H) = \left(\prod_{j=0}^{72} A^j(G,H) \right)^{1/73}.$$

Ces deux mesures lui permettent de confirmer que les modèles de graphes géométriques produisent des réseaux qui sont plus proches des réseaux d'interactions de protéines dont 14 ont en effet été à nouveau comparés à des réseaux issus d'une variété de modèles aléatoires.

4.6.3 graphlet correlation distance

En 2014, une publication pluridisciplinaire dont le premier signataire est Ömer Nebil Yaveroğlu, étudiant en thèse sous la direction de Nataša Pržulj propose une nouvelle méthode de comparaison des réseaux qui prend également en compte les positions de leurs sommets [Yaveroğlu et al., 2014]. Au lieu d'étudier la manière globale dont les graphlet-degrés sont distribués au sein du graphe en perdant, ce faisant, l'information relative aux sommets qui est agrégée, ils vont cette fois utiliser les vecteurs de graphlet-degrés de chaque sommet.

Ils commencent par procéder à des simplifications de cette liste de vecteurs, en remarquant notamment que le graphlet-degré d'un sommet dans l'unique position, qu'ils notent C_3 du triangle est déductible du nombre de fois où ce sommet apparaît en position centrale C_2 dans un chemin à trois sommets et dans l'unique position C_0 du graphlet-arête . Cette déduction se faisant selon la formule $\binom{C_0}{2} = C_2 + C_3$. À partir de remarques similaires, ils obtiennent 11 positions dites non redondantes parmi les graphlets de taille 4 ou moins et à 56 positions non redondantes pour les graphlets de taille 5 ou moins.

Une fois cette étape passée, ils construisent pour chaque réseau étudié, sa matrice de corrélation des positions. Celle-ci est calculée à partir des *coefficients de Spearman*, qu'il est possible de calculer puisque les graphlet-degrés de chaque sommet sont gardés séparément. C'est la distance entre ces matrices qui fait finalement office de distance entre les réseaux. Celle-ci, appelée *Graphlet correlation distance*, est calculée selon la distance euclidienne entre les triangles supérieurs des matrices.

Ils valident cette métrique en montrant que la distance est plus importante entre des réseaux issus de domaines différents ou bien de modèles aléatoires différents que s'ils viennent du même champ, par exemple deux réseaux sociaux (voir 4.21).

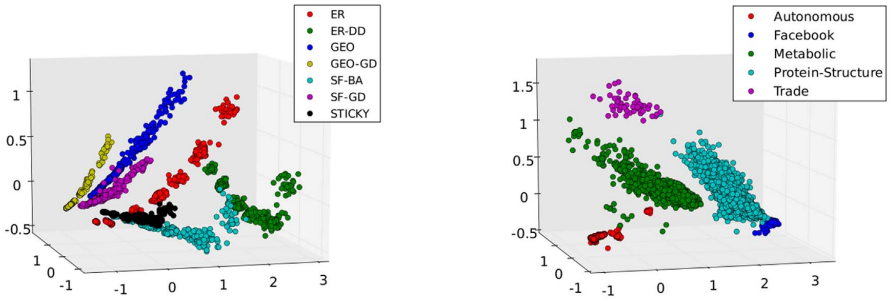


FIGURE 4.21 – Plongements, tirés de [Yaveroğlu et al., 2014], de réseaux distribués selon leurs *graphlet correlation distance*

4.6.4 NetDis

Le dernier modèle que j'aborde est également le plus récent d'entre eux et date de 2014. Une équipe du département de statistiques d'Oxford, travaillant en collaboration avec un programme de recherche en biologie de l'université de Caroline du Sud, propose une méthode de comparaison de réseaux biologiques basée sur les graphlets [Ali et al., 2014]. Ces chercheurs ne devaient pas avoir connaissance des derniers travaux portant sur la graphlet correlation distance [Yaveroğlu et al., 2014] qu'ils ne citent pas. Un article publié l'année suivante, par l'équipe de Nataša Pržulj [Yaveroğlu et al., 2015] les compare d'ailleurs en démontrant la supériorité de leur *graphlet correlation distance*. NetDis est néanmoins intéressante pour nous car elle partage plusieurs similarités avec la mesure que proposée dans cette thèse.

En effet, puisqu'énumérer les graphlets est long pour les grands réseaux que vise à caractériser NetDis, la stratégie adoptée consiste par découper celui-ci en autant de réseaux égocentrés qu'il contient de sommets (de protéines dans ce cas). Même si ceux-ci sont de ce fait très nombreux, il est beaucoup plus rapide de faire le calcul pour tous ces sous-réseaux égocentrés que pour le réseau complet. La méthode compare ensuite tous ces réseaux à des réseaux aléatoires, en invoquant un hypothétique *gold-standard model* problématique, pour finalement décrire le grand réseau.

On peut donc dans déjà voir une analogie entre ce travail et le notre. Le grand réseau d'interactions de protéines représentant alors le réseau de complet de Facebook et les réseaux égocentrés découpés, nos réseaux récupérés par l'application Algopol. À la différence de ce travail, nous nous interdisons cependant de rechercher d'éventuels sommets qui appartiendraient à plusieurs des réseaux à notre disposition, et ce pour des raisons de protection de la vie privée.

Par ailleurs, ces chercheurs ont fait le choix de se limiter aux graphlets d'une taille donnée pour construire leur mesure. Dans les métriques évoquées plus haut, les gra-

phlets considérés étaient tous ceux jusqu'à une certaine taille. À l'inverse, Ali et ses collègues définissent donc une mesure qui prend en quelque sorte la taille des graphlets en paramètre. On verra dans la section suivante que c'est également le choix que j'ai fait.

J'ai présenté dans ce chapitre le concept des graphlets, ces petits réseaux connexes qui une fois énumérés permettent de caractériser les réseaux selon le maillage de leur structure locale. On a vu comment les énumérer et quelles sont les contraintes inhérentes à cette énumération, en temps de calcul, du fait de l'explosion combinatoire associée.

J'ai également abordé comment ces éléments, dont les prémisses de l'étude sont apparues presque aussi rapidement que l'analyse des réseaux en étude des groupes d'individus avant de se structurer dans les années 2000, sont maintenant interprétés dans l'analyse des réseaux.

Chapitre 5

Représentativité des graphlets

Le chapitre précédent expose le concept d'énumération de graphlets. Afin de regrouper les réseaux égocentrés de notre panel selon ce principe, je propose dans ce chapitre une nouvelle métrique d'interprétation du résultat des énumérations de graphlets, que j'appelle la *représentativité des graphlets* ou, plus simplement dans la suite, la *représentativité*.

Je présente dans un premier temps la mesure proposée dans cette thèse afin d'interpréter une énumération des graphlets, dans le but de comparer entre eux des réseaux d'un large corpus. Je présente ensuite son application à notre panel, et également une caractérisation des graphlets eux-mêmes, dans un effort de compréhension plus poussée de l'outil graphlet en soi.

5.1 Les motivations derrière la représentativité

Assez différent de ceux dont on a déjà parlé dans le chapitre précédent, le contexte d'application de l'énumération des graphlets sur un grand nombre de réseau égocentré soulève des contraintes sur lesquelles il semble important de s'attarder quelque peu. A contrario, il offre également des opportunités qui n'apparaissaient pas dans ces contextes précédents.

5.1.1 Rapport aux graphes aléatoires

La majorité des métriques qui utilisent les graphlets se basent, à l'image des motifs de Milo, sur leur nombre moyen d'apparitions dans des réseaux générés aléatoirement

et selon différents modèles. Finalement, ces méthodes finissent parfois par nous en apprendre autant sur ces modèles de génération aléatoires que sur les réseaux de terrain étudiés eux-mêmes.

L'étude de Yaveroğlu montre bien ce problème à travers la figure 4.21. Chacun des différents modèles aléatoires de la projection de gauche a son comportement propre vis-à-vis des graphlets et ces comportements sont également différents de ceux des graphes de terrains qui ont été utilisés pour la partie de droite. Ce problème est également discuté par Artzy-Randrup et ses collègues, en commentaire des articles de l'équipe de Milo [Artzy-Randrup et al., 2004] et mis en avant par la méthode Netdis, évoquée en section 4.6.4, qui n'a d'autre choix que d'invoquer un *gold-standard network* [Ali et al., 2014]. Dans notre cas, le grand nombre de réseaux dont nous disposons nous permet d'éviter d'avoir à employer des modèles de génération aléatoire de graphes.

J'ai plutôt opté pour une comparaison strictement entre eux des réseaux égocentrés. Il est en effet possible qu'ils partagent tous une spécificité par rapport aux réseaux aléatoires qu'on ne souhaite pas particulièrement voir apparaître. Une possibilité serait alors de les comparer à plusieurs réseaux aléatoires construits selon une variété de modèles différents, mais il n'est encore une fois pas certain que cette solution soit optimale puisque les groupes qui émergeraient ainsi seraient dépendant de l'ensemble des modèles choisis.

5.1.2 Données uniformes

Au contraire des études précédentes, les données sur lesquelles nous travaillons sont homogènes, en ce sens que toutes sont des réseaux sociaux, qui plus est construits selon la même manière : une capture de données Facebook. Il ne s'agit donc pas pour nous de capturer un score qui révélerait quoi que ce soit de régulier sur les réseaux de relations sociales mais bien de décrire nos réseaux les uns par rapport aux autres.

Nous partons ainsi d'une idée assez simple qui consiste à penser qu'il est souvent plus intéressant, lorsqu'on étudie un ensemble d'éléments homogène décrits par quelques indicateurs, d'étudier ces valeurs les unes par rapport aux autres et non pas simplement les valeurs brutes. C'est ce principe qui sera appliqué pour les graphlets puisque la littérature a très bien montré que l'énumération de ceux-ci est fortement dépendante du type de réseau auquel on s'intéresse.

5.1.3 Un indicateur simple

Un autre point est le fait que la science des réseaux est un domaine hautement pluridisciplinaire, comme on l'a vu. Il en résulte que des chercheurs issus de disciplines variées peuvent être amenés à s'intéresser aux travaux sur les graphes. Dans cette optique, il nous semble important d'attacher une importance toute particulière au fait de construire des indicateurs qui soient les plus accessibles possible afin de diminuer le coût d'entrée aux chercheurs souhaitant tester de nouvelles méthodologies de recherche.

Nous pensons ainsi que la représentativité est, au delà des méthodes d'énumération parfois peu évidentes, une métrique simple d'accès. Faust a par exemple montré que la simple énumération des dyades agrémentée d'une collection de métriques classiques de l'analyse des réseaux était aussi efficace que l'énumération de leurs triades pour les caractériser [Faust, 2007]. Il semble cependant pertinent que n'étudier qu'un seul indicateur (en considérant par exemple que les 6 graphlets de taille 4 ou les 21 de taille 5 constituent un seul indicateur) améliore la lisibilité de l'analyse plutôt que de combiner de nombreuses métriques.

5.2 Définition formelle

Calcul des fréquences

Soit $\mathcal{G} = \{G_1, ...G_{|\mathcal{G}|}\}$ un ensemble de graphes.

Notons encore $N_i(G)$ comme étant le nombre de fois que le $i^{\text{ème}}$ graphlet est énuméré dans le graphe G. La figure 5.1 récapitule ces indices, tel qu'on va les utiliser dans la suite.

Pour une taille k donnée, on va dénoter par I_k l'ensemble des indices tel que les graphlets correspondants ont taille k.

On note également $k_i = k \,|\, i \in I_k$, la taille du graphlet i.

Pour une valeur de k donnée, on peut alors compter le nombre total de graphlets de taille k dans G :

$$T_k(G) = \sum_{i \in I_k} N_i(G),$$

Cette définition du nombre total de graphlets d'un réseau diffère de celle utilisée par l'équipe de Pržulj et présentée dans la section 4.6.1 [Pržulj et al., 2004]. Celle-ci était

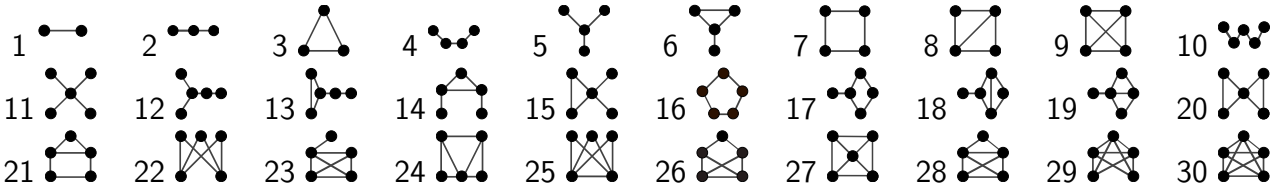


FIGURE 5.1 – Les 30 graphlets jusqu'à la taille 5.

effectivement construite de telle manière que tous les graphlets de taille inférieure ou égale à k étaient comptabilisés.

De manière analogue, on nomme

$$N_i(\mathcal{G}) = \sum_{G \in \mathcal{G}} N_i(G)$$

le nombre total de fois où le $i^{\text{ème}}$ graphlet apparaît dans le corpus $\mathcal{G}$ et

$$T_k(\mathcal{G}) = \sum_{G \in \mathcal{G}} T_k(G)$$

son nombre total de graphlets de taille k.

À partir de ces valeurs on construit la *fréquence globale* du $i^{\text{ème}}$ graphlet définie par le rapport entre la proportion d'apparitions d'un graphlet dans le corpus, par rapport aux autres graphlets de même taille :

$$R_i = \frac{N_i(\mathcal{G})}{T_{k_i}(\mathcal{G})}.$$

La *fréquence locale* du $i^{\text{ème}}$ graphlet dans G est son pendant au niveau d'un réseau :

$$r_i(G) = \frac{N_i(G)}{T_{k_i}(G)}.$$

Les fréquences globale et locale sont toutes les deux comprises entre 0 et 1. Pour une taille donnée, les fréquences globales des graphlets de cette taille pour l'ensemble du corpus somment à 1. Il en va de même pour les fréquences locales des graphlets de même taille d'un réseau.

Fréquences globales de deux corpus

Comme dans la section 3.1.1, on va comparer les réseaux de l'étude de Facebook avec ceux du panel de Caen afin d'avoir de voir si et comment les graphlets permettent de distinguer ces deux corpus fournis par des études différentes. Avant cela je vais faire un court aparté sur la représentation graphique que j'utiliserai dans la suite du manuscrit.

Représentation graphique

Une visualisation accessible et didactique de la description d'un réseau ou d'un ensemble de réseaux, en termes de graphlets, constitue un élément important de la conception de ces métriques, et j'ai eu l'occasion dans le chapitre précédent d'agrémenter le discours de quelques images empruntées aux publications historiques qui montrent que cette visualisation a effectivement évolué au fil du temps.

Les résultats des calculs qu'on va mettre en œuvre seront donc présentés sous la forme de diagrammes de Kiviat, ou diagramme radar, comme dans l'exemple de la figure 5.3. Ces diagrammes placent les différentes variables, c'est-à-dire les 6 graphlets de taille 4 ou les 21 graphlets de taille 5 dans notre cas, en cercle. Chaque individu (chaque réseau égocentré) est représenté par un polygone dont chaque sommet indique la valeur d'une des variables pour cet individu. Le sommet du polygone se rapprochant de l'extérieur du cercle lorsque la valeur grandit et de l'intérieur quand celle-ci est plus proche du minimum.

Les diagrammes radar sont connus pour être délicats à interpréter dans le cas où les valeurs voisines varient fortement les unes par rapport aux autres. Afin d'éviter cet écueil, les graphlets seront positionnés, dans ces diagrammes, selon un ordre permettant de mettre côte à côte ceux ayant des comportements similaires. La méthode pour choisir l'ordre des graphlets est présentée dans la section 5.6.

Des réseaux Facebook moins denses

La figure 5.2 indique ainsi les représentativités globales du panel de Caen ainsi que de notre corpus de réseaux de Facebook. On peut remarquer des différences notables qui recoupent celles qu'on a déjà observées dans la section 3.1.1. Les réseaux de Facebook ont une plus grande proportion de 4-chemins ce qu'on peut rapprocher de leur diamètre nettement plus élevé. Pour les deux graphlets ayant un alter central , c'est, à l'inverse, le panel de Caen qui en possède le plus. Le cintre est d'ailleurs le graphlet plus fréquent au sein des deux corpus. On imagine bien qu'ils sont à rapprocher de la centralisation d'intermédiarité, plus élevée des réseaux de Caen. Dans les deux cas, la fréquence des carrés est très faible, faiblesse qui apparaît être un des marqueurs des réseaux sociaux. Pour ce qui est des deux graphlets les plus denses, le diamant est plus présent dans Facebook et la 4-clique dans les réseaux de Caen, qui sont effectivement très denses, comme on l'a vu. La plus faible présence des diamants, soulève des interrogations puisque ce graphlet est néanmoins l'un des plus denses. On peut imaginer que c'est la structure particulière des petits réseaux personnels qui en est la cause, ce que confirment les graphlets de taille 5.

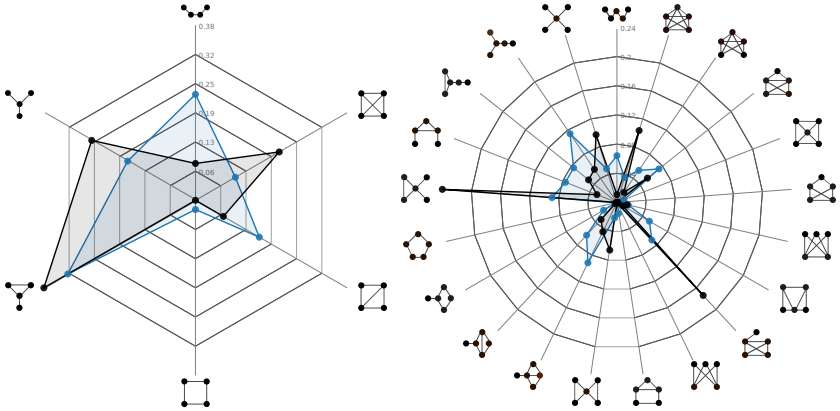


FIGURE 5.2 – La forme bleue représente le corpus de Facebook et la noire celui de Caen. Chaque ligne partant du centre représente la fréquence globale du graphlet vers lequel elle pointe, plus l'intersection avec la forme étant éloignée du centre et plus elle est élevée.

La fréquence globale de ceux-ci pour le panel de Caen est très hétérogène. Certains, comme l'étoile , la fausse étoile , la clique-plus-un et la 5-clique sont beaucoup plus fréquents que les autres. Les graphlets capturent donc bien le fait que les réseaux de Caen sont généralement formés de communautés très denses reliées entre elles par un ou quelques alters très centraux et retranscrivent la particularité des réseaux personnels construits par questionnaires. On a déjà parlé, dans la section 1.4.3, de leur propension à se centrer autour de sphères sociales au sein desquelles ego considère que la majorité des membres interagissent entre eux. De l'autre côté, les réseaux de Facebook, plus grands, sont plus équilibrés avec beaucoup de réseaux qu'on pourrait qualifier d'intermédiaires et moins de présence des graphlets archétypaux .

On peut par ailleurs noter que la représentation en radar que nous conservons par la suite ne s'adapte pas bien au fait que les valeurs des fréquences de graphlets de taille 5 soient très différentes les unes des autres, et on discerne mal les différences pour les moins fréquents d'entre eux. La représentativité des graphlets, dont je reprends le fil de la construction à la suite de cet aparté, va gommer ce problème.

Calcul de la représentativité

L'étape suivante de la construction de la *représentativité* du $i^{\text{ème}}$ graphlet dans le réseau G consiste à en calculer la représentativité non normalisée, le rapport entre sa

fréquence locale dans G et sa fréquence globale.

$$\rho'_i(G) = \frac{r_i(G)}{R_i}.$$

Certains graphlets ont une très faible fréquence globale mais peuvent parfois avoir une fréquence locale sinon importante, au moins nettement plus élevée. Cela impliquerait alors que le calcul du ratio entre ces deux valeurs aboutirait à un score qui, lui, pourraient être très important. C'est pourquoi on applique une étape de normalisation de cette valeur pour la restreindre entre 0 et 2. Cette valeur, la *représentativité*, se définit ainsi finalement, pour le $i^{\text{ème}}$ graphlet, dans le réseau G, comme suit :

$$\rho_i(G) = \begin{cases} \rho'_i(G) & \text{si } \rho'_i(G) \leqslant 1 \\ 2 - \frac{1}{\rho'_i(G)} & \text{sinon.} \end{cases}$$

Si $\rho_i(G) = 1$, le graphlet i apparaît dans G avec la même proportion exactement que dans $\mathcal{G}$. Si cette valeur est inférieure à 1, on dit alors que le graphlet est sous-représenté dans ce graphe par rapport au corpus et dans le cas contraire, qu'il est sur-représenté.

Adaptations de la mesure

La représentativité des graphlets, telle que définie, peut aisément être adaptée à un ensemble de graphes plutôt qu'à un seul réseau. Il faut pour cela repartir d'une définition de la fréquence locale d'un graphlet, qu'on applique cette fois à un ensemble de réseaux. Si $\mathcal{G}' \subset \mathcal{G}$ est un sous-ensemble du corpus de graphes, alors

$$r_i(\mathcal{G}') = \frac{N_i(\mathcal{G}')}{T_{k_i}(\mathcal{G}')}$$

est la représentativité locale du $i^{\text{ème}}$ graphlet dans le sous-ensemble $\mathcal{G}'$ de $\mathcal{G}$. On peut alors la diviser par la même valeur R_i puis la normaliser comme ci-dessus pour obtenir la représentativité d'un graphlet dans un sous-ensemble.

De la même manière, on peut également considérer la représentativité d'un groupe de graphlets de même taille plutôt que d'un graphlet. Notons I l'ensemble des indices de ces graphlets et k_I leur taille. On doit alors prendre comme représentativité locale de ces graphlets

$$r_I(G) = \frac{\sum_{i \in I} N_i(G)}{T_{k_I}(G)}$$

et comme représentativité globale

$$R_I(\mathcal{G}) = \frac{\sum_{i \in I} N_i(\mathcal{G})}{T_{k_I}(\mathcal{G})}$$

Finalement, on peut combiner ces deux adaptations pour, comme on le fera dans le chapitre 6, calculer la représentativité d'un groupe de graphlets dont I est l'ensemble des indices, dans un groupe de réseaux $\mathcal{G}'$. Dans ce cas,

$$r_I(\mathcal{G}') = \frac{\sum_{i \in I} N_i(\mathcal{G}')}{T_{k_I}(\mathcal{G}')}$$

et la représentativité globale est la même que précédemment

$$R_I(\mathcal{G}) = \frac{\sum_{i \in I} N_i(\mathcal{G})}{T_{k_I}(\mathcal{G})}$$

5.3 Menaces à la validité

La représentativité des graphlets, on va le voir, répond parfaitement à nos attentes en ce qui concerne l'étude de nos réseaux et leur regroupement selon leur forme. On peut néanmoins soulever quelques questions concernant son utilisation dans d'autres contextes. Nous ne sommes malheureusement pas encore en mesure d'apporter une réponse à toutes.

5.3.1 Taille du corpus

La taille minimale du corpus peut-être une limite à l'efficacité de la métrique. On peut imaginer qu'en dessous d'une dizaine ou de quelques dizaines de réseaux à comparer, la représentativité des graphlets ne permette pas d'observer des différences pertinentes. On perd en effet nécessairement en précision par rapport à ce à quoi ressemble un réseau social sur Facebook lorsqu'on diminue leur nombre.

Plusieurs expériences peuvent permettre de répondre à cette question : on peut essayer de reconstruire les clusters de l'ensemble du corpus en prenant plusieurs sous-ensembles de réseaux et voir si les groupes obtenus sont similaires.

Une fois les groupes de réseaux similaires construits, il serait intéressant de voir comment la représentativité de ces réseaux se comporte si on calcule la représentativité globale à partir de ce groupe au lieu de prendre l'ensemble du corpus. Est-ce qu'on retrouverait les mêmes groupes que dans l'ensemble du corpus ou bien y aurait-il des sous-groupes spécifiques à ce groupe ?

5.3.2 Outliers potentiels

On parle d'un *outlier* d'une métrique pour un objet ayant une valeur très significativement différente à celles de la majorité des autres. Dans le cas des graphlets, on a déjà dit que les réseaux avec beaucoup de composantes connexes pouvaient se comporter de manière particulière. De la même manière que la représentativité des graphlets est plus parlante lorsque la taille du corpus augmente, la fréquence des graphlets d'un réseau individuel est plus précise lorsque la taille de celui-ci croît.

D'autres réseaux qui pourrait être délicats à interpréter sont les réseaux très petits, ou très peu denses. Ces réseaux ont en effet un plus petit nombre de graphlets induits dont certains pourrait alors rapidement se retrouver sous- ou sur- représentés. Une vérification ultérieure de cette hypothèse semble pertinente. Elle pourrait alors s'opérer par la comparaison des valeurs de représentativité selon la taille des réseaux.

C'est en partie cette intuition, confirmée par nos observations qui nous a amené à procéder à un recentrage des valeurs de sur-représentation entre 1 et 2. Certains petits réseaux se retrouvaient ainsi avoir quelques représentativités qui explosaient (avec parfois des résultats à 5 ou 6 chiffres) et rendaient ainsi moins efficaces les algorithmes de clusterings.

5.3.3 Réseaux hétérogènes

La représentativité a été construite pour être utilisée dans le cadre d'une étude portant sur beaucoup de réseaux de même origine et elle n'a pas encore été utilisée pour un corpus hétérogène de réseaux modélisant des données d'origines diverses. Il serait, de fait, très intéressant d'avoir l'occasion de tenter l'expérience et de comparer ces résultats à ceux de Yaveroglu, rapportés en figure 4.21 [Yaveroğlu et al., 2014] ou de Milo, en figure 4.9 [Milo et al., 2004]. Quelle que soit l'issue d'une telle expérience, il serait pertinent de faire l'analyse du comportement de la métrique.

5.4 Retour sur la représentation graphique

C'est désormais la représentativité des graphlets de chaque réseau ou de chaque groupe de réseaux qu'on va utiliser comme valeur d'entrée des diagrammes radar. Celle-ci étant normalisée entre 0 et 2, elle permet d'éviter d'avoir des variables trop différentes entre elles et le diagramme radar correspond bien à ce cas de figure où toutes les variables ont la même échelle.

La figure 5.3 représente un exemple de représentativité d'un réseau. la valeur 1 qui correspond à une équivalence de représentativité, pour un graphlet, entre l'individu et le corpus est désormais mise en valeur par une ligne noire plus marquée que les autres.

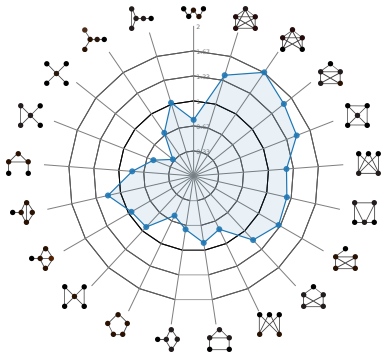


FIGURE 5.3 – Dans ce diagramme qui présente les représentativités des motifs de taille 5 d'un réseau, le cercle noir foncé indique le 1 correspondant à une représentativité identique à celle du corpus. On voit que le réseau a une sur-représentation des graphlets affichés en haut à droite du cercle et a une sous-représentation de et par exemple.

Pour présenter les différents groupes de réseaux que les algorithmes de clustering vont fournir, on va simplement afficher une forme d'une couleur différente pour chaque ensemble qui suivra ses valeurs de représentativité de cet ensemble, comme on l'a déjà fait pour les panels de Caen et Facebook et comme dans la figure 5.4.

Dans les deux radars présentés en figures 5.3 et 5.4, c'est donc la représentativité de chaque graphlet qui sert à calculer les intersections entre les polygones et les axes de chaque variable. En effet, on a vu que les réseaux sociaux ont, entre eux, des valeurs de nombres de graphlets globalement similaires. Il en résulte que les indicateurs, par ailleurs utiles pour les comparer à d'autres types de réseaux, ne permettent pas de bien distinguer leurs différences. La figure 5.5 présente ainsi deux autres visualisations, employant d' autres mesures envisageables et pour lesquelles on remarque qu'il est malaisé de lire des différences pour certains graphlets, bien que la normalisation dans la figure de droite améliore déjà légèrement le rendu.

5.5 Description générale du corpus

Le temps de calcul très important du nombre de graphlets de taille 5 nous a malheureusement imposé de ne prendre en compte que les réseaux de moins de 150 sommets.

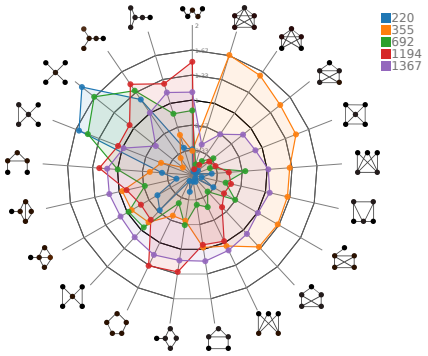
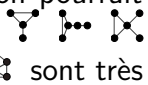
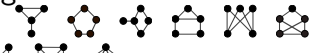
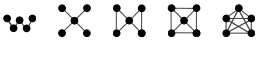


FIGURE 5.4 – Dans ce clustering en 5 groupes, chacun est représenté par une de forme de couleur différente et le nombre de réseaux qu'il compte est indiqué dans la légende à droite. Par exemple, le cluster orange à 355 réseaux a une forte représentativité en graphlets denses.

Il est évident que la mise à l'écart des grands réseaux est un regrettable mais la section 6.1 propose une méthode pour en réintégrer un grand nombre. De plus nous souhaitions que nos analyses puissent être faites sur des ordinateurs classiques et l'intérêt de l'ajout de très grands réseaux relativement rares en sociologie ne nous a pas poussé à effectuer le saut technique que nécessiterait la parallélisation des opérations.

Par ailleurs, nous avons également décidé d'exclure les comptes Facebook possédant moins de 15 amis, car nous avons jugé qu'en dessous de ce seuil, la méthode n'était pas particulièrement adaptée, du fait du trop petit nombre de graphlets à y énumérer. Au final, le corpus que nous allons étudier comporte 3 694 réseaux.

Comme on l'a déjà mentionné, les réseaux sociaux, *a fortiori* lorsqu'ils sont tous construits de manière homogène comme dans notre cas, ont généralement des particularités communes qui se traduisent sur les graphlets qu'on y trouve de manière fréquente (les motifs) mais également sur ceux qui en sont généralement absents.

La figure 5.6 met ainsi en avant les disparités entre graphlets dans les réseaux de Facebook. On voit bien que les plus présents ont souvent des profils qu'on pourrait qualifier de mixtes, avec des triangles induits mais également des chemins  . À l'inverse, les graphlets avec des trous induits  sont très rares, tandis que les motifs archétypaux  sont généralement peu présents.

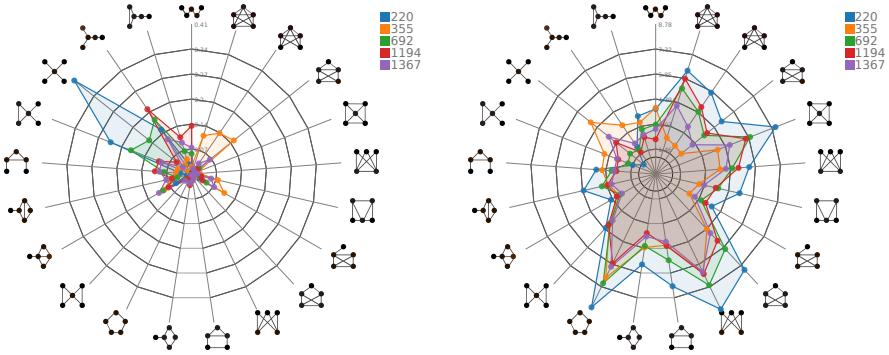


FIGURE 5.5 – Deux diagrammes présentant le même regroupement que celui de la figure 5.4 mais construits avec une autre métrique que la représentativité. À gauche, les valeurs moyennes des fréquences relatives de chaque graphlet ont été calculées pour chaque cluster, à droite c'est cette valeur à laquelle on a appliqué une normalisation en $-log(x)$. Ces deux mesures sont celles utilisées par Pržulj [Pržulj et al., 2004, Pržulj, 2007].

5.6 Les groupes de graphlets

La première utilisation que nous allons faire de la représentativité des graphlets de nos réseaux consiste en l'analyse des graphlets eux-même, et ce dans le but de réduire la difficulté à venir de l'interprétation des réseaux personnels.

On imagine ainsi qu'il est possible de regrouper des graphlets qui sont souvent sur-représentés ou sous-représentés de manière concomitante dans les réseaux égocentrés. Il sera en effet beaucoup plus simple d'interpréter la sur-représentation d'un groupe de graphlets dans un réseau plutôt que d'un seul. De la même manière, on sera plus sensible à la vue de différence de représentativités entre plusieurs graphlets d'un même groupe.

Certains graphlets sont signifiants, comme l'étoile qui suggère une relation spéciale entre ego et l'alter central, ou bien la clique qui lorsqu'elle est sur-représentée dénote l'extrême densité du réseau, marquent des situations qu'il est naturel de qualifier. Pour la majorité des autres, à l'inverse, la lecture des représentativités est nettement moins intuitive et le fait de trouver des similarités entre ceux-ci et les graphlets plus lisibles sera donc d'une aide précieuse dans l'interprétation des résultats.

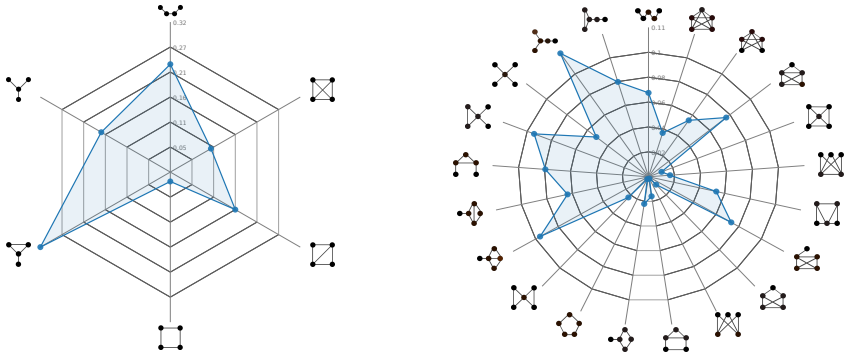


FIGURE 5.6 – Fréquences globales des motifs de taille 4 à gauche et 5 à droite. Ce sont donc ces valeurs qui vont servir de base au calcul des représentativités des graphlets dans nos réseaux.

5.6.1 Méthode d'analyse

Puisqu'on a calculé la représentativité de chaque graphlet de taille 5 pour nos réseaux, on peut appliquer la méthode des kMeans, vue en section 1.6, aux 21 vecteurs de taille 3 694, qui représentent chacun un graphlet vu au prisme de son nombre d'apparitions dans chaque réseau.

Puisque la méthode des kMeans n'est pas déterministe, nous avons lancé 100 itérations pour chaque valeur de k entre 4 et 6. Pour chacune d'entre elles, nous avons calculé la valeur de sa *silhouette*.

Pour chaque valeur de k testée, nous avons conservé le regroupement avec la plus haute silhouette moyenne. Ces groupes sont présentés en figure 5.7. Il est à noter que le score de silhouette maximum correspond au cas où chaque cluster est composé d'un seul membre. Compte tenu de cela on préfère utiliser la silhouette pour départager les clusterings de même taille plutôt que pour choisir le nombre de différents clusters qu'on conserve.

L'application des méthodes du coude et de la silhouette sur les regroupements en différents nombres de groupes sont présentées en figure 5.8. Elles semblent indiquer, bien que sans paraître déterminantes, que le regroupement en 5 catégories de graphlets constitue un bon choix. Plusieurs raisons empiriques nous poussent également à le conserver.

Le regroupement en 4 classes, de silhouette similaire, est très déséquilibré avec un dernier groupe qui comprend près de la moitié des graphlets tandis que le regroupement en 6 classes a une catégorie qui ne comprend que la fourche , ce qui traduit probablement un comportement particulier de ce graphlet, à l'intermédiaire entre les

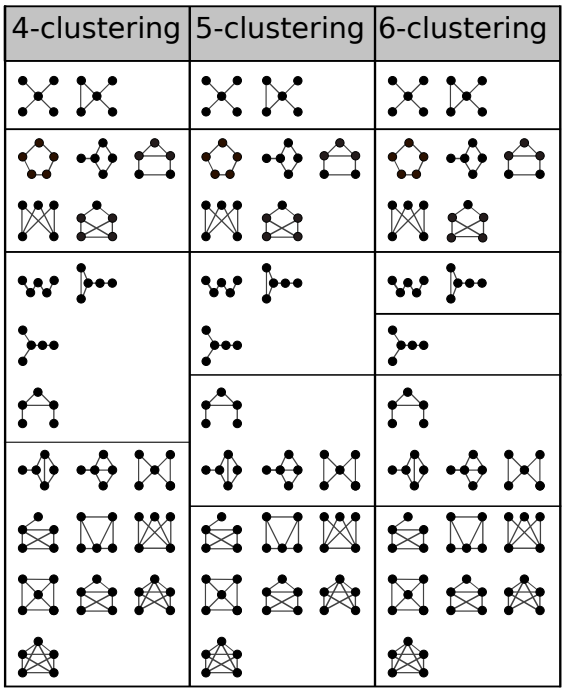


FIGURE 5.7 – Les meilleurs clusters de graphlets obtenus avec la méthode des kMeans for k = 4,5,6, selon leur score de silhouette. Les lignes verticales séparent les clusterings de taille 4, 5 et 6 tandis que les lignes horizontales séparent les groupes de chacun d'eux.

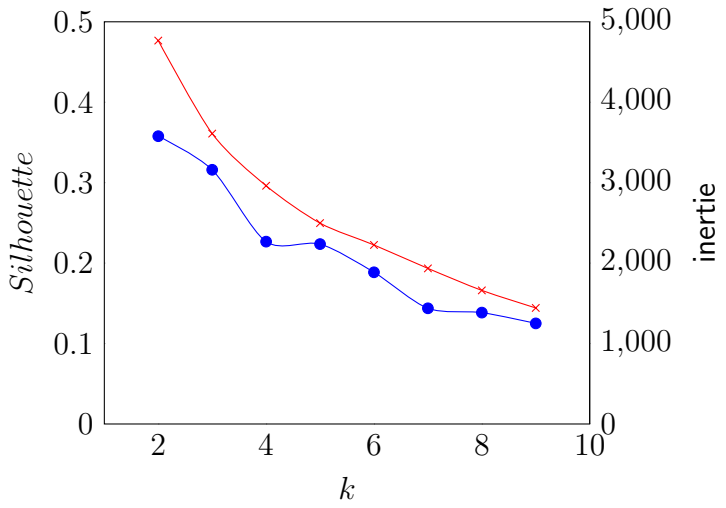


FIGURE 5.8 – La silhouette (en bleu, marquée par des cercle) et l'inertie (en rouge, marquée par des croix) des clusterings de graphlets, selon leur nombre de groupes.

fausses étoiles et les chemins . Le déséquilibre qu'il induit entre les tailles des groupes ne paraît en tous cas pas souhaitable.

De manière générale, pour ce regroupement, ainsi que pour les autres du manuscrit, une possibilité future de choix du nombre de clusters est d'appliquer préalablement à nos données un algorithme de regroupement de type hiérarchique, sur lequel la méthode du coude est connue pour fonctionner efficacement, afin de sélectionner la valeur k d'un kMeans effectué à la suite.

5.6.2 Les groupes des graphlets

Je vais donc maintenant présenter plus en détail ces groupes. Avant tout, notons qu'il ne faut pas surinterpréter ces catégories de graphlets, il ne s'agit *a priori* pas de catégories que l'on retrouverait de manière transversale dans la science des réseaux, mais bien de groupes construits à la lumière d'un corpus de réseaux sociaux égocentrés. Malgré cela, leur répartition suggère une certaine consistance de ces groupes, comme on le verra. Le processus de construction devrait cependant être reproduit sur d'autres types de données afin de valider ces groupes au delà de notre cas d'étude. La table 5.1 indique les moyennes des valeurs de quelques indicateurs pour chacun de ces clusters, tandis que la table 5.2 dénombre l'ensemble des graphlets de taille 4 qui leur sont induits.

Les chemins

Ce cluster contient les graphlets ressemblant à de longs chemins et tous induisent exactement deux fois le chemin de taille 4 . Aucun d'entre eux ne contient cependant de cycle de taille 4 comme sous-graphe, et le porte-manteau est même le seul à comporter un triangle induit. Ces motifs sont ainsi très peu denses et ceux avec les diamètres d'assez loin les plus importants. Ils ont en outre une centralisation d'intermédiarité plus élevée que d'autres groupes et la plus forte modularité du clustering.

La fourche est détachée des deux autres et forme seule un des six groupes pour $k = 6$. Comme mentionné plus haut, cela semble dû au fait que celle-ci est également proche des étoilés et qu'il doit exister un nombre relativement important de réseaux pour lesquelles elle a ses représentativités à un niveau intermédiaire entre ces deux familles, bien qu'elle soit dans la majorité des cas, plus proche des chemins, comme on le verra dans la suite.

Les étoilés

Deux graphlets sont regroupés quelle que soit la valeur de k que l'on ait évalué, et ce groupe doit donc être consistant. Ils ont tous les deux un alter très central entouré d'amis qui ne se connaissent pas, ou alors peu entre eux. L'étoile contient quatre étoiles à trois branches induites et la fausse étoile , deux. Ces deux graphlets n'ont par ailleurs aucun carré et un seul triangle induits à eux deux. Comme les chemins, ils sont peu denses, mais leur principale spécificité réside dans leur très forte centralisation d'intermédiarité.

On aurait de prime abord pu imaginer que le nœud papillon soit dans le même groupe car il paraît assez similaire mais cette absence souligne sa différence en termes de sous-graphes induits. Il ne possède en effet aucune étoile à trois branches induite et 4 cintres .

Les triangulés

Ils possèdent tous au moins un triangle induit mais ont malgré tout une relativement faible densité, mais une transitivité nettement plus importante. En effet, avec au

maximum deux triangles induits par graphlet, ils ne peuvent pas avoir plus d'une fois le diamant comme sous-graphe alors que celui-ci est le second graphlet de taille 4 le plus dense. Ils ont également une quelques sommets centraux , ce que traduit leur centralisation d'intermédiarité relativement élevée.

La présence du nœud papillon parmi eux laisse à penser que celui-ci ne marque pas particulièrement une centralité importante de son sommet joignant les deux ailes, sauf éventuellement au sein d'un groupe d'alters relativement peu dense.

L'interprétation de ces graphlets pose question, et notamment pour savoir si, lorsqu'ils apparaissent, ils se retrouvent le plus souvent être circonscrits à un groupe peu dense ou bien si les sommets qui les composent appartiennent à des communautés différentes du réseau, ce qui justifierait leur absence de corde.

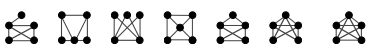
Les troués

Ces graphlets sont ceux qui contiennent ce qu'on appelle des cycles sans corde, ou trou, c'est-à-dire une succession de sommets de degré 2 qui forment un cycle. Cette structure est particulière dans le cas des réseaux sociaux où elle est rare, comme on l'a vu dans la figure 5.6 notamment. Tous les graphlets induisant un cycle sans corde sont d'ailleurs dans cette catégorie, à l'exception de la pyramide , avec ses 4 diamants induits, et dont on peut imaginer qu'elle apparaît alors dans d'autres circonstances.

Il en résulte naturellement la plus faible transitivité parmi les groupes ainsi qu'une centralisation d'intermédiarité également faible. Avec peu de sommets de degré inférieurs à deux, ces graphlets ont néanmoins une densité relativement importante.

On notera qu'une particularité intéressante des troués dans le fait que le cycle de taille 5 soit le seul à n'avoir aucun carré induit. On imagine donc que les gens ayant plus de cycles de taille 4 dans leur réseau personnel ont également plus de cycles de taille 5.

les densifiés



Ces graphlets sont ceux avec la plus forte densité et transitivité. Ils sont les seuls à induire des cliques ainsi que la majorité des diamants . Ce cluster est également celui qui regroupe le plus de graphlets et on aurait pu supposer qu'il ait été recomposé en deux dans le kMeans à $k = 6$. La clique et la clique moins un n'ont en

	chemins	étoilés	triangulés	troués	densifiés
diamètre	3.33	2.0	2.5	2.2	1.86
densité	0.43	0.45	0.58	0.58	0.8
transitivité	0.17	0.19	0.54	0.16	0.77
centralisation d'intermédiarité	0.37	0.61	0.34	0.11	0.11
modularité	0.22	0.04	0.08	0.06	0.01

TABLE 5.1 – Les moyennes de quelques variables classiques décrivant les clusters de graphlets

chemins	6	1	1	0	0	0
étoilés	0	6	2	0	0	0
triangulés	3	1	9	0	2	0
troués	9	3	4	7	1	0
densifiés	1	2	7	1	14	9

TABLE 5.2 – Les sous-graphlets induits de chaque famille de graphlets

effet pas de diamant induit tandis que l'enveloppe fermée , l'opéra de Sydney et la pyramide n'ont pas de 4-cliques induites.

Les graphlets densifiés ont la plus faible modularité moyenne des graphlets ainsi que la plus faible centralisation d'intermédiarité. Leurs sommets sont donc globalement comparables les uns aux autres, il est difficile de les hiérarchiser ou de dégager des communautés différentes dans ces graphlets.

En se rappelant que la clique , comme les autres motifs archétypaux, n'a pas une représentativité globale très importante, on peut déduire que les groupes denses d'alters, dont cette sur-représentation est un marqueur, sont généralement composites. Ceux-ci sont alors composés d'individus plus centraux se trouvant éventuellement au centre de pyramides et ayant même parfois présenté leurs propres amis à *ego* ou bien rencontré les siens issus d'autres sphères sociales . Il y a donc également des personnes plus isolées au sein des groupes et favorisant la sur-représentation d'opéras de Sydney et d'enveloppes ouvertes .

5.6.3 Deux niveaux de relations entre graphlets

Globalement, on remarque une tendance forte des graphlets à se regrouper selon leurs sous-graphes induits avec notamment quelques graphlets de taille 4 qui semblent, d'une

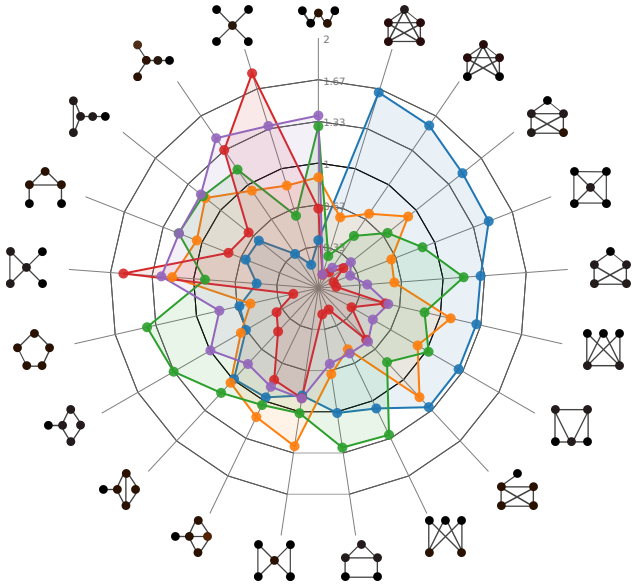


FIGURE 5.9 – Un diagramme radar autour duquel les graphlets ne sont pas positionnés selon un ordre facilitant la lecture.

certaine manière, imposer le regroupement des graphlets concernés. C'est par exemple le cas de la clique mais on note également certaines exceptions comme le 5-cycle ou les deux sous-familles rassemblées de graphlets denses.

Une future et intéressante hypothèse de travail pourrait être de vérifier une conjecture proche de celle vérifiée par Yaveroglu et al. [Yaveroğlu et al., 2014] concernant les positions. On pourrait très bien imaginer que, quel que soit le corpus de réseaux étudié, on retrouve généralement les mêmes clusters de graphlets déterminés par leurs sous-graphes mais que la nature du corpus influe, à la marge, sur certains de ces regroupements.

Les diagrammes radar présentés font la part belle à ces clusters de graphlets puisque les membres d'un même groupe sont systématiquement positionnés côte à côte autour du cercle. Cette astuce permet de mieux voir comment les représentativités des réseaux se distribuent parmi les clusters. La figure 5.9 présente ainsi un diagramme radar où les graphlets sont positionnés selon l'ordre usuel, qui dépend de leur densité. On remarque que seul le groupe de réseaux qui possèdent les graphlets les plus denses, qui sont placés ensemble, est aisé à lire.

5.7 Les formes de réseaux personnels

Dans cette section, on va enfin voir de quelle manière les réseaux sociaux se regroupent en familles selon la représentativité de leurs graphlets.

5.7.1 Le regroupement par les métriques de la littérature

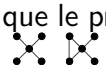
On a vu dans le chapitre précédent que plusieurs grilles de lecture de l'énumération des graphlets avaient déjà été proposées dans la littérature, et avec elles autant de métriques utilisables. La page internet de Nataša Pržulj héberge un logiciel qui, à partir de l'énumération des graphlets d'un ensemble de réseaux, produit plusieurs grilles de distances entre eux[1]. Nous avons grâce à cela pu appliquer plusieurs méthodes de clustering à ces résultats pour comparer les différents regroupements.

Relative graphlet frequency

Le clustering effectué à partir des distances entre les réseaux selon leurs fréquences relatives de graphlets, première méthode proposée par Pržulj [Pržulj et al., 2004] est représenté en figure 5.10, à gauche. La méthode a déjà été présentée en section 4.6.1.

Ce clustering effectue des découpages efficaces entre les réseaux, si on en croit la représentativité de ses groupes. Un groupe intéressant est par exemple celui en orange, de 355 membres avec une sur-représentativité des graphlets denses, où il est le seul à avoir un score important. Malheureusement, on voit que le cluster le plus important, en violet avec 1 1367 membres a la majorité de ses représentativités qui sont proches de 1.

En fait ce groupe est le seul, parmi les quatre à être sous-représentés en graphlets denses, à être également sous-représentés en étoiles . Le problème est donc le suivant : un grand nombre de réseaux sont regroupés car ils n'ont pas beaucoup de graphlets denses et peu d'étoiles, et ces réseaux sont différents les uns des autres pour la majorité des autres graphlets, puisque la représentativité du groupe y est proche de 1.

Par ailleurs, les groupes 220-bleu et 692-vert sont très proches l'un de l'autre mis à part le fait que le premier concentre ceux qui ont une encore plus forte représentativité des étoilés  et donc une moins forte des autres.

1. http ://www0.cs.ucl.ac.uk/staff/natasa/GCD/

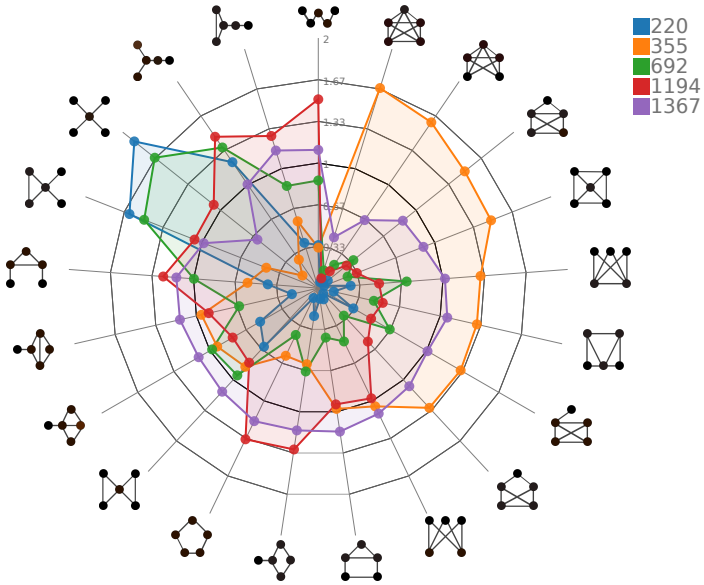


FIGURE 5.10 – Les groupes obtenus à partir des fréquences relatives des graphlets de chaque réseau.

Graphlet degree distribution agreement

Le graphlet degree distribution agreement, également proposé par Pržulj et son équipe [Pržulj, 2007] découpe les groupes de réseaux comme présentés en figure 5.11). Le cluster le plus important, dont les 1 360 réseaux ont une sur-représentation des cliques et des quasi-cliques , mais pas particulièrement des autres graphlets densifiés a néanmoins encore la majorité de ses représentativités autour de la valeur normale 1. Et à quelques rares exemples près, les autres groupes ne s'en écartent guère plus, le groupe violet à 817 réseaux n'offre même aucune accroche à l'interprétation, selon ce modèle qu'on a choisi.

On retrouve néanmoins un groupe en rouge, avec 940 réseaux qui ont une sur-représentation de réseaux étoilés , ainsi qu'à moindre mesure de chemins . Le groupe orange à 379 membres a lui beaucoup de chemins mais peu d'étoiles. Ces deux groupes ont par ailleurs peu de graphlets densifiés. Un dernier cluster, en vert et avec 332 réseaux membres se distingue surtout par une sous-représentation forte en étoiles mais ses autres valeurs ne permettent pas de conclure outre mesure.

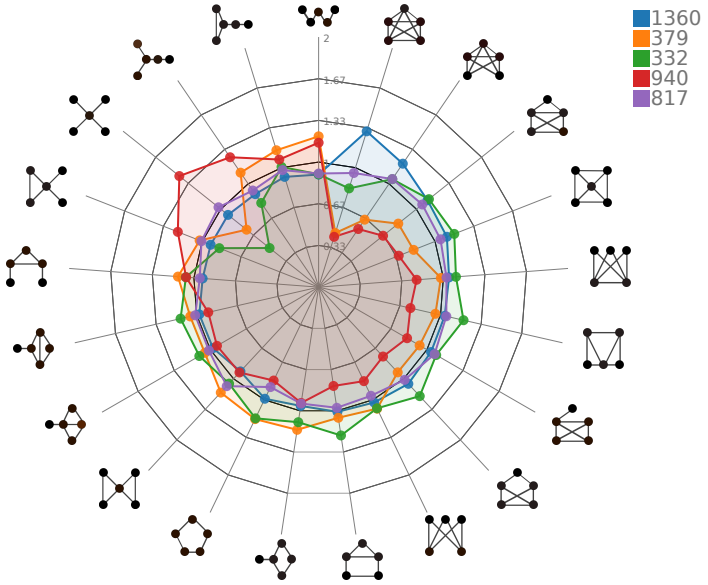


FIGURE 5.11 – Les groupes obtenus à partir des graphlets degree distribution agreements entre les réseaux.

Graphlet correlation distance

Les clusters obtenus par la métrique basée sur les corrélations entre les positions, mise au point par Yaveroğlu au sein de l'équipe de Pržulj [Yaveroğlu et al., 2014] est présentée en figure 5.12. On y retrouve deux groupes de réseaux avec des graphlets densifiés, un qui a plus de chemins et deux qui ont un peu plus de graphlets étoilés.

Comme pour les clusterings précédents, ce dernier pose certains problèmes pour analyser nos réseaux. Le groupe orange à 525 réseaux et le vert à 1 078 membres ont tous les deux beaucoup de représentativités autour de 1. D'un autre côté, les deux groupes rouge à 164 et violet à 427, qui contiennent les graphlets densifiés, se ressemblent beaucoup. Le plus petit des deux semblant être le regroupement de réseaux qui auraient pu être dans le cluster violet mais qui sont encore plus densifiés (et de fait ont des représentativités plus faibles des autres graphlets).

5.8 Les clusters de la représentativité

Maintenant qu'on a vu les regroupements obtenus par l'application de plusieurs des métriques développées par les équipes de Pržulj, on va pouvoir comparer à ceux produits

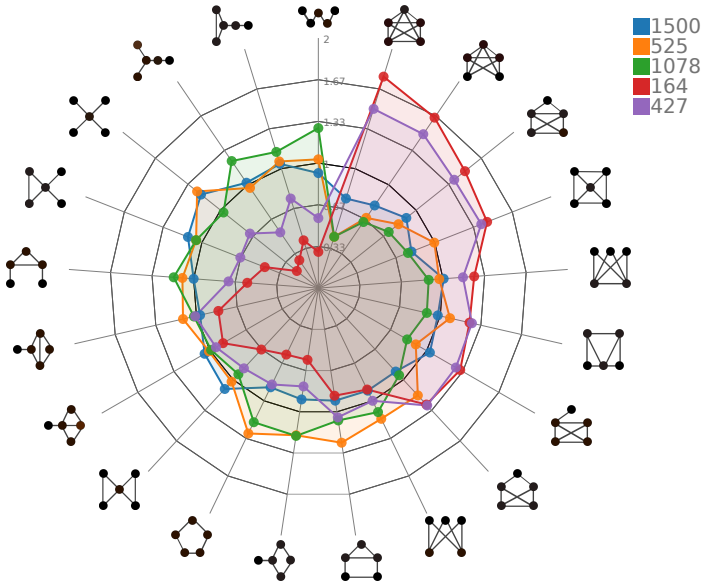


FIGURE 5.12 – Les groupes obtenus à partir des graphlets correlation distances entre les réseaux.

par la représentation des réseaux par leurs représentativités de graphlets. Encore une fois, la silhouette et le coude semblent converger vers un choix en cinq groupes de réseaux, comme l'indique la figure 5.13.

5.8.1 Les clusters

Nous avons choisi de découper notre ensemble de réseaux en cinq groupes. En appliquant toujours 100 occurrences du kMeans aux représentativités des graphlets et en conservant le regroupement avec le meilleur score de silhouette, comme on l'a déjà fait, on obtient les groupes visuellement présentés en figure 5.14.

Contrairement aux clusterings vus précédemment, chacun des clusters de ce dernier a un ensemble de valeurs de représentativités qui lui est propre et s'écartent généreusement de la ligne noire foncée du 1 pour la majorité des variables. Les réseaux sont en outre distribués de manière équitable entre les différents groupes.

Dans la suite, je vais décrire ces groupes de réseaux, en m'appuyant sur le tableau 5.3, qui décrit les valeurs moyennes des indicateurs classiques des réseaux pour chaque groupe, et sur une visualisation du réseau le plus proche de son centroïde, c'est-à-dire le vecteur ayant pour coordonnées les moyennes de celles de tous les membres du groupe (les points noirs de la figure 1.20). Les réseaux seront présentés avec la taille

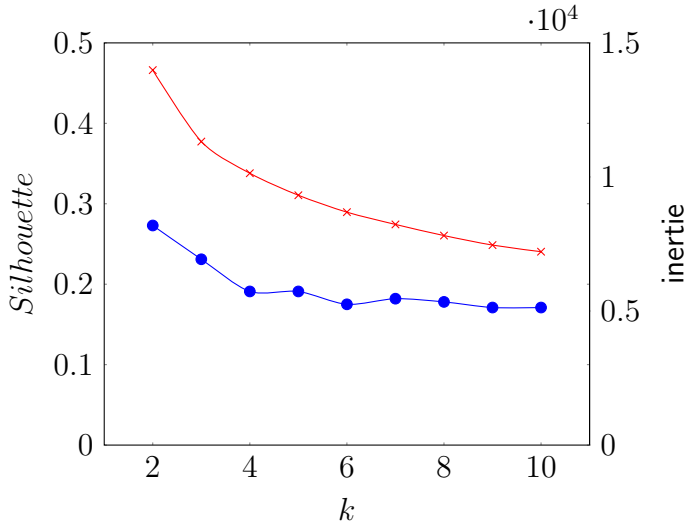
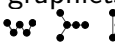


FIGURE 5.13 – La silhouette (en bleu, marquée par des cercle) et l'inertie (en rouge, marquée par des croix) des clusterings de réseaux, selon leur nombre de groupes.

des sommets dépendante de leur centralité d'intermédiarité et la couleur, de leur degré. Les radars qui les accompagneront montrent, en couleur la représentativité du cluster, et en noir celle du réseau-centroïde.

Cluster densifié - 493 (13 % du corpus)

Ce cluster, en bleu sur le diagramme, regroupe des réseaux à forte représentativité des graphlets densifiés . Certains des graphlets troués, ceux aux plus fortes densités , sont également sur-représentés. Un futur travail à réaliser consiste à se pencher sur l'hypothèse que ces graphlets troués sont induits par les sommets périphériques des communautés denses, en étudiant l'énumération de leurs positions. Pour terminer, ils ont une faible représentativité des autres graphlets troués, ainsi que des graphlets triangulés et même une très faible représentativité du groupe des chemins  et des étoilés .

Ces réseaux personnels sont donc les plus denses du corpus et ont également la plus forte transitivité, ce qui n'est pas une surprise mais qui amène déjà l'idée que les valeurs des indicateurs classiques calculés sur un graphe pourraient être corrélées à celles de ses graphlets sur-représentés et anti-corrélées à celles de ceux sous-représentés. En effet, à l'inverse des graphlets étoilés et des chemins, les réseaux densifiés ont une faible centralisation d'intermédiarité (voir figure 5.2). Encore à l'inverse des chemins, ils ont un faible diamètre et une faible modularité.

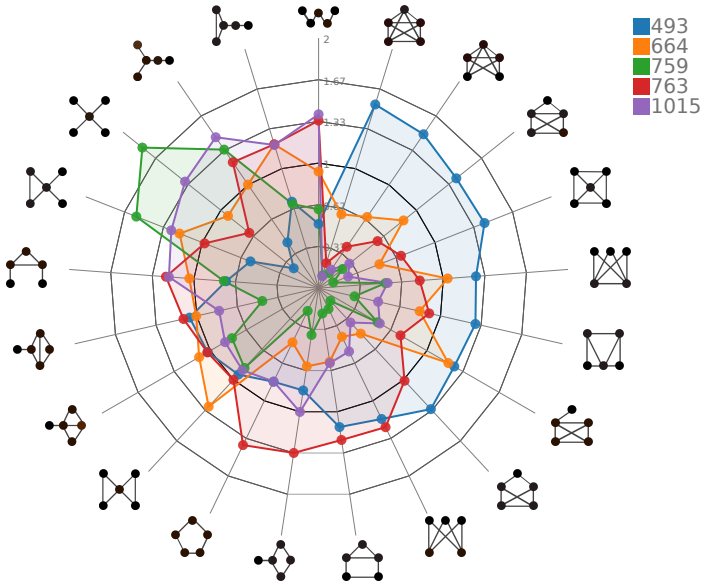


FIGURE 5.14 – Les cinq clusters obtenus avec la représentativité des graphlets de taille 5.

.

Ces réseaux ont donc rarement un sommet extrêmement central, qui pourrait faire le lien entre des communautés, qui sont de toute façon généralement mal définies dans ce cas. Le faible nombre de sommets des réseaux densifiés est commun pour les réseaux denses puisque moins on a d'alters et plus un nouvel ami a de chance d'être présenté à une fraction importante d'entre eux.

Cluster nœud-papillon - 664 (18 % du corpus)

Ce cluster est un peu particulier car il est moins que les autres associé à une famille de graphlets, et n'est présent dans aucune autre mesure testée plus haut. Son nom lui vient de sa sur-représentation notable de ces graphlets , qui, pour tous les autres groupes est proche de 1.

Il a également une sur-représentation de la clique-plus-un dont la forte représentativité globale, décrite en figure 5.6, indique qu'elle est un marqueur efficace des petites communautés très denses ou bien des communautés peu denses, très présentes dans le corpus. On imagine alors que cela doit également être le cas, à moindre mesure peut-être, pour les autres graphlets densifiés moins sous-représentés que les autres .

Parmi les graphlets des autres familles, ce sont à chaque fois les plus denses qui

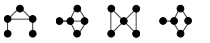
	denses	nœuds	étoilés	troués	fourchus
$\|V\|$	84	81	89	98	94
$\|E\|$	625	365	291	530	327
densité	0.18	0.12	0.08	0.11	0.07
transitivité	0.7	0.64	0.46	0.52	0.47
diamètre	5.63	6.26	6.3	6.47	7.06
centralisation d'intermédiarité	0.15	0.2	0.35	0.17	0.23
nombre de sommets dans la composante connexe principale	58.1	58.2	70.8	75.5	74.1
nombre de communautés de Louvain de taille > 5	4.01	4.4	5.09	4.74	5.16
modularité	0.39	0.52	0.56	0.46	0.58
taille des communautés de 5 sommets ou plus	17.9	15.2	18.1	14.4	15.1

TABLE 5.3 – Quelques variables classiques descriptives des clusters

sont sur-représentés. Ces réseaux ont quelques communautés d'alters, peu importantes en taille mais relativement denses, ce que confirme le tableau des indicateurs. En effet, ce sont, après les réseaux densifiés, ceux avec la plus importante densité, même si elle demeure nettement plus faible, et surtout à la plus haute transitivité, la densité locale. Ces communautés sont probablement reliées entre eux par plusieurs alters, qui apparaissent au centre de nœuds-papillon et sûrement dans les positions centrales d'autres graphlets .

Le cluster a une forte sous-représentation en graphlets troués, ce qui va encore dans le sens de l'idée que ceux-ci sont généralement induits dans les larges communautés d'alters, dont la taille est plutôt faible pour les réseaux de ce groupe.

Cluster étoilé - 759 (20 % du corpus)

Ce cluster a une sur-représentation des deux graphlets en étoile et à moindre mesure de la fourche dont on a déjà vu en section 5.6.2 qu'elle était assez proche de ceux-ci. Les autres graphlets sont à l'inverse sous-représentés. Pour chaque famille de graphlets, ceux avec un sommet nettement plus central sont moins sous-représentés que les autres  ainsi qu'un avec deux sommets centraux . Que ce soit pour les deux sommets centraux de l'opéra ou celui de la clique-plus-un on voit donc que les sommets des graphlets densifiés ne sont pas tous « coincés » au

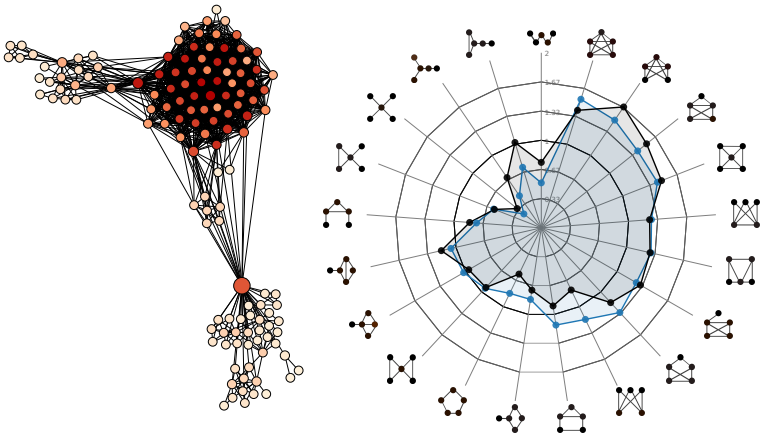


FIGURE 5.15 – **centroïde densifié**
Le réseau le plus proche du centre du groupe densifié possède une très importante et très dense communauté principale d'alters, ce qui explique en grande partie sa sur-représentation des graphlets densifiés. Il a une seconde communauté d'amis, moins dense, en bas de la figure, et elles sont reliées par un alter avec beaucoup de connexions dans chacune, en particulier celle du bas. Cet alter ne fait cependant pas croître la représentativité des motifs étoilés, mais plutôt celle de la fausse étoile , puisqu'il ne connaît des gens que de ces deux groupes à fortes transitivités. Finalement, un dernier groupe d'alters, en haut à gauche est relié à quelques membres de la communauté principale. La présence de trois communautés fait légèrement augmenter le nombre de chemins , et surtout de cintres qui ont la même représentativité que l'ensemble du corpus. Par ailleurs, la très forte densité, ici, de la communauté principale fait diminuer la représentativité des graphlets troués, puisqu'elle contient donc peu de sommets périphériques et probablement beaucoup de cordes.

centre de communautés denses.

Ces réseaux ont tous au moins un sommet très central, un *alter-ego*, connecté à la majorité, sinon toutes les sphères sociales d'*ego*, qui sont, selon Louvain, les plus grandes du corpus. Le cluster a, de manière attendue, la plus haute centralisation d'intermédiarité, ce qui induit une faible transitivité puisque les triangles contenant le sommet central, ou éventuellement un des sommets centraux, ne sont généralement pas fermés. L'hypothèse la plus probable est que l'un de ces alters soit le conjoint de

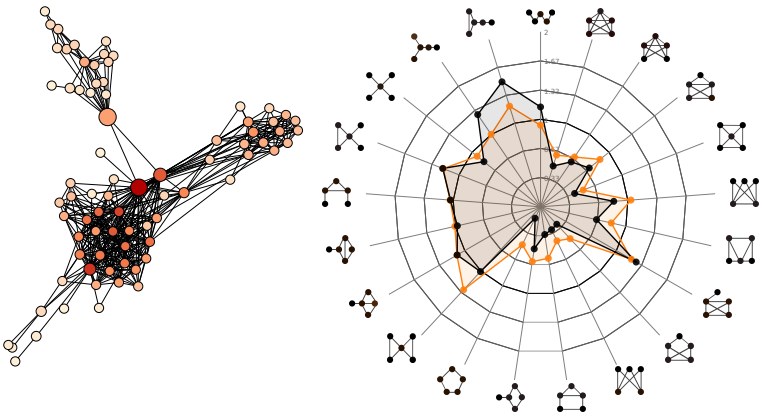


FIGURE 5.16 – **centroïde nœud-papillon**
Ce centroïde est composé de trois communautés qui sont reliées par trois sommets centraux, chacun plutôt issu d'une d'entre elles. Deux de ces communautés sont assez denses et fournissent les cliques-plus-un , dont le sommet périphérique est alors généralement situé dans une des autres, et où le sommet faisant le lien étant alors un de ces trois alters centraux. La troisième communauté, en haut à gauche, est plus aérée. On peut penser que l'un des deux triangles de la fausse étoile , qui est cependant moins sur-représentée ici que dans le cluster, est généralement le triangle constitué par ces trois sommets. Les deux sommets restants variant alors entre les alters de la communauté du sommet qui se retrouve au centre du nœud.

l'enquêté et que ses meilleurs amis de chaque communauté soient parmi ceux reliés à cet alter, et donc en périphérie des étoiles. En dehors du ou des quelques alters centraux, les sphères sociales de ces enquêtés sont en général plutôt déconnectées les unes des autres, comme l'indiquent la modularité forte et la densité faible.

Une question intéressante avec ces réseaux est de se demander si le sommet central de la clique-plus-un est central dans le réseau ou bien si c'est paradoxalement son sommet périphérique qui n'apparaît pas plus souvent dans la position centrale des étoilés , auxquels cas les sommets centraux des cliques-plus-un seraient les leaders de groupe.

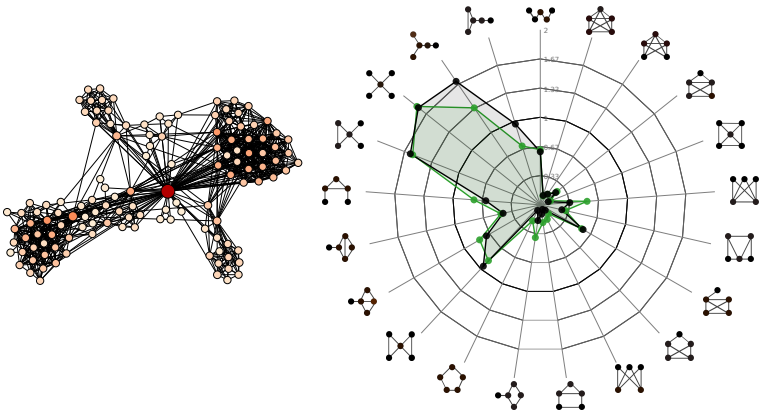


FIGURE 5.17 – **centroïde étoilé**
On remarque tout de suite le sommet très central de ce réseau, qu'on suspecte de représenté l'alter-amour, et qui est au centre de la majorité des étoiles et fausses étoiles. Il a beaucoup de liens avec les deux groupes principaux ainsi qu'avec un troisième et a quelques liens avec le dernier groupe, également plus petit, en bas à droite. Quelques alters de différentes communautés sont également connectés entre eux et sont très probablement les relations les plus fortes de l'enquêté dans chaque communauté. Ces sommets apparaissent certainement au centre étoiles à trois branches et donc de fausses étoiles , de fourches , ainsi que de cintres , qui sont ici plus sur-représentés que dans le groupe, peut-être à cause de la faible densité apparente de la communauté de gauche qui favoriserait alors l'induction de chemins.

Cluster troué - 763 (21 % du corpus)

la majorité des graphlets troués , et à moindre mesure l'enveloppe ouverte sont tous sur-représentés dans ce groupe. Les graphes de cette famille contiennent évidemment toujours assez peu de graphlets troués, mais donc nettement plus que les autres. Ils contiennent donc des chaînes fermées d'amis d'égo qui s'entre-connaissent de manière assez singulière : il n'y a pas d'amis plus centraux que d'autres dans les cycles, et les triangles ne sont jamais fermés. C'est d'ailleurs le 5-cycle, archétypal de la famille, qui est le graphlet le plus sur-représenté dans le groupe.

Les graphlets chemins ainsi que d'autres induisant des 4-chemins sont sur-représentées dans ce groupe ou alors moins sous-représentés que

les autres graphlets de la même famille . Par ailleurs, les graphlets denses de chaque famille sont ceux qui en sont généralement les moins représentés alors que les graphlets densifiés eux-mêmes sont moins sous-représentés dans cette famille que dans les graphes étoilés ou chemins.

Le groupe est celui dont les membres ont le plus de sommets, réparties dans des communautés assez difficilement discernables induisant une modularité relativement faible. Ces communautés sont généralement reliées entre elles par quelques connexions sans qu'un alter central émerge. Une hypothèse liée à ces réseaux sont qu'ils appartiennent à des individus dont les sphères sociales sont assez homogènes, ce qui explique que certains de leurs membres s'entre-connaissent, sans que cela ne soit forcément du fait d'*ego*.

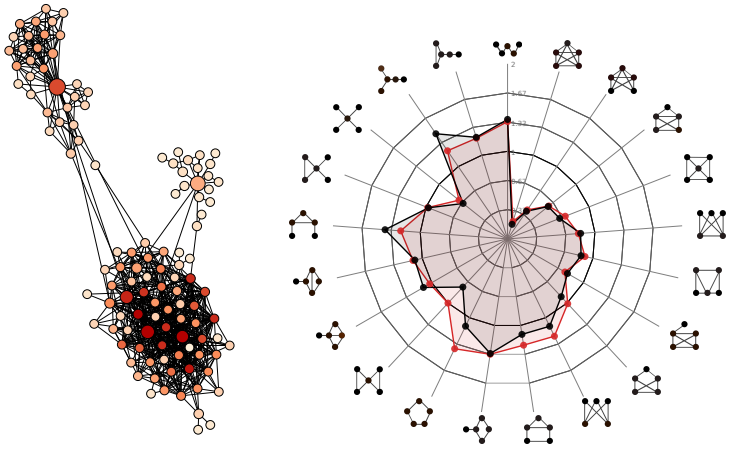


FIGURE 5.18 – **Centroïde troué**
Le réseau est composé de trois groupes qui sont assez peu denses. Le principal a une répartition équilibrée des centralités et degrés, et les deux plus petits contiennent chacun un sommet localement central. Plusieurs facteurs expliquent la sur-représentation des graphlets troués dans ce réseau. D'un côté il a une grande communauté relativement dense mais qui contient beaucoup de sommets périphériques qui sont peu connectés entre eux. De l'autre, plusieurs connexions existent entre ses communautés avec plusieurs alters qui sont à l'interface entre celles-ci.

Cluster fourchu - 1015 (28 % du corpus)

Ce dernier groupe partage quelques similitudes avec le précédent, avec une sur-représentation des graphlets chemins , peu de graphlets triangulés, et

âge	densifiés	nœuds	étoilés	troués	fourchus	tous
18-25	0.42	0.26	0.11	0.42	0.26	0.28
26-40	0.32	0.45	0.56	0.31	0.46	0.43
41-60	0.22	0.27	0.3	0.24	0.25	0.26
60+	0.03	0.02	0.03	0.03	0.03	0.03
pop.	322	401	487	494	614	2318

TABLE 5.4 – Répartition des classes d'âge au sein des familles de réseaux personnels. Hors erreur d'arrondis, les colonnes se somment à 1.

encore moins les plus denses d'entre eux, et, de fait, peu de graphlets densifiés. Il a par contre plus de motifs étoilés indiquant que ses réseaux ont des sommets faisant le pont entre plusieurs communautés d'alters. Ces graphes contiennent également peu de graphlets troués, à l'exception du cycle-plus-un légèrement sur-représenté.

Le groupe a le diamètre moyen le plus long du corpus, ce qui montre que ces réseaux sont traversés par des chemins naviguant à travers leurs différentes communautés d'alters. Ces dernières, nombreuses, sont les plus dessinées du corpus mais sont plutôt petites, malgré le fait qu'elles soient généralement assez peu denses.

La sur-représentation relativement forte des étoiles à quatre branches démontre l'existence d'alters centraux entre les communautés. Ces sommets centraux sont d'ailleurs probablement inclus au centre des chemins et il est alors imaginable que les quatre branches de l'étoile soient terminées par des alters de la communauté de cet alter central et d'alters d'autres communautés.

5.9 Quels individus pour quels réseaux ?

On peut désormais combiner ces familles de réseaux personnels avec les données socio-démographiques qu'on a récupérées lors de l'enquête Algopol ainsi qu'avec les familles d'usages qu'on a construites, ces deux axes étant présentés dans le chapitre 2.

5.9.1 Croisement avec les données socio-démographiques

Dans un premier temps on va regarder comment les classes d'âge et les types de relations amoureuses que nos enquêtés ont déclarées à Facebook se distribuent dans nos catégories de réseaux. Ces répartitions sont respectivement présentées en figure 5.4 et figure 5.5.

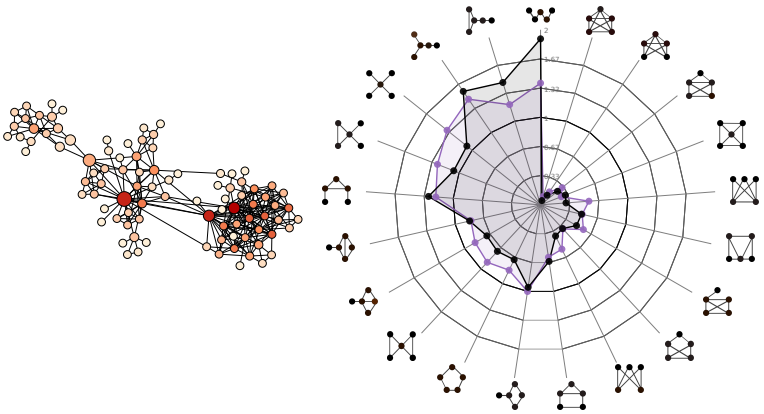


FIGURE 5.19 – **centroïde fourchu**
Le réseau est composé de groupes hétérogènes : un dense, un dispersé autour d'un alter central et un moyennement dense. Bien que les communautés soient assez séparées les unes des autres, elles sont tout de même connectées par quelques liens. Certaines connexions apparaissant d'ailleurs entre des alters qui sont peu centraux au sein de leur propre communauté. On peut alors imaginer des rencontres fortuites ou en tous cas peu structurantes. La faible densité de la communauté centrale participe à la très forte représentativité en chemins du réseau puisqu'elle induit le fait qu'il y ait beaucoup de manières de la traverser. À l'inverse, l'absence de connexions entre les alters centraux de chaque groupe fait baisser la représentativité des étoiles et des autres graphlets contenant un sommet central

On retrouve encore un résultat classique déjà discuté dans le chapitre 3 : les jeunes ont des réseaux plus denses, avec pas moins de 42% des réseaux densifiés qui appartiennent à des moins de 25 ans, tandis qu'ils sont sous-représentés dans les réseaux fourchus et étoilés, les moins denses du corpus.

Comme cela a été mentionné dans la description du cluster troué, également composé à 42% de jeunes adultes, on fait l'hypothèse qu'une haute représentativité des graphlets à trou peut souvent être expliquée par l'homogénéité des différentes communautés qui composent ces réseaux et par l'homophilie entre leurs alters. Ces derniers appartiendraient alors majoritairement aux mêmes catégories d'âges ou partageraient des origines sociales similaires. Ce cas de figure semble être plus courant dans le cas où *ego* est jeune et n'a pas encore diversifié les sphères sociales qu'il fréquente.

Un autre résultat attendu est la proéminence des personnes mariées dans le cluster

relation Facebook	densifiés	nœuds	étoilés	troués	fourchus	tous
célibataire	0.42	0.31	0.14	0.4	0.27	0.29
en couple	0.27	0.32	0.36	0.34	0.39	0.35
marié	0.31	0.37	0.5	0.26	0.34	0.36
pop.	197	278	414	328	450	1667

TABLE 5.5 – Répartition des différentes relations amoureuses, telles que déclarées à Facebook, au sein des familles de réseaux personnels. Hors erreur d'arrondis, les colonnes se somment à 1.

étoilé qui semble confirmer l'hypothèse selon laquelle l'alter au centre de l'étoile est généralement l'alter-amour. On peut noter la sur-représentation assez forte des 26-40 ans parmi les étoilés qui correspond peut-être au fait que parmi les populations dont l'âge les rend plus susceptibles d'être mariés (au delà de 25 ans), les moins de 40 ans sont certainement ceux ayant la plus forte présence sur la Facebook. À partir de là, on imagine que la présence des deux membres du couple sur la plateforme augmente les chances pour que les deux réseaux adoptent cette forme en étoile. Cette classe d'âge est par ailleurs sous-représentée dans les réseaux troués, qu'on suppose à forte homophilie, ce qui va dans le sens des hypothèses émises en section 3.2.2 comme quoi cette période correspond souvent à un pic de sociabilité.

Une dernière observation vient du fait que nous nous sommes basés sur les descriptions de statuts amoureux donnés à Facebook, cette information n'étant pas demandée dans le questionnaire Algopol. Par ailleurs certaines personnes n'ont pas répondu à toutes les questions de l'enquête et c'est pourquoi nous n'avons pas l'âge de toutes. Il est intéressant de noter que le ratio entre le nombre de personnes qualifiant leur relation à Facebook et celui du nombre de répondants à notre question concernant leur âge varie significativement selon les clusters de réseaux. De 0.72 sur l'ensemble du corpus, il passe à 0.61 pour les densifiés et à 0.66 pour les troués, ce qui suggère que les jeunes, plus présents dans ces deux groupes, ont plus de réticences à déclarer leur situation amoureuse en ligne. À l'inverse, le cluster étoilé possède, avec 0.85, le plus haut ratio du corpus, ce qui montre que les personnes mariées sont plus susceptibles de le préciser que les célibataires.

5.9.2 Graphlets et usages

La table 5.6 décrit les répartitions des enquêtés selon les deux entrées que sont leur usage et la famille de leur réseau personnel. Le tableau indique également les scores de χ^2 de chaque case, qui augmente lorsque la valeur s'éloigne de celle attendue, positive lorsque l'effectif est supérieur à celui attendu, et négative dans le cas contraire. On

	Densifiés		Noeuds-pap.		Étoilés		Troués		Fourchus	
Non actifs	63	(-0.93)	99	(0.05)	148	(2.48)	121	(0.36)	139	(-1.86)
Conv. Groupes	29	(0.86)	35	(0.1)	32	(-1.57)	50	(1.43)	53	(-0.51)
Conv. Distribués	53	(0.98)	61	(-0.43)	61	(-2.03)	102	(2.9)	96	(-1.02)
Egocentrés	175	(1.18)	258	(2.37)	270	(-0.18)	218	(-2.86)	367	(-0.04)
Egovisibles	60	(-0.58)	73	(-1.78)	124	(1.31)	98	(-0.85)	165	(1.36)
Partageurs	37	(-1.88)	54	(-1.91)	76	(-1.06)	100	(1.85)	138	(2.08)

TABLE 5.6 – Répartition des populations entre catégories d'usages et familles de réseaux et entre parenthèses : résidus de χ^2par rapport aux effectifs attendus.

voit ainsi émerger quelques tendances à la lecture de ces résultats.

La sur-représentation des étoilés au sein du groupe des non-actifs peut être le marqueur d'une baisse de la fréquentation de Facebook chez les gens mariés, qui comme on l'a vu, forment une partie importante des étoilés. Le fait d'être installé en couple, et peut-être des enfants à charge limiteraient alors probablement le temps disponible pour communiquer sur la plateforme.

Beaucoup de conversants distribués ont des réseaux troués. Cela n'est pas étonnant puisqu'on a vu que ces deux populations rassemblaient une grande parties des jeunes de notre corpus, et il semble que cette concentration soit plus importante qu'un éventuel lien de cause à effet pour justifier cette sur-représentation. À l'inverse, ils sont naturellement moins représentés parmi les étoilés, qui sont plus âgés et généralement installés en couple.

Les égocentrés, qui représentent l'usage a priori le plus commun de Facebook, ont plus souvent des réseaux de la famille des nœuds-papillon que des troués. Les partageurs ont quant à eux une répartition inattendue en familles de réseaux, ce qui confirme le caractère assez particulier de cet usage de Facebook.

Une fois de plus, les partageurs ont un distribution particulière. Peu présents au sein des étoilés et à plus forte raison encore parmi les nœuds-papillons et les densifiés, ils sont à l'inverse nombreux à avoir des réseaux fourchus ou troués.

Dans ce chapitre, j'ai présenté la méthode d'analyse des réseaux par les graphlets, la

représentativité des graphlets, mise au point durant mon doctorat.

En utilisant cette méthode, on a commencé par voir le comportement des graphlets les uns par rapport aux autres, dans les réseaux personnels de Facebook, mettant en évidence des relations structurelles liées à leurs propres graphlets induits. Probablement robuste de ce fait, la typologie proposée présente néanmoins très certainement des spécificités propres au type de réseaux analysés et une étape suivante serait de faire le même travail à partir de réseaux d'origines variées.

Après avoir vu quelques regroupements produits par des métriques de la littérature, j'ai analysé celui produit par la représentativité des graphlets, exhibant cinq familles de réseaux qui constituent des ensembles très identifiables les uns par rapport aux autres. J'ai ensuite confronté ces familles de réseaux aux données socio-démographiques et d'usages abordées dans la première partie de cette thèse, suggérant quelques liens possibles.

Si ces liens existent, il apparaît cependant que la forme qu'adopte le réseau social de chaque usager de Facebook ne peut probablement pas être directement déduite de ces quelques indicateurs. Une étape suivante de cette recherche consisterait alors à analyser les différentes catégories socio-professionnelles de ces 30 sous-familles, obtenues par le croisement des deux typologies de la thèse.

Chapitre 6

Approfondissements

Dans ce chapitre, je présente quelques études en cours visant à approfondir les résultats précédents, soit par un questionnement sur la taille des graphlets étudiés, soit au travers des essais de constructions différentes des groupes des clusterings du chapitre précédent.

6.1 Taille des graphlets

Dans le chapitre précédent, on a vu que la représentativité des graphlets permet de regrouper des réseaux selon les différences et les similitudes de leurs formes. Au regard du temps de calcul important de cet indicateur, on peut cependant s'interroger sur l'intérêt de pousser l'algorithme jusqu'à énumérer tous les graphlets de taille 5. Dans cette section, on va voir des résultats obtenus en limitant l'analyse aux six graphlets de taille 4, et on va également mettre en lumière les intérêts complémentaires des deux options.

6.1.1 Intuition : des groupes structuraux de graphlets

Ce sont les familles de graphlets qui suggèrent la possibilité de se limiter aux graphlets de taille 4 dans l'étude des réseaux sociaux. On a en effet vu que celles-ci sont en partie déduites des sous-graphlets de taille 4 inclus dans chaque groupe, comme le montre la figure 5.2. On peut donc imaginer que ceux-ci permettent également de caractériser les réseaux de notre panel.

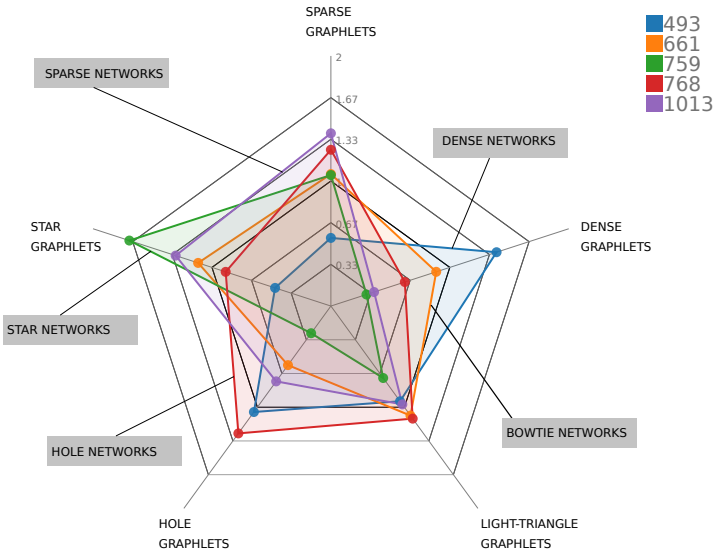


FIGURE 6.1 – La représentativité des familles de graphlets pour chaque famille de réseaux

Par ailleurs, on a aussi vu que la majorité des familles de réseaux pouvaient ainsi être associées, dans une certaine mesure, à une, parfois deux, familles de graphlets. La figure 6.1 montre que les groupes densifiés et étoilés sont associés aux familles de graphlets éponymes et ont des sous-représentations des graphlets des autres familles. Les familles des troués et des chemins ont chacune plusieurs types de graphlets sur-représentés et quelques autres nettement sous-représentées. Seule la famille des nœuds papillon est plutôt équilibrée, n'ayant qu'une sous-représentation remarquable des graphlets troués.

À l'exception de cette famille des réseaux en nœuds papillon qui est beaucoup moins précisément décrite qu'avec les 21 graphlets, toutes les autres peuvent être appréhendées, de manière certes succincte, selon ce diagramme et on peut donc imaginer que les représentativités des six graphlets de taille 4 peuvent suffire pour retrouver une partie importante des interprétations faites pour la taille 5 des graphlets, peut-être même permettre pour le cluster nœuds papillon.

6.1.2 Clustering depuis la taille 4

On a donc procédé au calcul des représentativité des graphlets de taille 4. Comme le calcul est nettement plus rapide que dans le cas de la taille 5, il a été possible de faire des regroupements avec beaucoup plus de réseaux, et notamment avec des réseaux plus grands, ayant jusqu'à 250 sommets. La même méthode de regroupement des réseaux a été utilisée : un kMeans répété cent fois, en ne conservant finalement que celui avec

le plus haut score de silhouette. La figure 6.2 montre les clusters obtenus de cette manière. Les couleurs des clusters précédents ont été conservées.

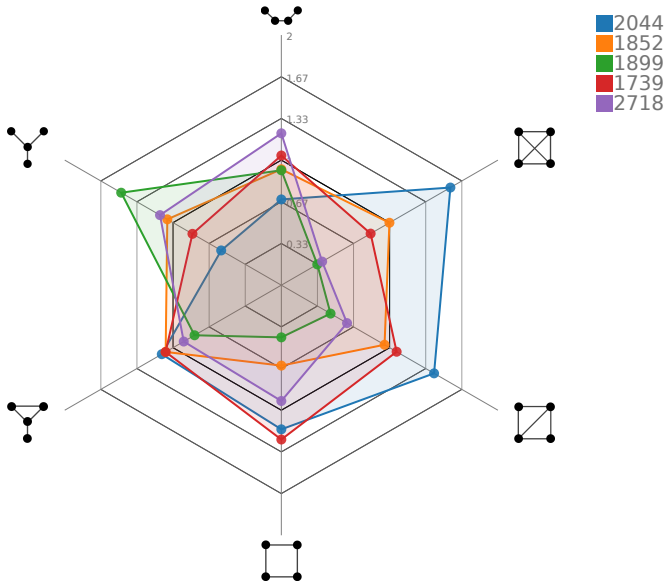


FIGURE 6.2 – Les clusters obtenus par la méthode des kMeans appliquée aux valeurs de représentativités des graphlets de taille 4.

On retrouve, avec 2 044 réseaux, soit nettement plus que précédemment, le cluster densifié. Il a toujours une très forte sur-représentation des graphlets les plus denses et une sous-représentation des deux graphlets les moins denses . De manière un peu surprenante, le groupe a également une sur-représentation en carrés , ce que suggère néanmoins déjà la figure 6.1, dans laquelle le cluster densifié est également sur-représenté en graphlets troués. À ce niveau, les graphlets de taille 5 apportent une précision puisque dans le groupe ayant des motifs denses, seuls certains graphlets troués sont également sur-représentés, tandis que les moins denses d'entre eux y demeurent sous-représentés.

Le cluster des nœuds papillon, avec 1 852 réseaux, demeure le moins aisé à interpréter. Il a en effet une représentation normale, proche de 1, de tous les graphlets à l'exception du carré , sous-représenté. On peut néanmoins noter qu'il a la deuxième plus importante représentativité des cliques , ce qui suggère qu'il ait des communautés de sommets assez denses, ou de nombreuses communautés moyennement denses, malgré sa légère sous-représentativité des diamants . Ici, le passage de la taille 4 à la taille 5 apporte des précisions importantes puisqu'il permet de voir que certains graphlets particuliers sont importants pour ce groupe de réseaux.

Le cluster étoilé (1 899 réseaux) a les mêmes caractéristiques de représentativité qu'à la taille 5. Il a beaucoup d'étoiles et très peu de trous et de graphlets denses . Il a ici une légère sur-représentation des chemins qui sont induits deux fois de la fourche , elle-même sur-représentée pour la taille 5, et qui contrebalance ainsi la sous-représentation des deux autres graphlets chemins .

Le groupe des réseaux troués est nettement moins représenté que dans le cas précédent avec maintenant 1 739 membres. Il est beaucoup moins sous-représenté en graphlets denses que pour la taille 5 et a même une sur-représentation des diamants difficilement discernable au niveau 5, mis à part pour sa sur-représentativité de certains graphlets desquels il est induit . Le groupe est surtout moins dominant que précédemment dans le cas du seul graphlet à trou ayant 4 sommets et bien qu'il y soit toujours le plus sur-représenté, le groupe densifié a presque la même représentativité.

Finalement, le cluster des chemins, qui demeure le plus fourni avec ses 2 718 réseaux, conserve les propriétés remarquables qu'on lui avait trouvées avec les graphlets de taille 5. Il est le plus sur-représenté en chemins traversant le réseau via les sommets centraux d'étoiles et de fausses étoiles légèrement sur-représentées également. Il est également toujours très peu densifié .

6.1.3 Similitude entre les clusterings

Il convient également de vérifier si ce clustering est semblable ou non au précédent, réalisé avec les 21 graphlets de taille 5. Il faut également voir si le fait d'ajouter des réseaux plus grands modifie les regroupements de manière importante. Puisqu'ils ne sont pas réalisés sur les mêmes ensembles de réseaux (beaucoup plus nombreux dans ce dernier cas) nous allons en effet devoir nous restreindre à la comparaison de la répartition des réseaux inclus dans les deux cas, en adoptant plusieurs astuces pour tenter d'intuiter le comportement de ces clustering.

Pour comparer deux regroupements, on utilise l'indice de Rand qui leur donne un score de similitude. Il est calculé à partir du nombre m de couples d'éléments qui sont dans le même groupe dans les deux clusterings et du nombre s de couples qui sont dans deux groupes séparés dans chaque cas. Si N est le nombre total de couples d'éléments qu'on peut construire, alors l'indice de Rand est donné par :

$$R = \frac{m+s}{N}.$$

Il vaut 1 quand les deux regroupements sont exactement les mêmes et décroît jusqu'à 0 avec l'augmentation de la différence entre les deux. La figure 6.3 décrit l'évolution

des indices de Rand des clusterings pour les graphlets de taille 4 et ceux de taille 5 en fonction de la taille maximale des réseaux étudiés dans trois situations :

— **a)** En ne prenant que les réseaux qu'on a classés selon leurs représentativités de graphlets de taille 5, on va effectuer la même opération avec les représentativités de taille 4. On va ainsi pouvoir comparer les regroupements de ces réseaux entre les tailles 4 et 5.
— **b)** En prenant ces mêmes réseaux mais en ne conservant que les plus petits d'entre eux, on peut effectuer plusieurs clusterings en ajoutant à chaque fois les réseaux plus grands, par exemple par tranche de 20, selon leur nombre de sommets. Ainsi on procédera à un clustering des réseaux de taille 0 à 20, puis 0 à 40, 0 à 60, etc. On espère ainsi voir si le fait d'ajouter des réseaux plus grands rend le découpage moins précis
— **c)** Comme ci-dessus, on peut faire une opération similaire mais en ne regroupant que les réseaux ayant plus ou moins la même taille, on commence ainsi cette fois par les réseaux de taille 0 à 20, puis 20 à 40, 40 à 60, etc.

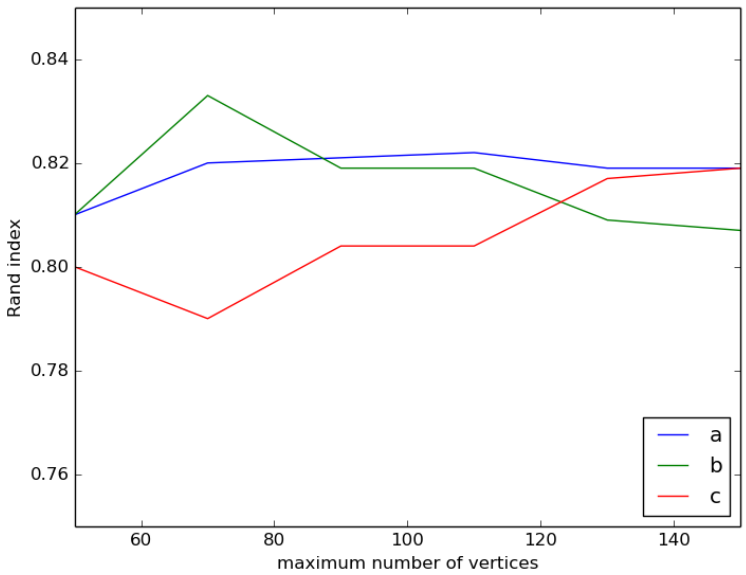


FIGURE 6.3 – Les indices de Rand des clustering de taille 4 et 5 selon les cas **a)**, **b)** et **c)**

La valeur de l'indice de Rand, autour de 0.8 est généralement légèrement supérieure est plutôt encourageante, surtout que ces trois courbes ne semblent pas connaître une diminution trop importante de leurs valeurs lorsque la taille des réseaux augmente. Il est cependant trop tôt pour apporter des conclusions très précises concernant l'efficacité

d'un clustering restreint à la taille 4, par rapport à ceux qui utilisent les graphlets de taille 5.

6.1.4 Des clusterings aux intérêts complémentaires

Le regroupement des réseaux selon leurs représentativités des graphlets de taille 4 offre en tous cas moins d'intérêt que celui de taille 5 lorsqu'il s'agit d'interpréter son résultat, moins parlant. Il y manque en effet les graphlets emblématiques des familles, qui n'ont ici le plus souvent qu'un unique représentant. Ces graphlets sont en effet très utiles pour guider l'interprétation des clusters. De plus, certaines subtilités de la taille 5, comme les différences de représentation des graphlets troués, ou encore la représentativité de certains graphlets comme la clique-plus-un dans le cluster des nœuds papillon qui permet de mieux le comprendre est ici tout à fait invisible.

Si la taille 5 est la plus à même de permettre une analyse poussée des groupes de réseaux, la répartition qui semble similaire de ceux-ci selon qu'on utilise 21 représentativités ou bien simplement les 6 graphlets à quatre sommets permet néanmoins d'envisager d'utiliser cette méthode afin de dispatcher les grands réseaux pour lesquels il n'est pas raisonnablement possible d'énumérer les sous-graphes induits de taille 5. Ces réseaux seraient alors assignés à un cluster construit à partir des graphlets de taille 5, selon son cluster de taille 4.

6.2 Positions et lien social

Comme cela a déjà été mentionné dans le chapitre 4, certains algorithmes permettent, durant l'énumération des sous-graphes induits d'un réseau, de conjointement calculer les positions dans lesquelles apparaissent ses sommets. C'est-à-dire que pour chaque graphlet énuméré, qui correspond donc au graphe formé par, disons, 5 sommets, pour chacun d'entre eux est notée la position qu'il occupe dans le graphlet.

Une intuition forte existe, selon laquelle ces positions sont très efficaces pour déterminer la nature du lien qui existe, dans le cas des réseaux personnels, entre l'enquêté et un de ses alter. Dans le cas des réseaux appartenant à la famille des étoilés, on se doute que l'alter-amour, s'il existe, est celui qui sera dans la position centrale du plus d'étoiles . Au-delà de cet exemple légèrement trivial, il doit être possible d'envisager une typologie structurelle fine de la position des alters dans un réseau qui soit plus complexe qu'un score de centralité, mais également plus simple à lire qu'une collection de scores de différentes centralités.

Avec l'aide de Nicolas Rosset, un élève-ingénieur de Télécom ParisTech qui réalisait un projet dans le laboratoire SES, nous avons entrepris de premiers travaux exploratoires dans ce sens, avec les 11 positions des graphlets de taille 4.

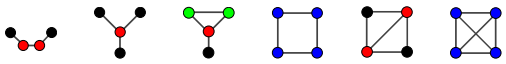


FIGURE 6.4 – Les 11 positions des 6 graphlets de taille 4.

Nous avons poursuivi avec la représentativité comme mesure, mais cette fois des positions. Elle est définie de manière similaire à la représentativité des graphlets. Cette fois la *fréquence globale* d'une position est le nombre de fois où elle a été énumérée dans l'ensemble du corpus et sa *fréquence locale*, pour un alter, est le nombre de fois où, lui, a été trouvé dans cette position. Les détails des calculs sont fournis en section 5.2.

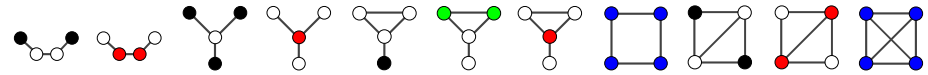
Il est à noter qu'un choix a dû être fait pour le calcul de la fréquence globale des positions. Celle-ci pouvait être faite aussi bien à partir de l'ensemble des alters de tous les réseaux ou bien simplement à partir de l'ensemble des sommets du réseau d'un alter. Nous avons choisi de prendre l'ensemble des réseaux plutôt qu'un seul à chaque fois en ayant en tête le fait qu'un alter pourrait rapidement avoir une sur-représentativité de la position centrale de l'étoile, par exemple, car il est le plus central de son réseau, et non pas car il est réellement en position d'alter-ego, qui connecterait entre elles les différentes communautés du réseau.

Comme pour les graphlets, nous avons construit un regroupement, non pas cette fois de réseaux, mais des alters de ces réseaux. À partir d'un échantillon de 100 réseaux choisis aléatoirement dans notre panel et de la méthode des kmeans, nous avons obtenu les sept groupes d'alters dont les corrélations sont indiquées dans la figure 6.5.

Ce clustering regroupe :

- 5 815 alters très périphériques
- 7 845 alters qui semblent appartenir à des communautés peu denses
- 6 213 alters très centraux dans leur réseau
- 20 118 alters qui sont amis avec ces alters très centraux
- 5 054 alters qui sont dans des communautés très denses
- 10 379 alters qui ressemblent aux 20 118 mais qui ont l'air d'être plus souvent inclus dans des communautés
- 12 409 alters similaires mais qui créent des ponts entre des alters ne se connaissant pas

Cette étude débute à peine et mérite d'être poursuivie, en identifiant beaucoup plus précisément ces groupes d'alters. Puis en poursuivant ce travail avec les 58 positions des graphlets de taille 5, qui permettent une lecture riche des différents rôles qu'on les



1,73	0,25	0,45	0,04	0,51	0,33	0,06	0,08	0,15	0,08	0,14	5815
1,01	0,68	0,8	0,19	0,91	1,28	0,49	0,96	1,3	0,89	1,05	7845
0,75	1,36	0,59	1,29	0,47	0,89	1,29	0,87	0,54	0,91	0,46	6213
1,44	0,3	1,49	0,02	1,34	0,35	0,03	0,16	0,2	0,03	0,03	20118
0,5	0,74	0,3	0,67	0,29	1,13	1,21	0,71	0,87	1,53	1,5	5054
1,34	0,41	1,3	0,04	1,35	0,99	0,11	0,55	1	0,22	0,32	10379
1,38	1,16	1,14	0,38	1,06	0,77	0,41	0,84	0,51	0,29	0,16	12409

FIGURE 6.5 – La table des corrélations aux positions des 7 groupes d'alters. Les nombres à droite de la figure correspondent aux nombres d'alters dans chaque groupe.

alters dans la sociabilité de nos egos afin d'enfin en obtenir une caractérisation précise, dont la figure 6.6, au sein de laquelle la couleur des sommets dépend de leur cluster de position.

6.3 Une méthode de sélection du nombre de clusters du kMeans

Dans cette ultime section, je propose une astuce pour sélectionner le nombre de clusters à obtenir à partir de l'algorithme des kmeans, très utilisée dans ce travail de thèse.

La méthode des kMeans est très efficace pour regrouper des éléments mais son principal défaut réside, on l'a dit, dans l'obligation qu'elle impose de choisir le nombre de groupes du clustering. Le chercheur qui l'utilise doit donc empiriquement décider quel est selon lui celui qui fait le plus de sens. En l'absence d'outil déterministe, ce choix peut s'avérer délicat et il n'est pas toujours aisé de justifier une option plutôt qu'une autre.

La figure 6.7 présente un diagramme de Sankey, c'est-à-dire une visualisation de groupes et de flux entre ces groupes, régulièrement utilisée pour représenter, par exemple, les rapports de voix entre deux tours d'une élection. Ce sont ici les regroupements de nos réseaux par des méthodes des kMeans pour k valant entre 2 et 15, c'est-à-dire pour regrouper nos réseaux en 2, 3, ... ou 15 groupes. Chaque niveau vertical correspond à une valeur de k et chaque rectangle représente un cluster. Les deux rectangles les plus à gauche, en vert et bleu, représentent donc les deux groupes

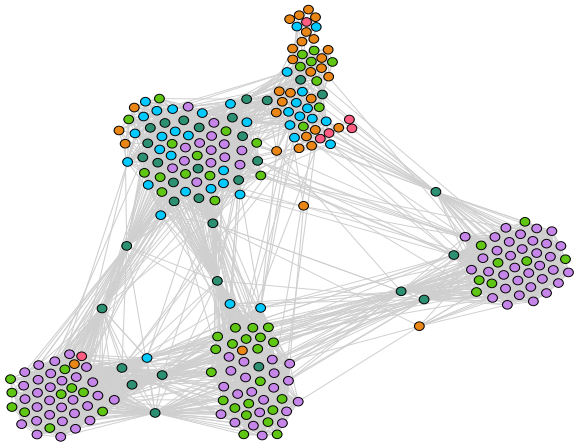


FIGURE 6.6 – Un réseau personnel dont les couleurs des sommets dépendent de leur cluster, selon les représentativités des positions.

obtenus lorsqu'on sépare nos réseaux en deux ensembles, celui du dessus étant plus important, car composé de plus de réseaux. Les flux indiquent la manière dont ces réseaux se séparent ou se rassemblent pour former de nouveaux groupes, lorsqu'on passe à la valeur de k suivante.

On remarque plusieurs comportements aux interfaces entre deux valeurs de k. Certains groupes sont conservés, presque à l'identique. Le phénomène est assez visible pour les groupes placés tout en haut ou tout en bas de la figure. À l'inverse, certains réseaux formant un groupe à un niveau donné se séparent en deux nouveaux groupes, éventuellement avec d'autres pour former de nouveaux clusters.

L'idée proposée est d'observer les évolutions de ces groupes afin de décider quels sont ceux qui vont être conservés dans le clustering final. On va en fait considérer qu'un groupe est *cohérent* lorsqu'il n'évolue pas trop quand on fait varier k. On décide qu'un groupe, issu de la décomposition en k' clusters, est *cohérent* si un autre cluster, pris parmi ceux de la décomposition en $k' + 2$ clusters, a un certain nombre minimal, noté S, qu'on a arbitrairement fixé ici à 75%, de réseaux en commun avec lui.

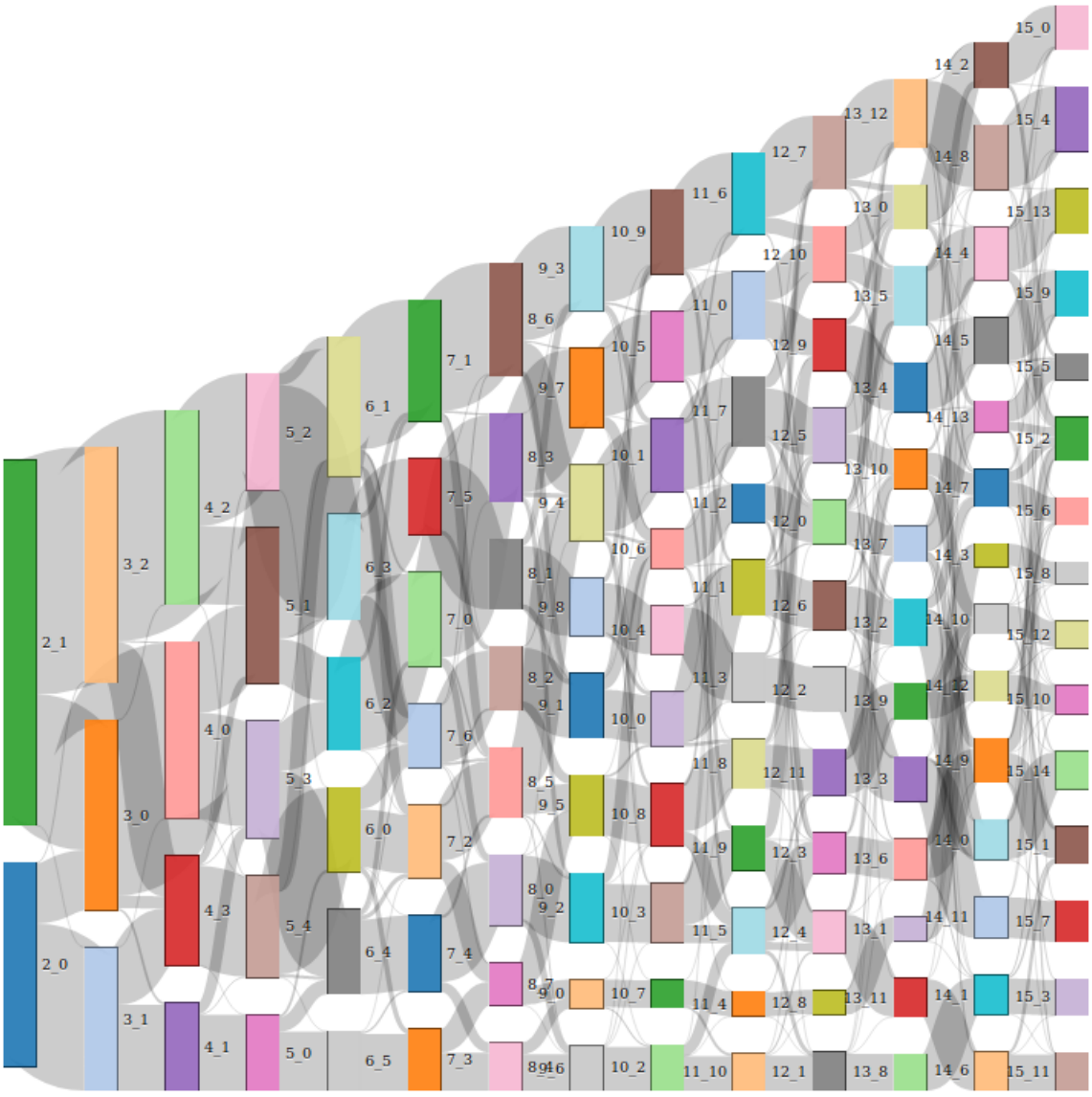


FIGURE 6.7 – Le diagramme de Sankey des clusters de kMeans pour k variant entre 2 et 15.

L'algorithme part d'un dictionnaire de clusterings ClustersParK, dont chaque élément représente les clusters pour une valeur de k et est représenté par une liste de l groupes, pour l variant entre k_{min} et k_{max} les valeurs minimale et maximale de k qu'on considère (2 et 15, dans notre cas). Chacun de ces groupes est représenté par la liste des éléments qui le composent.

On commence par construire deux dictionnaires, PartsSC et PartsCS (S pour source et C pour cible, représentant les clusters sources et cibles des différents flux du diagramme) qui, à un groupe font correspondre respectivement la part de ses éléments dans chaque groupe de taille directement supérieure et directement inférieure. L'algorithme employé pour la construction de ces deux dictionnaires est l'Algorithme 4.

Algorithme 4 : ConstructionParts

Données : ClustersParK, k_{min}, k_{max}
Résultat : PartsSC, PartsCS

```
/* Initialisation */
```

PartsSC $= \{\}$
pour $k' \in [k_{min}; k_{max} - 1]$ **faire**
  pour $c \in$ *ClustersParK*$[k']$ **faire**
    PartsSC$[c] = \{\}$

PartsCS $= \{\}$
pour $k' \in [k_{min} + 1; k_{max}]$ **faire**
  pour $c \in$ *ClustersParK*$[k']$ **faire**
    PartsCS$[c] = \{\}$

```
/* Construction des dictionnaires */
```

pour $k' \in [k_{min}; k_{max} - 1]$ **faire**
  pour $s \in$ *ClustersParK*$[k']$ **faire**
    ElemK$' =$ ClustersParK$[k'][s]$
    pour $c \in$ *ClustersParK*$[k' + 1]$ **faire**
      ElemK$'' =$ ClustersParK$[k' + 1][c]$
      PartsSC$[s][c] = \frac{|\text{ElemK}' \cap \text{ElemK}''|}{|\text{ElemK}'|}$
      PartsCS$[c][s] = \frac{|\text{ElemK}' \cap \text{ElemK}''|}{|\text{ElemK}''|}$

retourner *PartsCS, PartsSC*

L'algorithme 5 permet, une fois ces deux dictionnaires construits, de calculer quels sont les clusters cohérents de ces différents clusterings.

La figure 6.8 présente le diagramme de Sankey qui s'arrête au niveau de ces groupes qu'on qualifie de cohérents. Dans notre cas, la méthode aboutit à 6 groupes et un de ses avantage est que ces groupes ne sont pas forcément sélectionnés depuis le même niveau de clustering. On obtient d'ailleurs dans notre cas deux groupes issus du kMeans à $k = 4$, deux de $k = 5$ et deux de $k = 6$.

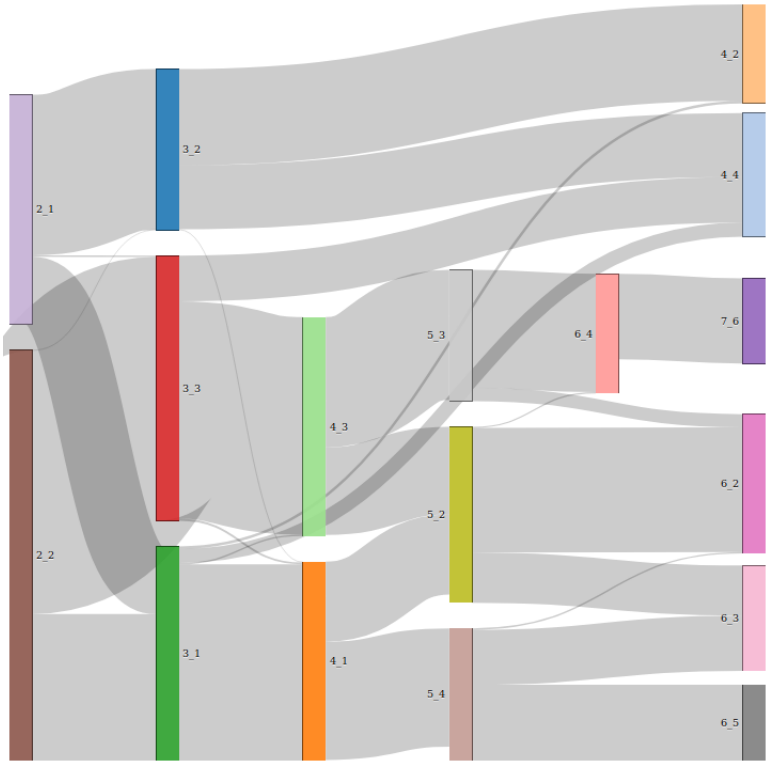


FIGURE 6.8 – Le diagramme des Sankey entre $k = 2$ et les groupes cohérents.

Certains réseaux peuvent n'être inclus dans aucun de ces 6 groupes, on peut dans ce cas soit choisir de les exclure de l'étude, soit trouver un moyen de les réintégrer, en prenant par exemple cluster avec la plus haute valeur de k dans lequel ils apparaissent puis en les incorporant au cluster cohérent ayant le plus de réseaux issus de celui-ci. De manière similaire, il est possible qu'un réseau soit inclus dans deux groupes cohérents conservés puisque ceux-ci sont issus de deux clusterings de différents. À nouveau, solution possible, celle choisie ici, est d'inclure le réseau dans le cluster obtenu avec la plus grande valeur de k qui est a priori plus précis.

La figure 6.9 présente les groupes de réseaux obtenus par cette méthode. Le nombre de 6 groupes est d'ailleurs probablement le maximum que cette visualisation permette d'étudier et en rajouter la rendrait illisible. On retrouve les groupes densifiés, étoilés, en

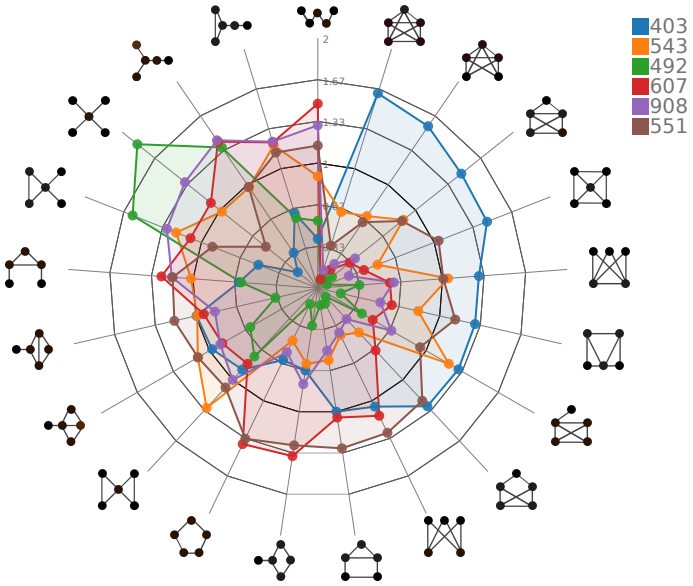


FIGURE 6.9 – Les représentativités en graphlets des 6 clusters cohérents.

nœud papillon et des chemins de la Section 5.7. C'est le groupe des troués qui change le plus et se scinde en deux, un groupe de 607 réseaux et un groupe de 551 éléments, auquel sont également adjoints certains réseaux du groupe des chemins.

Dans ce dernier chapitre, j'ai passé en revue quelques prolongements des méthodes vues dans les chapitres précédents. De niveaux d'avancement variés et de bases bibliographiques elles aussi inégales, il semble néanmoins que toutes offrent des perspectives de recherche aux belles promesses.

Algorithme 5 : TrouverClustersCoherents

Données : ClustersParK, S, k_{min}, k_{max}, PartsSC, PartsCS
Résultat : Liste des clusters cohérents

```
/* Initialisation */
```

ClustersCoherents = []
SourcesPossibles = ClustersParK[k_{min}]
Banis = [] `/* Les groupes dont qui ressemblent au delà du seuil à leur père cohérent */`

```
/* Calculs */
```

pour $s \in SourcesPossibles$ **faire**
- Retirer s de SourcesPossibles
- SEstCoherent = Faux
- **si** $s \in Bannis$ **alors**
  - Continuer
- **si** $s \in$ *ClustersParK*$[k_{max}]$ **alors**
  - Ajouter s à ClustersCoherents
  - Continuer
- **si** $s \in$ *ClustersParK*$[k_{max} - 1]$ **alors**
  - **pour** $c \in$ *PartsSC*$[s]$ **faire**
    - **si** *PartsSC*$[s][c] \geq S$ **et** *PartsCS*$[c][s] \geq S$ **alors**
      - Ajouter s à ClustersCoherents
      - Ajouter c à Bannis
  - Continuer
- **pour** $c \in$ *PartsSC*$[s]$ **faire**
  - **si** *PartsSC*$[s][c] \geq S$ **et** *PartsCS*$[c][s] \geq S$ **alors**
    - **pour** $c' \in$ *PartsSC*$[c]$ **faire**
      - **si** *PartsSC*$[c][c'] \geq S$ **et** *PartsCS*$[c'][c] \geq S$ **alors**
        - Ajouter s à ClustersCoherents
        - Ajouter c et c' à Bannis
        - SEstCoherent = Vrai
        - Sortir de la boucle
    - **si** *SEstCoherent est Vrai* **alors**
      - Sortir de la boucle
    - `/* Si c est similaire à s mais qu'aucun c' n'est similaire a c */`
    - Ajouter c à SourcesPossibles si il n'y est pas déjà
  - **sinon**
    - `/* Si c n'est pas suffisament similaire à s */`
    - **si** *PartsSC*$[s][c] > 1 - S$ **alors**
      - `/* Si la part de s dans c n'est pas trop petite */`
      - Ajouter c à SourcesPossibles si il n'y est pas déjà

retourner *ClustersCoherents*

Conclusion

Pour conclure cette thèse de doctorat, je vais tout d'abord revenir sur les principales contributions apportées par ce manuscrit, dont l'objectif est l'étude des modes de sociabilités en ligne, à partir de données collectées depuis les comptes Facebook de milliers d'utilisateurs dont les habitudes sont analysées dans une perspective dite égocentrée, c'est-à-dire qui prend leur point de vue comme référence. Dans un second et dernier temps, je présente quelques pistes potentielles de recherches futures.

Contributions

Quel que soit le média en question, il est clair qu'une étude des interactions passant par son biais doit tenir compte d'une éventuelle pluralité de ses usages. C'est pourquoi, dans un premier temps, le chapitre 2 présente une typologie des usages de Facebook, agrémentée d'une étude des caractéristiques socio-professionnelles de chaque famille. La description des usages est construite à partir d'un découpage de l'ensemble des actions de nos enquêtés en une collection d'activités élémentaires, permettant de déterminer les habitudes de ces derniers. Il en résulte la détection de 6 catégories d'usage réparties en 3 grandes familles : ceux qui publient majoritairement sur leur propre mur, ceux dont une partie importante de l'activité est située ailleurs que sur leur espace dédié, et les non-actifs.

La première famille, dont les membres publient principalement sur leur mur personnel est répartie entre les égocentrés, les égovisibles et les partageurs. Les égocentrés, plutôt discrets, constituent la figure ordinaire de l'usage adulte de Facebook. Les égovisibles rassemblent les utilisateurs les plus dynamiques, dont l'activité pourrait être rapprochée d'une quête de visibilité personnelle. Finalement, les partageurs forment un groupe au profil singulier et dont l'activité est principalement orientée vers le partage de liens, statuts, photos qu'ils vont piocher sur d'autres pages Facebook que la leur. Deux profils d'usage présentent une diversité de leurs lieux d'activités sur Facebook qui partagent leur très grande proportion de jeunes de moins de 25 ans. Les conversants

distribués passent une bonne partie de leur temps en ligne à écrire et commenter sur les pages de leurs amis, qui n'hésitent pas eux-mêmes à venir discuter sur leur page. Les conversants de groupes concentrent eux une grande part de leur activité sur des pages de groupes, qu'on suppose être des groupes d'interconnaissances plus que d'intérêts communs. Finalement, les non-actifs regroupent des utilisateurs pouvant avoir des profils divers. Ils peuvent aussi bien utiliser des fonctionnalités du site que pour des raisons de protection de la vie privée nous n'avons pas à disposition, comme la messagerie instantanée, ou bien se contenter de regarder les publications de leurs amis. Toujours est-il qu'ils ne publient pas ou alors rarement eux-mêmes.

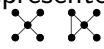
Les interactions entre les enquêtés et les membres de leurs réseaux sont analysées, dans le chapitre 3, à partir d'une batterie d'indicateurs originaux qui qualifient la manière dont les alters commentent et likent les publications de nos enquêtés. Ils permettent par exemple de déterminer si une publication cible une sphère sociale spécifique ou bien est destinée à l'ensemble des amis, si elle génère une discussion animée ou bien si chacun se contente de déposer un unique commentaire. L'ensemble de ces indicateurs sont ensuite agrégés au niveau de chaque ego pour étudier ses habitudes. Il est montré que selon la catégorie d'usage d'un enquêté, Facebook revêt une forme polysémique, véritable plateforme à discussions avec ses amis pour certains ou support à l'auto-publication, à la manière des blogs, pour d'autres. Cette analyse permet également de voir que l'âge d'ego est un facteur déterminant de son utilisation de Facebook, qui varie au cours des étapes de la vie.

Ce chapitre est également l'occasion d'explorer la manière dont le lien social s'exprime en ligne. Une première étude montre que les usages des amis de nos enquêtés, qu'on ne peut pas analyser avec nos données, est néanmoins déterminante dans cette optique. En effet, sont mises en lumière l'absence de corrélation déterminante entre rôle de l'alter dans le réseau égocentré et l'intensité de l'interaction en ligne avec ego, déterminée par le nombre de likes et de commentaires. Une observation du réseau égocentré peut ainsi très bien suggérer qu'un alter particulier a un lien très fort avec l'enquêté, si cet alter est non-actif sur Facebook, l'analyse de leur interaction en ligne dira l'inverse. Finalement, on montre que le succès d'une recherche, dans le réseau, de l'alter représentant le conjoint dépend largement du type de relation, et donc probablement de la durée du couple. Les métriques classiques sont ainsi beaucoup plus efficaces lorsqu'il s'agit de retrouver un alter qualifié comme mari ou femme qu'un alter qualifié comme petit copain ou petite copine.

La thèse explore ensuite les différentes formes de structures relationnelles égocentrées qui se construisent en ligne. Pour cela, elle mobilise l'énumération des sous-graphes induits à ces réseaux personnels, qu'on appelle les graphlets. Après la présentation de quelques éléments historiques de cette mesure dans le chapitre 4, le chapitre 5 présente la représentativité des graphlets. C'est une nouvelle métrique mise au point durant mon

travail de doctorat dont l'intérêt réside dans son efficacité à caractériser, les uns par rapport aux autres, un grand nombre de réseaux qui sont proches entre eux. C'est le cas des réseaux égocentrés du panel Algopol, qui ont tous été collectés de la même manière.

Une catégorisation des 21 graphlets de taille 5 est proposée. Elle montre que les graphlets qui se comportent de manières similaires dans les réseaux personnels ont des caractéristiques structurelles communes, et notamment leurs propres sous-graphes induits de taille 4, ce qui suggère qu'une catégorisation des graphlets d'autres types de réseaux aurait été semblable. Quelques spécificités propres aux réseaux de sociabilité apparaissent cependant, comme le fait que les cycles de taille 5 appartiennent à la même catégorie que les graphlets induisant un cycle de taille 4 . Cette particularité ne serait certainement pas apparue avec des graphes contenant plus de cycles sans corde que les réseaux sociaux, pour lesquels cette sous-structure est rare.

Après avoir catégorisé les graphlets, on applique la représentativité pour catégoriser les réseaux sociaux, en employant la méthode des kMeans, une méthode de clustering. Cette catégorisation fait ressortir 5 groupes de réseaux. Les réseaux densifiés possèdent beaucoup de graphlets denses induits et ont généralement au moins une communauté d'alters très dense. Ces réseaux regroupent un bon nombre de ceux qui appartiennent aux plus jeunes de nos enquêtés, ce qui n'est pas une surprise puisque les jeunes ont souvent des réseaux plus denses, ayant fréquenté moins de sphères de sociabilité. Les réseaux en nœuds papillon sont généralement composés de petites communautés plutôt denses, souvent reliées deux à deux par des alters qui ne sont pas extrêmement centraux par ailleurs. Le groupe des réseaux étoilés ont, dont les egos sont sur-représentés parmi les non-actifs, une très forte sur-représentation des graphlets étoilés  , dans lesquels un alter est très central. Ces réseaux adoptent en quelque sorte la même forme et sont composés de plusieurs communautés peu interconnectées à l'exception d'un sommet qui a beaucoup de liens avec chacune d'entre elles. On a vu que ce groupe d'enquêtés est celui dont les membres se sont le plus déclarés comme en couple ou mariés, et tout indique que c'est cet alter central qui est leur conjoint. Les graphlets troués sont assez spécifiques aux mesures qui comme la représentativité des graphlets pointent des différences de fréquence de graphlets, bien qu'ils soient globalement moins représentés que les autres. Ce groupe, également composé en bonne partie de jeunes répondants, correspond probablement à des enquêtés dont les communautés d'amis sont peu denses et assez proches les unes des autres. Elles sont souvent reliées entre elles par des alters qui ne sont pourtant pas centraux dans le réseau. Ces egos apparaissent souvent dans le groupe des conversants distribués, qui sont également assez jeunes. Finalement, le groupe des fourchus est composé d'une succession de petites communautés relativement peu denses et reliées deux à deux. Ils peuvent dans certains cas avoir un alter central mais ces derniers ne le sont néanmoins pas suffisamment

pour apparaître au centre d'un grand nombre de graphlets étoilés.

Perspectives de développements futurs

Ce travail de thèse ouvre sur une large variété de potentielles suites. Comme tout travail quantitatif lié à l'étude des sociabilités, un complément par une approche qualitative permettrait dans un premier temps d'appréhender avec plus de certitudes les comportements assignés aux différentes classes d'usages et aux familles de réseaux. La question de la protection des données privées, centrales dans ce travail, qui a été discutée avec la CNIL, ne nous permet cependant pas de recontacter nos enquêtés et une nouvelle étude devra donc être mise en place sur ce sujet.

Comme cela a été dit dans le chapitre 3, les indicateurs de statuts n'ont jusqu'à présent été utilisés qu'à un niveau agrégé pour qualifier les egos. Une prochaine étape naturelle, et déjà entamée, consiste alors à étudier les publications elles-mêmes, afin de pouvoir analyser les relations entre caractère de la publication et réactions des amis. Une telle étude devrait alors tenir compte de la catégorie d'usages de l'enquêté par qui elle a été faite afin de contrebalancer les effets de leur usage habituel. On imagine en effet mal qu'une publication, disons politique, reçoive le même accueil par les amis d'ego selon que celui-ci soit un non-actif ou bien un egovisible. On pourrait alors envisager de normaliser les valeurs des indicateurs pour chaque statut selon ces mêmes valeurs pour l'ensemble des publications de l'enquêté.

Une autre étude possible, concernant les catégories d'usages, concerne la prise en compte de la dynamique des comptes Facebook. Puisqu'on a vu que l'âge apparaît être, parmi les indicateurs socio-démographique, le plus explicatif de l'appartenance à l'une de ces classes, on peut très bien imaginer que celle-ci change avec le temps. Découper les activités des répondants en plusieurs périodes pourrait ainsi permettre de voir si des changements s'opèrent entre elles.

Le travail sur les graphlets ouvre également de multiples perspectives. Un premier travail pourrait être de faire un zoom sur les communautés d'alters en n'opérant les énumérations de graphlets qu'à l'intérieur de celles proposées par l'algorithme de Louvain. On perdrait dans ce cas toute information sur les alters centraux dans le réseau, ainsi que les graphlets fourchus, qui n'apparaîtraient alors plus, mais on aurait plus d'information sur certains graphlets. En effet, on peut imaginer qu'un lien qui relie deux communautés a une grande influence sur l'énumération de quelques graphlets, mais ne change rien pour les plus denses. On serait alors plus à même de comprendre comment les différences de structure entre les communautés font apparaître plutôt des graphlets densifiés ou des graphlets troués par exemple. Par ailleurs, la caractérisation

des réseaux par la caractérisation de leurs communautés et par celles de leurs alters centraux semble être une piste intéressante.

Une future amélioration possible de la représentativité consisterait à la remplacer par un dérivé des résidus du χ^2, assez similaire.En effet, puisque les valeurs non normalisées supérieures à 1 sont recentrées entre 1 et 2 afin d'obtenir l'indicateur final, ce dernier est en quelque sorte écrasé au-dessus de la représentativité normale. Il en résulte que certains réseaux ayant pourtant une représentativité élevée d'un graphlet ne ressortent pas avec autant de clarté qu'on l'aurait souhaité. Le χ^2 présente en effet l'avantage d'avoir des valeurs aussi bien en dessous qu'au-dessus de zéro, permettant certainement la normalisation plus simple à rendre symétrique par la suite.

L'analyse entamée de positions des alters dans les réseaux personnels, dont j'ai parlé dans le chapitre 6, en plus d'être très riche en problématiques sociologiques, pourrait également participer à mieux comprendre la répartition des graphlets dans le réseau. Avec cette information, on pourrait ainsi savoir par exemple quelle est la part d'alters troués qui sont à cheval sur plusieurs groupes, et combien sont ceux qui sont incluent dans la même communauté. Il faudrait néanmoins peut-être dans ce cas rajouter une routine à l'algorithme d'énumération des graphlets afin de noter quels sont les sommets qui partagent des graphlets en communs, et quels graphlets.

La consolidation de la catégorisation en 5 familles de réseaux à partir des graphlets pourrait se faire en appliquant à nouveau la méthode des kMeans sur des sous-ensembles aléatoires du corpus de réseaux utilisés afin de voir si on retrouve les mêmes catégories.

Enfin, on pourrait étudier les énumérations de graphlets de plusieurs autres sortes de réseaux. Il serait très utile de savoir si les catégories de graphlets qu'on a construites se retrouvent également avec des réseaux biologiques par exemple, ou bien si elles sont vraiment propres aux réseaux sociaux.

Bibliographie

[Adamic et al., 2001] Adamic, L. A., Lukose, R. M., Puniyani, A. R., and Huberman, B. A. (2001). Search in power-law networks. *Physical review E*, 64(4) :046135.

[Ali et al., 2014] Ali, W., Rito, T., Reinert, G., Sun, F., and Deane, C. M. (2014). Alignment-free protein interaction network comparison. *Bioinformatics*, 30(17) :i430–i437.

[Arthur and Vassilvitskii, 2007] Arthur, D. and Vassilvitskii, S. (2007). k-means++ : The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics.

[Artzy-Randrup et al., 2004] Artzy-Randrup, Y., Fleishman, S. J., Ben-Tal, N., and Stone, L. (2004). Comment on" network motifs : simple building blocks of complex networks" and" superfamilies of evolved and designed networks". *science*, 305(5687) :1107–1107.

[Backstrom et al., 2012] Backstrom, L., Boldi, P., Rosa, M., Ugander, J., and Vigna, S. (2012). Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 33–42. ACM.

[Backstrom and Kleinberg, 2014] Backstrom, L. and Kleinberg, J. (2014). Romantic partnerships and the dispersion of social ties : a network analysis of relationship status on facebook. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 831–841. ACM.

[Bansal et al., 2007] Bansal, S., Grenfell, B. T., and Meyers, L. A. (2007). When individual behaviour matters : homogeneous and network models in epidemiology. *Journal of the Royal Society Interface*, 4(16) :879–891.

[Barabási and Albert, 1999] Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439) :509–512.

[Barnes, 1954] Barnes, J. A. (1954). Class and committees in a norwegian island parish. *Human Relations*, 7(1) :39–58.

[Barrat and Weigt, 2000] Barrat, A. and Weigt, M. (2000). On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13(3) :547–560.

[Barthélemy, 2011] Barthélemy, M. (2011). Spatial networks. *Physics Reports*, 499(1-3) :1–101.

[Bastard et al., 2013] Bastard, I., Cardon, D., Fouetillou, G., Prieur, C., and Raux, S. (2013). Travail et travailleurs de la donnée.

[Battista et al., 1998] Battista, G. D., Eades, P., Tamassia, R., and Tollis, I. G. (1998). *Graph drawing : algorithms for the visualization of graphs*. Prentice Hall PTR.

[Bavelas, 1950] Bavelas, A. (1950). Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, 22(6) :725–730.

[Beaudouin, 2009] Beaudouin, V. (2009). Les dynamiques des sociabilités.

[Bidart et al., 2011] Bidart, C., Degenne, A., and Grossetti, M. (2011). *La vie en réseau. Dynamique des relations sociales*. Presses universitaires de France.

[Bidart and Lavenu, 2005] Bidart, C. and Lavenu, D. (2005). Evolutions of personal networks and life events. *Social networks*, 27(4) :359–376.

[Blondel et al., 2008] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics : theory and experiment*, 2008(10) :P10008.

[Boccaletti et al., 2006] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks : Structure and dynamics. *Physics reports*, 424(4-5) :175–308.

[Borgatti, 1997] Borgatti, S. P. (1997). Structural holes : Unpacking burt's redundancy measures. *Connections*, 20(1) :35–38.

[Borgatti and Everett, 2006] Borgatti, S. P. and Everett, M. G. (2006). A graph-theoretic perspective on centrality. *Social networks*, 28(4) :466–484.

[Bott, 1957] Bott, E. (1957). *Family and social network : Roles, norms and external relationships in ordinary urban families*. Routledge.

[Brandes, 2001] Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2) :163–177.

[Brooks et al., 2014] Brooks, B., Hogan, B., Ellison, N., Lampe, C., and Vitak, J. (2014). Assessing structural correlates to social capital in facebook ego networks. *Social Networks*, 38 :1–15.

[Burke et al., 2011] Burke, M., Kraut, R., and Marlow, C. (2011). Social capital on facebook : Differentiating uses and users. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 571–580. ACM.

[Burke et al., 2010] Burke, M., Marlow, C., and Lento, T. (2010). Social network activity and social well-being. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1909–1912. ACM.

[Burt, 1992] Burt, R. S. (1992). Structural hole. *Harvard Business School Press, Cambridge, MA*.

[Cardon, 2010] Cardon, D. (2010). *La démocratie Internet : Promesses et limites*. Paris.

[Cardon and Prieur, 2016] Cardon, D. and Prieur, C. (2016). Comment la conversation a façonné le web. *L'Ordinaire d'internet*, pages 226–248.

[Casilli, 2010] Casilli, A. A. (2010). *Les Liaisons numériques. Vers une nouvelle sociabilité ? : Vers une nouvelle sociabilité ?* Le Seuil.

[Clauset et al., 2004] Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, 70(6) :066111.

[Corel et al., 2016] Corel, E., Lopez, P., Méheust, R., and Bapteste, E. (2016). Network-thinking : Graphs to analyze microbial complexity and evolution. *Trends in Microbiology*, 24(3) :224–237.

[Cristofoli, 2008] Cristofoli, P. (2008). Aux sources des grands réseaux d'interactions. *Réseaux*, 1(6) :21–58.

[Cunningham et al., 2013] Cunningham, P., Harrigan, M., Wu, G., and O'Callaghan, D. (2013). Characterizing ego-networks using motifs. *Network Science*, 1(02) :170–190.

[Dagiral and Martin, 2017] Dagiral, É. and Martin, O. (2017). Liens sociaux numériques. pour une sociologie plus soucieuse des techniques. *Sociologie*, (1, vol. 8).

[Davis, 1963] Davis, J. A. (1963). Structural balance, mechanical solidarity, and interpersonal relations. *American Journal of Sociology*, 68(4) :444–462.

[Davis and Leinhardt, 1967] Davis, J. A. and Leinhardt, S. (1967). The structure of positive interpersonal relations in small groups.

[Donath and Boyd, 2004] Donath, J. and Boyd, D. (2004). Public displays of connection. *bt technology Journal*, 22(4) :71–82.

[Dunbar, 1992] Dunbar, R. I. (1992). Neocortex size as a constraint on group size in primates. *Journal of human evolution*, 22(6) :469–493.

[Dunne et al., 2002] Dunne, J. A., Williams, R. J., and Martinez, N. D. (2002). Food-web structure and network theory : the role of connectance and size. *Proceedings of the National Academy of Sciences*, 99(20) :12917–12922.

[Ellison et al., 2007] Ellison, N., Steinfield, C., and Lampe, C. (2007). Te benefits of facebook "friends" : Exploring the relationship between college students' use of

online social networks and social capital. *Journal of Computer-Mediated Communication*, 12(3).

[Erdos and Rényi, 1960] Erdos, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1) :17–60.

[Eubank et al., 2004] Eubank, S., Guclu, H., Kumar, V. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z., and Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988) :180.

[Faust, 2007] Faust, K. (2007). 7. very local structure in social networks. *Sociological Methodology*, 37(1) :209–256.

[Fischer, 1982] Fischer, C. S. (1982). *To dwell among friends : Personal networks in town and city*. University of chicago Press.

[Flake et al., 2002] Flake, G. W., Lawrence, S., Giles, C. L., and Coetzee, F. M. (2002). Self-organization and identification of web communities. *Computer*, 35(3) :66–70.

[Freeman, 2004] Freeman, L. (2004). The development of social network analysis. *A Study in the Sociology of Science*, 1.

[Freeman, 1977] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

[Freeman, 1978] Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, 1(3) :215–239.

[Friedkin, 1980] Friedkin, N. (1980). A test of structural features of granovetter's strength of weak ties theory. *Social networks*, 2(4) :411–422.

[Gini, 1921] Gini, C. (1921). Measurement of inequality of incomes. *The Economic Journal*, 31(121) :124–126.

[Granovetter, 1974] Granovetter, M. (1974). Finding a job : A study on contacts and careers.

[Granovetter, 1977] Granovetter, M. S. (1977). The strength of weak ties. In *Social networks*, pages 347–367. Elsevier.

[Graovac et al., 2012] Graovac, A., Gotman, I., and Trinajstic, N. (2012). *Topological approach to the chemistry of conjugated molecules*, volume 4. Springer Science & Business Media.

[Grossetti, 2014] Grossetti, M. (2014). Que font les réseaux sociaux aux réseaux sociaux ? *Réseaux*, (2) :187–209.

[Guimera et al., 2005] Guimera, R., Mossa, S., Turtschi, A., and Amaral, L. N. (2005). The worldwide air transportation network : Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, 102(22) :7794–7799.

[Hecker et al., 2009] Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E., and Guthke, R. (2009). Gene regulatory network inference : data integration in dynamic models—a review. *Biosystems*, 96(1) :86–103.

[Henry, 2008] Henry, Nathalie et Fekete, J.-D. (2008). Représentations visuelles alternatives pour les réseaux sociaux. *Réseaux*, (6) :59–92.

[Héran, 1988] Héran, F. (1988). La sociabilité, une pratique culturelle. *Economie et statistique*, 216(1) :3–22.

[Herring et al., 2005] Herring, S. C., Kouper, I., Paolillo, J. C., Scheidt, L. A., Tyworth, M., Welsch, P., Wright, E., and Yu, N. (2005). Conversations in the blogosphere : An analysis" from the bottom up". In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 107b–107b. IEEE.

[Hočevar and Demšar, 2014] Hočevar, T. and Demšar, J. (2014). A combinatorial approach to graphlet counting. *Bioinformatics*, 30(4) :559–565.

[Holland and Leinhardt, 1971] Holland, P. W. and Leinhardt, S. (1971). Transitivity in structural models of small groups. *Comparative Group Studies*, 2(2) :107–124.

[Holland and Leinhardt, 1974] Holland, P. W. and Leinhardt, S. (1974). The statistical analysis of local structure in social networks.

[Holland and Leinhardt, 1977] Holland, P. W. and Leinhardt, S. (1977). A method for detecting structure in sociometric data. In *Social Networks*, pages 411–432. Elsevier.

[Homans, 1950] Homans, G. C. (1950). The human group new york. *Harpers*.

[Janic, 2007] Janic, M. (2007). Modelling the full costs of an intermodal and road freight transport network. *Transportation Research Part D : Transport and Environment*, 12(1) :33–44.

[Janssen et al., 2012] Janssen, J., Hurshman, M., and Kalyaniwalla, N. (2012). Model selection for social networks using graphlets. *Internet Mathematics*, 8(4) :338–363.

[Javed et al., 2018] Javed, M. A., Younis, M. S., Latif, S., Qadir, J., and Baig, A. (2018). Community detection in networks : A multidisciplinary review. *Journal of Network and Computer Applications*.

[Jeong et al., 2000] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804) :651.

[Jeong et al., 2007] Jeong, S.-J., Lee, C.-G., and Bookbinder, J. H. (2007). The european freight railway system as a hub-and-spoke network. *Transportation Research Part A : Policy and Practice*, 41(6) :523–536.

[Kalmijn, 2012] Kalmijn, M. (2012). Longitudinal analyses of the effects of age, marriage, and parenthood on social contacts and support. *Advances in Life Course Research*, 17(4) :177–190.

[Kashtan et al., 2004] Kashtan, N., Itzkovitz, S., Milo, R., and Alon, U. (2004). Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11) :1746–1758.

[Kreutzberger, 2008] Kreutzberger, E. D. (2008). Distance and time in intermodal goods transport networks in europe : A generic approach. *Transportation Research Part A : Policy and Practice*, 42(7) :973–993.

[Kwak et al., 2010] Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media ? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

[Lagesse et al., 2016] Lagesse, C., Bonnin, P., Bordin, P., and Douady, S. (2016). Méthodologie de modélisation et de caractérisation des réseaux spatiaux. application au réseau viaire de paris. *Flux*, 1(3) :33–49.

[Lambiotte et al., 2008] Lambiotte, R., Blondel, V. D., De Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., and Van Dooren, P. (2008). Geographical dispersal of mobile communication networks. *Physica A : Statistical Mechanics and its Applications*, 387(21) :5317–5325.

[Lampe et al., 2006] Lampe, C., Ellison, N., and Steinfield, C. (2006). A face (book) in the crowd : Social searching vs. social browsing. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 167–170. ACM.

[Lazega, 2014] Lazega, E. (2014). *Réseaux sociaux et structures relationnelles : Que sais-je ? n3399*. Presses universitaires de France.

[Lazega et al., 2007] Lazega, E., Jourda, M.-T., Mounier, L., and Stofer, R. (2007). Des poissons et des mares : l'analyse de réseaux multi-niveaux. *Revue française de sociologie*, 48(1) :93–131.

[Leavitt, 1951] Leavitt, H. J. (1951). Some effects of certain communication patterns on group performance. *The Journal of Abnormal and Social Psychology*, 46(1) :38.

[Lemercier, 2005] Lemercier, C. (2005). Analyse de réseaux et histoire. *Revue d'histoire moderne et contemporaine*, (2) :88–112.

[Licoppe and Smoreda, 2000] Licoppe, C. and Smoreda, Z. (2000). Liens sociaux et régulations domestiques dans l'usage du téléphone. de l'analyse quantitative de la durée des conversations à l'examen des interactions. *Réseaux*, 18(103) :253–276.

[Manceron et al., 2002] Manceron, V., Lelong, B., and Smoreda, Z. (2002). La naissance du premier enfant. *Réseaux*, (5) :91–120.

[Marsden and Campbell, 1984] Marsden, P. V. and Campbell, K. E. (1984). Measuring tie strength. *Social forces*, 63(2) :482–501.

[McCallister and Fischer, 1978] McCallister, L. and Fischer, C. S. (1978). A procedure for surveying personal networks. *Sociological Methods & Research*, 7(2) :131–148.

[Mercklé, 2011] Mercklé, P. (2011). *Sociologie des réseaux sociaux*. La découverte.

[Milardo, 1988] Milardo, R. M. (1988). *Families and social networks*. Sage Publications, Inc.

[Milo et al., 2004] Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U. (2004). Superfamilies of evolved and designed networks. *Science*, 303(5663) :1538–1542.

[Milo et al., 2002] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs : simple building blocks of complex networks. *Science*, 298(5594) :824–827.

[Moreno et al., 1934] Moreno, J. L., Jennings, H. H., et al. (1934). *Who shall survive ?* Nervous and mental disease publishing co.

[Nasim et al., 2016] Nasim, M., Charbey, R., Prieur, C., and Brandes, U. (2016). Investigating link inference in partially observable networks : Friendship ties and interaction. *IEEE Transactions on Computational Social Systems*, 3(3) :113–119.

[Newman, 2010] Newman, M. (2010). *Networks : an introduction*. Oxford university press.

[Newman, 2003] Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2) :167–256.

[Newman and Girvan, 2004] Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2) :026113.

[Nguyen and Lethiais, 2016] Nguyen, G. D. and Lethiais, V. (2016). Impact des réseaux sociaux sur la sociabilité : le cas de facebook. *Réseaux*, 34(195) :165–195.

[Onnela et al., 2007] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the national academy of sciences*, 104(18) :7332–7336.

[Padgett and Ansell, 1993] Padgett, J. F. and Ansell, C. K. (1993). Robust action and the rise of the medici, 1400-1434. *American journal of sociology*, 98(6) :1259–1319.

[Page et al., 1997] Page, L., Brin, S., Motwani, R., and Winograd, T. (1997). Pagerank : Bringing order to the web. Technical report, Stanford Digital Libraries Working Paper.

[Paine, 1966] Paine, R. T. (1966). Food web complexity and species diversity. *The American Naturalist*, 100(910) :65–75.

[Pasquier, 2005] Pasquier, D. (2005). Cultures lycéennes : La tyrannie de la majorité. *Autrement. Série mutations*, (235) :4–180.

[Pimm et al., 1991] Pimm, S. L., Lawton, J. H., and Cohen, J. E. (1991). Food web patterns and their consequences. *Nature*, 350(6320) :669.

[Pržulj, 2007] Pržulj, N. (2007). Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2) :e177–e183.

[Pržulj et al., 2004] Pržulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling interactome : scale-free or geometric ? *Bioinformatics*, 20(18) :3508–3515.

[Pržulj et al., 2006] Pržulj, N., Corneil, D. G., and Jurisica, I. (2006). Efficient estimation of graphlet frequency distributions in protein–protein interaction networks. *Bioinformatics*, 22(8) :974–980.

[Putnam, 2000] Putnam, R. D. (2000). Bowling alone : America's declining social capital. In *Culture and politics*, pages 223–234. Springer.

[Rivière, 2000] Rivière, C. (2000). Les réseaux de sociabilité téléphonique. *Revue française de sociologie*, pages 685–717.

[Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20 :53–65.

[Rual et al., 2005] Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., et al. (2005). Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062) :1173.

[Sabidussi, 1966] Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4) :581–603.

[Scott, 2017] Scott, J. (2017). *Social network analysis*. Sage.

[Sharan et al., 2005] Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R. M., and Ideker, T. (2005). Conserved patterns of protein interaction in multiple species. *Proceedings of the National Academy of Sciences of the United States of America*, 102(6) :1974–1979.

[Shaw, 1954] Shaw, M. E. (1954). Group structure and the behavior of individuals in small groups. *The Journal of psychology*, 38(1) :139–149.

[Shen-Orr et al., 2002] Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, 31(1) :64.

[Smith, 2011] Smith, A. (2011). Why americans use social media. *Pew Internet & American Life Project*, pages 1–11.

[Stephen and Toubia, 2009] Stephen, A. T. and Toubia, O. (2009). Explaining the power-law degree distribution in a social commerce network. *Social Networks*, 31(4) :262–270.

[Stoica and Prieur, 2009] Stoica, A. and Prieur, C. (2009). Structure of neighborhoods in a large social network. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 26–33. IEEE.

[Stoica et al., 2013] Stoica, A., Smoreda, Z., and Prieur, C. (2013). A local structure-based method for nodes clustering : Application to a large mobile phone social network. In *The Influence of Technology on Social Network Analysis and Mining*, pages 157–184. Springer.

[Travers and Milgram, 1967] Travers, J. and Milgram, S. (1967). The small world problem. *Phychology Today*, 1(1) :61–67.

[Vazquez et al., 2003] Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nature biotechnology*, 21(6) :697.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, 393(6684) :440.

[Wellman, 1997] Wellman, B. (1997). Structural analysis : From method and metaphor to theory and substance. *Contemporary Studies in Sociology*, 15 :19–61.

[Wernicke, 2006] Wernicke, S. (2006). Efficient detection of network motifs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(4).

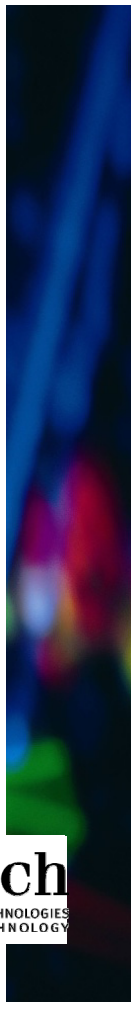
[White et al., 1976] White, H. C., Boorman, S. A., and Breiger, R. L. (1976). Social structure from multiple networks. i. blockmodels of roles and positions. *American journal of sociology*, 81(4) :730–780.

[Williams, 2006] Williams, D. (2006). On and off the'net : Scales for social capital in an online era. *Journal of computer-mediated communication*, 11(2) :593–628.

[Xie and Levinson, 2007] Xie, F. and Levinson, D. (2007). Measuring the structure of road networks. *Geographical analysis*, 39(3) :336–356.

[Yaveroğlu et al., 2014] Yaveroğlu, Ö. N., Malod-Dognin, N., Davis, D., Levnajic, Z., Janjic, V., Karapandza, R., Stojmirovic, A., and Pržulj, N. (2014). Revealing the hidden language of complex networks. *Scientific reports*, 4 :4547.

[Yaveroğlu et al., 2015] Yaveroğlu, Ö. N., Milenković, T., and Pržulj, N. (2015). Proper evaluation of alignment-free network comparison methods. *Bioinformatics*, 31(16) :2697–2704.



TELECOM ParisTech

EDiE

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY