Université de Lille

FACULTÉ DES SCIENCES ET TECHNOLOGIES
Département Mathématiques

Travail encadré de recherche

Étude d'un réseau social inconnu

Mohamed Metidji & Jean Hiouani

Encadré par

Chi Tran Cyrielle Dumont Anthony Cousien

Introduction

L'échantillonnage de populations dont il est impossible de connaître a priori les contours avec précision constitue un défit majeur. Ces populations sont difficiles à recruter dans des enquêtes épidémiologiques ou sociologique du fait de certaines de leurs caractéristiques potentiellement stigmatisantes (comme les usagers de drogues, les personnes LGBTQ...), ou plus simplement rares car trop spécifiques pour être saisies par les outils statistiques officiels.
Certaines solutions ont néanmoins été proposées pour tenter de produire des données fiables dans ce type de cas. Parmi les procédures d'échantillonnage alternatives, c'est indubitablement la méthode du Respondant Driven Sampling (RDS) [1] qui peut être traduit par «échantillonnage guidé par les répondants», introduite à la fin des années 1990 par Douglas Heckathorn qui a connu la plus forte popularité.

Dans ce TER, nous explorerons en profondeur les différents aspects de cette méthode par le biais d'une étude sur l'estimation de la population de consommateurs de drogues injectables à Paris.

Table des matières

Chapitre 1

Contexte et description des données

1.1 Estimation des usagers de drogues à Paris

Nos données sont issus d'une enquête regroupant plusieurs laboratoires de recherche dans le but d'estimer la population d'usagers de drogues injectable à Paris selon la méthode du RDS afin de pouvoir proposer des solutions pour endiguer la transmission de maladies infectieuses au sein de ce réseau.

Cette enquête et les données de cette étude ont été financées par l'Agence Nationale de Recherche sur le Sida et les Hépatites virales (ANRS, http ://www.anrs.fr, grant number 95146) ainsi qu'en partie par le Labex CEMPI (ANR-11-LABX-0007-01). Pour plus d'informations voir :[2].

1.2 La méthode du Respondent-driven sampling

Le protocole commence par la sélection d'un petit nombre c d'informateurs initiaux (ici $c = 3$), ces primo participants sont appelés «graines». Ces graines sont ensuite invitées à recruter deux ou trois de leurs contacts personnels correspondant aux critères de définition de la population. Pour cela, elles sont dotées de «coupons» que leurs contacts doivent retourner aux enquêteurs lors de leur interview de manière à s'assurer qu'il y a bien eu un échange entre les individus recruteurs et les futurs recrutés. Une autre recommandation pour s'assurer la participation active des recruteurs est de fournir une rétribution matérielle ou symbolique sous condition que les contacts participent effectivement à l'enquête et retournent les fameux «coupons» aux enquêteurs. Le même procédé est répété à chaque vague avec les personnes nouvellement contactées.
La technique du RDS est basée sur une idée assez simple et bien connue depuis les travaux de Stuart Milgram sur le phénomène du «small world» [3] : statistiquement, tous les individus d'un même groupe social sont liés entre eux par un nombre limité de liens. Autrement dit, quel que soit le point de départ des

chaînes de relations, on peut potentiellement atteindre n'importe quel individu du groupe en question au bout de n mises en relation successives. Fort de ce constat, dans le premier article qu'il consacre au RDS en 1997, D. Heckathorn défend donc l'idée suivante : lorsque la sélection des informateurs privilégiés choisis pour nouer des contacts dans la population cible est non-aléatoire et arbitraire, si l'on multiplie les vagues de recrutement à partir de ce petit groupe initial, la sélection des personnes intégrées dans l'échantillon dépendra de moins en moins des graines initiales et va tendre à devenir aléatoire.

1.3 Données récoltées au cours de l'enquête

Au cours de l'enquête, on a obtenu une base de données de 399 individus et de 8 variables. Pour les variables, on a :

— Le **n° égo** donnant l'identifiant de l'individu interrogé.

— Le **nombre de citations** donnant le nombre de fois où l'individu a été cité par d'autres participants à l'enquête.

— La **graine** qui donne un identifiant pour les individus correspondant aux «graînes» de la méthode RDS, cet identifiant nous sert à connaître quel est l'enquêteur qui l'a interrogé ainsi qu'à quelle zone s'est déroulé ce questionnaire.

— L'**âge** et le **sexe**.

— La variable **coupon/interrogé** qui permet de savoir si l'individu a été interrogé par un enquêteur ou bien a répondu au questionnaire après avoir reçu un coupon.

— Et enfin l'**identifiants des gens qui le citent** ainsi que le **numéro de questionnaire** qui sont suffisament explicites.

Chapitre 2

Analyse descriptive

Afin de nous appuyer sur de solides bases pour notre analyse, nous allons utiliser les outils développés en *théorie des graphes*.

2.1 Théorie des graphes

La théorie des graphes étant un domaine d'étude tellement vaste qu'elle pourrait faire l'objet entier de notre TER, nous n'en développerons que les éléments utiles à notre étude.
Tout d'abord définissons l'objet central de notre sujet :

Définition 2.1.1.
Un graphe $\mathcal{G} = (V, E)$ est un couple d'objets appelés sommets V (vertices) et de relations liant ces objets appelées arêtes E (edges). Si les arêtes sont non orientées, la relation va dans les deux sens et est donc symétrique, et le graphe est dit non orienté. Sinon les arêtes sont orientées et appelées flèches, et la relation va dans un seul sens et est donc asymétrique, et le graphe est dit orienté.

La plupart du temps, on associe des informations aux sommets et aux arêtes. Dans notre étude, les sommets V correspondent aux usagers de drogues et les arêtes E signifient que les sommets qui ont pratiqués l'injection ensemble au cours du dernier mois.

Historiquement, les graphes sont surtout étudiés en mathématiques discrètes. En parallèle, les réseaux (sociaux, par exemple) sont étudiés en sciences sociales. Typiquement, en sciences sociales, on s'intéresse aux questions de :

— *centralité*, c'est-à-dire qui est influent et le mieux connecté.

— *connexité*, c'est-à-dire comment les individus sont reliés les uns aux autres.

Depuis récemment, on s'intéresse aux statistiques à grande échelle de ces réseaux. On se pose par exemple la question suivante : «Quel sommet est le plus central dans ce réseau s'il était supprimé ?». Une autre question est : «Combien de sommets doit-on supprimer pour rendre le graphe non connexe ?»

Dans ce sens, nous étudierons tout d'abord une propriété essentielle des graphes, leur *connexité* :

Définition 2.1.2.
Un graphe non orienté est dit connexe si quels que soient les sommets u et v, il existe une chaîne reliant u à v. Pour un graphe orienté, on parle de connexité si en oubliant l'orientation des arêtes le graphe est connexe. On parle de forte connexité s'il existe un chemin orienté de u vers v.
Une composante connexe d'un graphe est un sous-graphe connexe de ce graphe.

Dans notre étude ces composantes représentent des groupes de toxicomanes se droguant régulièrement ensemble selon les relations définies par les arêtes du graphe.

Une virus se propagera plus rapidement s'il n'y a que 6 intermédiaires plutôt que 100 intermédiaires. Pour cela, nous nous intéresserons par la suite à la *longueur géodésique* :

Définition 2.1.3.
Une géodésique de u à v est une chaîne de u à v de longueur $d(u, v)$ qui correspond au nombre minimal de sommets par lequel il est nécessaire de passer pour relier u à v. Si des sommets ne sont pas reliés par un chemin, alors la distance entre les 2 sommets sera infinie.

Un paquet d'informations sera transmis plus rapidement sur Internet si la distance géodésique moyenne est petite. Dans un réseau de transport, cela permet de minimiser les correspondances.

Il est alors naturel de se demander l'effet que peut avoir la suppression de certains sommets dans la structure du réseau. Cela nous conduira à nous pencher sur les *points d'articulation* de notre graphe :

Définition 2.1.4.
Un point d'articulation est un sommet d'un graphe qui, si on le retire de ce dernier, augmente le nombre de composantes connexes. Si le graphe était connexe avant de retirer ce sommet, il devient donc non connexe. Ces points peuvent selon les cas révéler une vulnérabilité ou une position stratégique dans un réseau.

Lorsqu'on souhaite modéliser la propagation d'une maladie, la suppression d'un sommet peut correspondre à la vaccination d'une personne contre la maladie en question.

Enfin, nous étudierons le *degré* des sommets de notre graphe :

Définition 2.1.5.
Dans le cas d'un graphe orienté on parle du degré entrant d'un sommet, c'est-à-dire le nombre d'arcs dirigés vers le sommet, et du degré sortant de ce sommet, c'est-à-dire le nombre d'arcs sortant de ce dernier. Le degré total du sommet est la somme du degré sortant et du degré entrant.

2.2 Description globale du graphe

Pour analyser et visualiser l'ensemble des propriétés de notre graphe, nous nous aiderons du package igraph[4] de R. On peut espérer «bien» visualiser des graphes de quelques milliers de sommets. En effet, l'oeil humain permet souvent de capturer des propriétés plus difficiles à détecter de façon informatique. En revanche, lorsqu'on a plusieurs millions ou même milliards de sommets, ce n'est pas envisageable. Pour se donner une idée globale du réseau que l'on va étudier nous allons tracer le graphe associé :

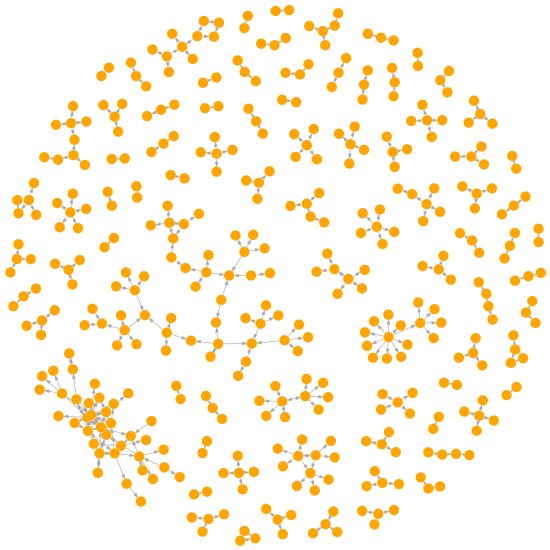

FIGURE 2.1 – Visualisation du réseau

On remarque d'abord que le graphe n'est pas d'un seul tenant, il n'est pas connexe. En effet, notre graphe est divisé en 82 composantes connexes de tailles variables. Essayons d'expliquer ce qui peut justifier cette topologie particulière.

2.2.1 Composantes connexes

On peut representer graphiquement le nombre de composantes connexes en fonction du nombre de sommets qu'elles incluent.

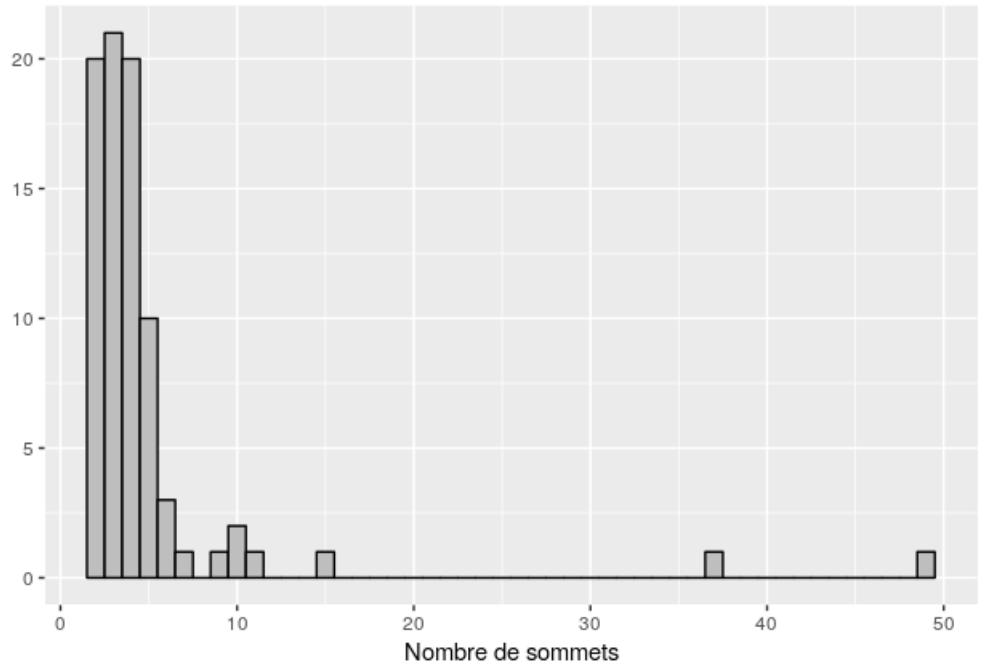

FIGURE 2.2 – Distribution de la taille des composantes connexes

On observe que dans notre réseau, on a un nombre important de petites composantes connexes de taille 2 à 11 qui représente environ 74,69 % des 399 sommets de notre graphe et 2 autres composantes de grandes tailles qui represent 21,55 % du nombre de sommets. On pourrait interpréter cela du fait que la majorité des gens préfèrent se droguer en petit comité avec des amis ou peut-être qu'il existe une sorte de relation élève-professeur qui existe au sein du réseau.

Par ailleurs, la nature de notre méthode d'échantillonage pourrait expliquer le fait d'avoir de nombreuses composantes connexes. Traçons alors notre graphe en mettant en évidence les «graines» :

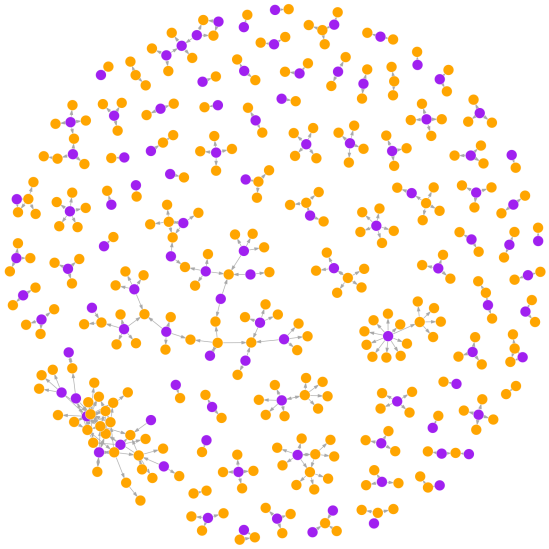

FIGURE 2.3 – Visualisation du les graines

Nous avons mis en violet les sommets du graphe représentant les 104 graines. Croisons alors la taille des composantes avec le nombre de graines qu'elles contiennent :

Taille composante \ Nombre de graine	0	1	2	4	8	14
2	2	17	1	0	0	0
3	2	19	0	0	0	0
4	0	19	1	0	0	0
5	0	10	0	0	0	0
6	0	3	0	0	0	0
7	0	1	0	0	0	0
9	0	0	1	0	0	0
10	0	1	0	1	0	0
11	0	1	0	0	0	0
15	0	1	0	0	0	0
37	0	0	0	0	1	0
49	0	0	0	0	0	1

En faisant un test d'indépendance du χ^2, on obtient une p-valeur inférieur à $2.35e^{-24}$ et on rejette alors l'hypothèse d'indépendance. Par conséquent, la taille des composantes connexes pourrait être expliquée par le nombre de graines qui la compose. Cependant, on peut naturellement se dire que plus une composante connexe est importante et plus elle a de chances de contenir un grand nombre de graines.

Analysons de plus près les deux plus grandes composantes connexes.

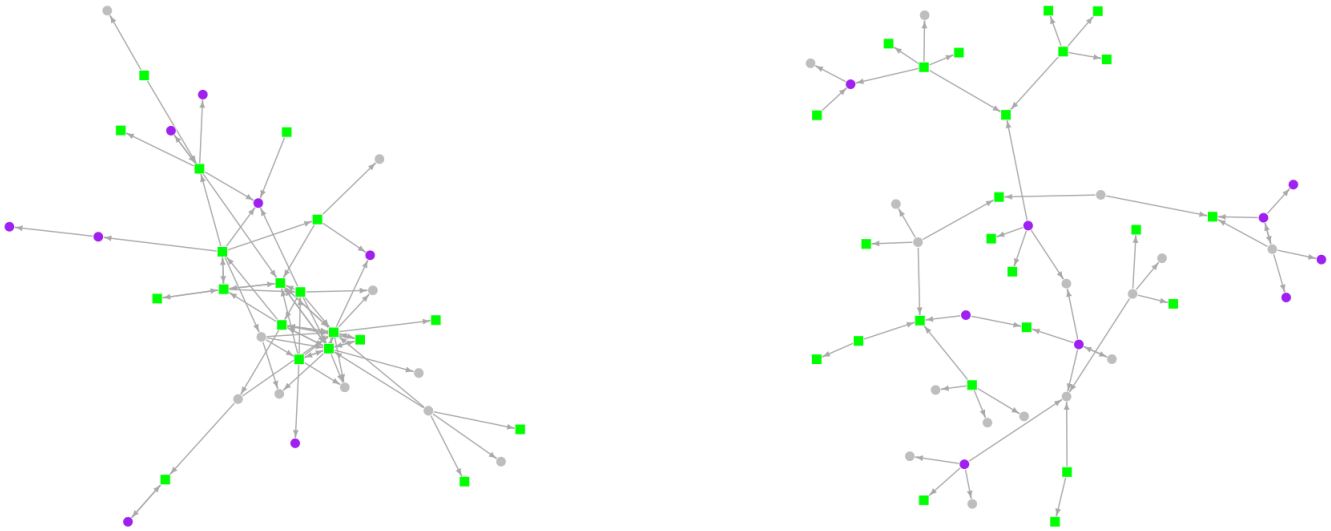

FIGURE 2.4 – Visualisation des composantes principales selon la graine

Ces groupes sont composés de 49 individus dont 14 graines pour la plus grande composante (à gauche) et de 37 dont 8 graines pour la seconde (à droite). On ne distingue pas graphiquement de topologie particulière pour la composante de gauche, quant à celle de droite on peut discerner une certaine arborescence. Dans ce sens, on remarque que la première composante connexe possède certains sommets qui sont en relations avec beaucoup d'autres alors que pour la deuxième il y en a très peu.

Par ailleurs, au vu de la géométrie de notre graphe et de ses composantes, on peut s'intéresser à une autre propriété qui caractérise les liens entre les différents individus : la longueur géodésique.

2.2.2 Longueur géodésique

Nous pouvons calculer la *longueur géodésique moyenne*, qu'on note ℓ, qui permet de qualifier le nombre d'arrête nécessaire pour que l'on puisse relier deux sommets en moyenne.
Dans les réseaux présentant la propriété du «petit monde» de Milgram, la longueur géodésique moyenne est souvent très petite [5] :

Réseau	n	m	ℓ
IMDb	449 913	25 516 482	3.48
Articles(math)	253 339	496 489	7.57
Articles(biologie)	1 520 251	11 803 064	4.92
Citations	783 339	6 716 198	8.57
Réseau peer-to-peer	880	1 296	4.28
Réseau électrique	4 941	6 594	2.67
Chaîne alimentaire marine	135	598	4.43

Si des sommets ne sont pas reliés par un chemin, alors la distance entre les 2 sommets sera infinie. De fait lorsque l'on a plusieurs composantes connexes comme c'est le cas ici, on privilégiera cette formule pour le calcul de notre longueur geodesique moyenne :

$$\ell^{-1} = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d(i,j)^{-1}$$

De plus, nous supposerons le graphe non orienté car l'intérêt est de connaître la distance entre les individus. L'interprétation dans notre contexte est que si ℓ est petit, les individus se connaissent tous entre eux à quelques personnes près. On trouve $\ell = 4.4$ pour notre graphe, donc en moyenne il faut quatre arrêtes pour lier deux individus d'une même composante connexe c'est-à-dire que les personnes d'un même groupe se connaissent tous à quatre personnes près. Dans le cadre de notre étude, si l est petit, une infection va se propager plus rapidement au sein du groupe. Ce résultat est biaisé à cause du nombre important de composantes connexes. Si on recalcule cette longeur en se limitant seulement aux deux composantes principales, on trouve $\ell = 3.0$ pour la plus grande et $\ell = 6.6$ pour la seconde La différence de structure entre ces deux composantes influence donc la proximité entre les sommets de manière conséquente. Dans ce sens, intéressons nous à l'importance de certains points dans la structure de notre graphe.

2.2.3 Points d'articulation

Dans la plupart des réseaux, la connexité entre différents sommets est fondamentale. Regardons maintenant les *points d'articulation* du graphe :

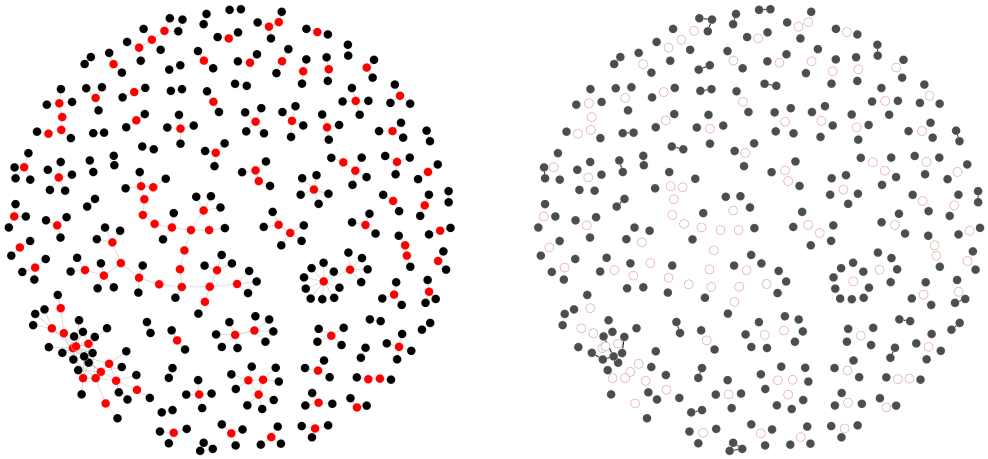

FIGURE 2.5 – Visualisation des points d'articulation

La première chose qui nous vient à l'esprit quand on regarde le graphe ainsi représenté, c'est que la plupart des points d'articulations (en rouge) semblent être les graines que l'on a visualisé plus haut. Croisons alors ces deux variables :

Point d'articulation \ Graine	Non	Oui
Non	257	33
Oui	38	71

En faisant un test d'indépendance du χ^2, on obtient une p-valeur inférieur à $1.2e^{-27}$ et on rejette alors l'hypothèse d'indépendance. On perçoit donc l'importance des graines dans notre méthode d'échantillonage pour estimer des communautés cachés. En effet, ces graînes nous permettent de recruter par la suite les individus cachés que nous ne pourrions pas atteindre avec des méthodes de recrutement habituelles.

De plus, on constate l'influence des points d'articulation dans la structure de notre graphe. Ainsi, si on isole ces points en supprimant leurs liens avec tout autre sommet le nombre de composantes connexes de notre graphe passe de 82 à 368. Par conséquent, ces points d'articulations sont d'une importance capitale dans la lutte contre la propagation de maladies au sein du réseau.

2.2.4 Exploration noeuds-centrés

Dans cette section, nous nous intéressons à la distribution du *degré* de chaque sommet.
Visualisons comment les degrés de notre graphe sont distribués :

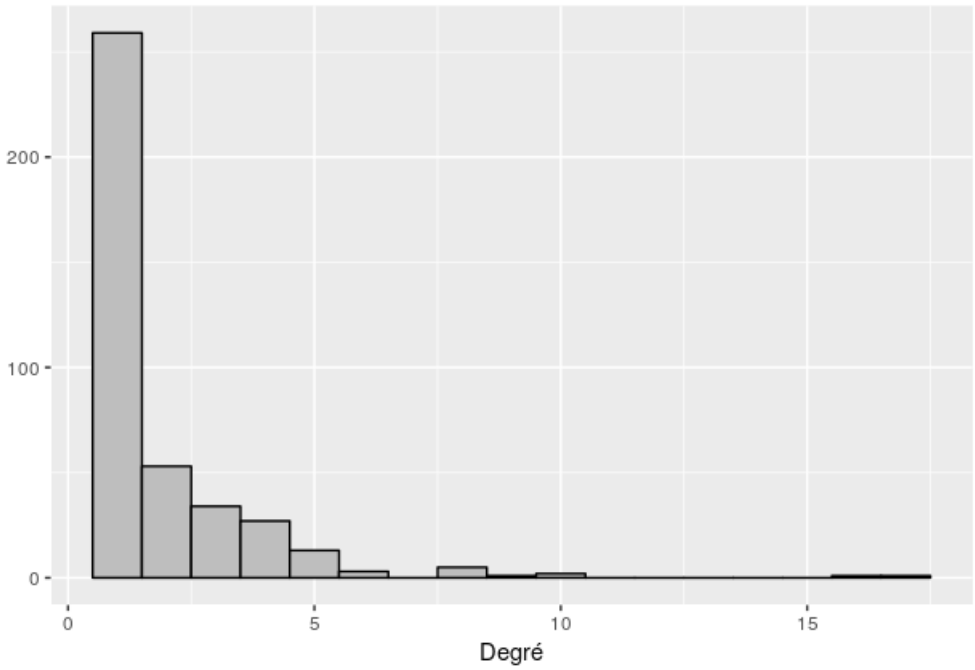

FIGURE 2.6 – Histogramme de la distribution du degré des sommets

Pour un graphe aléatoire, on peut modéliser le degré d'un sommet par une variable aléatoire D. De plus, beaucoup de graphes réels ont une distribution des degrés des noeuds qui s'ajuste correctement sur une *loi de puissance discrète*, c'est-à-dire telle que :

$$\forall k \geq 1 \qquad \mathbb{P}(D = k) = \frac{C_\alpha}{k^\alpha}$$

Avec C_α une constante de normalisation. De plus, on doit avoir $\alpha > 1$ car la série $\sum_{k=1}^{+\infty} \mathbb{P}(D = k)$ divergerait. En pratique, peu de données réelles suivent une loi puissance pour toutes les valeurs de k. Le plus souvent, la loi puissance ne s'applique que pour des valeurs supérieures à un certain k_{min}. Dans de tels cas, on dit que la queue de distribution suit une loi de puissance.

Proposition 2.2.1. *La constante de normalisation vaut :*

$$C_\alpha = \frac{1}{\zeta(\alpha, k_{min})} \quad \textit{avec} \quad \zeta(\alpha, k_{min}) = \sum_{n=0}^{+\infty} (n + k_{min})^{-\alpha}$$

Démonstration.

$$\sum_{i=k_{min}}^{+\infty} \mathbb{P}(D=i) = 1$$
$$\sum_{i=k_{min}}^{+\infty} C_\alpha i^{-\alpha} = 1$$
$$\sum_{i=k_{min}}^{+\infty} i^{-\alpha} = C_\alpha^{-1}$$

En posant $n = i - k_{min}$, on obtient :

$$\sum_{n=0}^{+\infty} (n + k_{min})^{-\alpha} = C_\alpha^{-1}$$

□

En pratique, peu de données réelles suivent une loi puissance pour toutes les valeurs de k. Le plus souvent, la loi puissance ne s'applique que pour des valeurs supérieures à un certain k_{min}. Dans de tels cas, on dit que la queue de distribution suit une loi de puissance.

Estimation du paramètre α

Supposons que D suit une loi puissance. Tout d'abord, tâchons de faire une estimation du paramètre α. Pour estimer α nous avons besoin, comme nous allons le voir, du paramètre k_{min}. Pour le moment, considérons que cette valeur est connue. Dans le cas où ce paramètre n'est pas connu, nous pouvons l'estimer à partir de nos données.
Nous allons estimer α par la méthode du maximum de vraissemblance.

Proposition 2.2.2. *L'estimateur du maximum de vraissemblance de α est solution de l'équation :*

$$\frac{\zeta'(\hat{\alpha}, k_{min})}{\zeta(\hat{\alpha}, k_{min})} = -\frac{1}{n}\sum_{i=1}^{n} \ln(x_i)$$

Démonstration.
Soit $X_1, \ldots, X_n$ un n-échantillon de variables aléatoires indépendantes et de même loi que D, on calcule alors la vraissemblance :

$$\begin{aligned}\mathcal{L}(\alpha, X_1, \ldots, X_n) &= \prod_{i=1}^{n} \mathbb{P}_\alpha(X_i = x_i) \\ &= \prod_{i=1}^{n} \frac{x_i^{-\alpha}}{\zeta(\alpha, k_{min})}\end{aligned}$$

En passant à la log-vraissemblance, on obtient :

$$\begin{aligned}\mathcal{L}og(\alpha) &= \ln \prod_{i=1}^{n} \frac{xi^{-\alpha}}{\zeta(\alpha, k_{min})} \\ &= -n \ln \zeta(\alpha, k_{min}) - \alpha \sum_{i=1}^{n} \ln\end{aligned}$$

En calculant $\dfrac{\partial \mathcal{L}og(\alpha)}{\partial \alpha} = 0$, on trouve :

$$\frac{\zeta'(\hat{\alpha}, k_{min})}{\zeta(\hat{\alpha}, k_{min})} = -\frac{1}{n} \sum_{i=1}^{n} \ln(x_i)$$

□

Il est possible de résoudre cette équation numériquement ou on peut aussi directement maximiser la log-vraissemblance. Bien qu'il n'y ait pas d'expression exacte de $\hat{\alpha}$, on peut en trouver une expression approximative :

Proposition 2.2.3. *Une approximation de l'estimateur du maximum de vraissemblance de α est :*

$$\hat{\alpha} \simeq 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{k_{min} - \frac{1}{2}} \right]^{-1}$$

Démonstration.

Soit f une fonction dérivable et F une primitive de f telle que $F'(x) = f(x)$.

$$\int_{k-\frac{1}{2}}^{k+\frac{1}{2}} f(t)dt = F(k + \frac{1}{2}) - F(k - \frac{1}{2})$$

En utilisant la formule de Taylor-Young, on a :

$$\begin{aligned}\int_{k-\frac{1}{2}}^{k+\frac{1}{2}} f(t)dt &= [F(k) + \frac{1}{2}F'(k) + \frac{1}{8}F''(k) + \frac{1}{48}F'''(k)] - [F(k) - \frac{1}{2}F'(k) + \frac{1}{8}F''(k) - \frac{1}{48}F'''(k)] + \cdots \\ &= f(k) + \frac{1}{24}f''(k) + \cdots\end{aligned}$$

En sommant sur k, on obtient :

$$\int_{k_{min}-\frac{1}{2}}^{+\infty} f(t)dt = \sum_{k=k_{min}}^{+\infty} f(k) + \frac{1}{24} \sum_{k=k_{min}}^{+\infty} f''(k) + \cdots$$

Pour $f(k) = k^{-\alpha}$ avec α constant, on a :

$$\begin{aligned}\int_{k_{min}-\frac{1}{2}}^{+\infty} t^{-\alpha}dt &= \frac{(k_{min} - \frac{1}{2})^{-\alpha+1}}{\alpha - 1} \\ &= \sum_{k=k_{min}}^{+\infty} k^{-\alpha} + \frac{\alpha(\alpha+1)}{24} \sum_{k=k_{min}}^{+\infty} k^{-\alpha-2} + \cdots \\ &= \zeta(\alpha, k_{min})[1 + \mathcal{O}(k_{min}^{-2})]\end{aligned}$$

En utilisant le fait que $k^{-2} \leq k_{min}^{-2}$ pour tous les termes de la seconde série. Par conséquent :

$$\zeta(\alpha, k_{min}) = \frac{(k_{min} - \frac{1}{2})^{-\alpha+1}}{\alpha - 1}[1 + \mathcal{O}(k_{min}^{-2})]$$

On dérive l'expression ainsi obtenue selon α :

$$\zeta'(\alpha, k_{min}) = -\frac{(k_{min} - \frac{1}{2})^{-\alpha+1}}{\alpha - 1}\left[\frac{1}{\alpha - 1} + \ln(k_{min} - \frac{1}{2})\right][1 + \mathcal{O}(k_{min}^{-2})]$$

On obtient alors :

$$\frac{\zeta'(\alpha, k_{min})}{\zeta(\alpha, k_{min})} = -\left[\frac{1}{\alpha - 1} + \ln(k_{min} - \frac{1}{2})\right][1 + \mathcal{O}(k_{min}^{-2})]$$

L'équation que l'on devait résoudre pour trouver l'estimateur de α devient :

$$-\left[\frac{1}{\hat{\alpha} - 1} + \ln(k_{min} - \frac{1}{2})\right][1 + \mathcal{O}(k_{min}^{-2})] = \frac{1}{n}\sum_{i=1}^{n}\ln(x_i)$$

Et en négligeant les quantités d'ordre k_{min}^{-2} en comparaison des quantités d'ordre 1, on trouve alors :

$$\hat{\alpha} \simeq 1 + n\left[\sum_{i=1}^{n}\ln\frac{x_i}{k_{min} - \frac{1}{2}}\right]^{-1}$$

□

Pour notre graphe, en considérant $k_{min} = 1$, on trouve $\hat{\alpha} = \simeq 1.91$.

Une autre manière d'estimer le paramètre α pour notre loi puisssance est d'utiliser la régression linéaire sur nos données passées en échelle logarithmique.[6]

On récupère alors la distribution des degrés, puis en faisant une régression linéaire on obtient une estimation du paramètre α de la loi puissance, celle-ci étant la valeur absolue de la pente de la droite obtenue par régression linéaire.

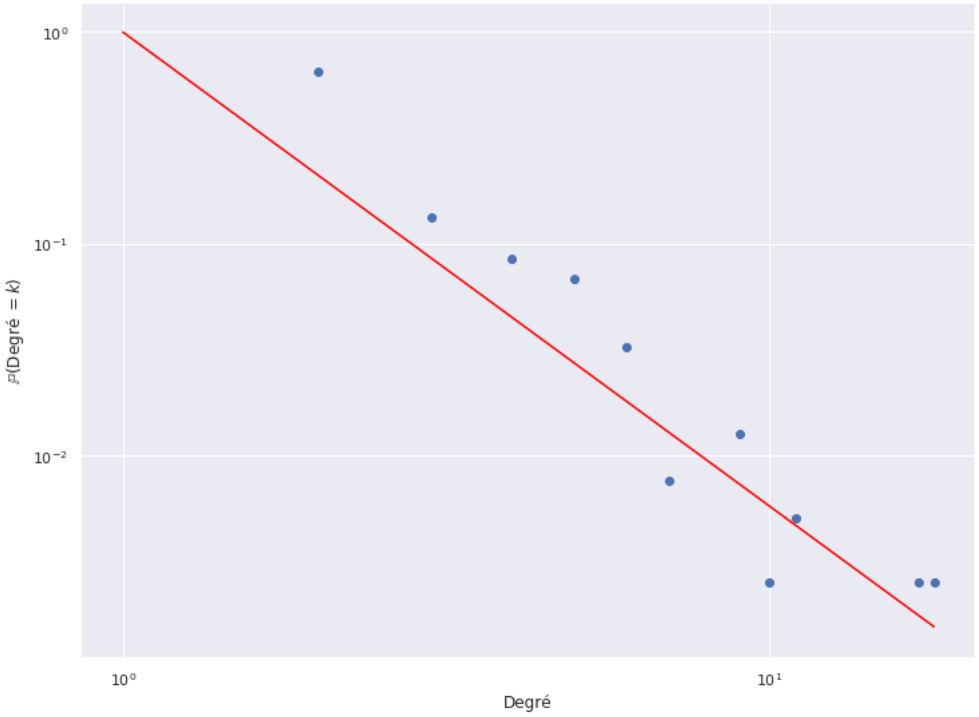

FIGURE 2.7 – Récupération du paramètre α

On obtient $\tilde{\alpha} = 2.44$ avec $R^2 = 0.9323$, notre régression est donc très bonne car elle explique plus de 93% de la variance de nos données et on suppose alors que la distribution du degré semble suivre une loi puissance discrète de paramètre $\alpha = 2.44$. Pour avoir une première impression on trace alors la distribution des degrés et on lui superpose la droite d'équation $y = \dfrac{1}{x^\alpha}$.

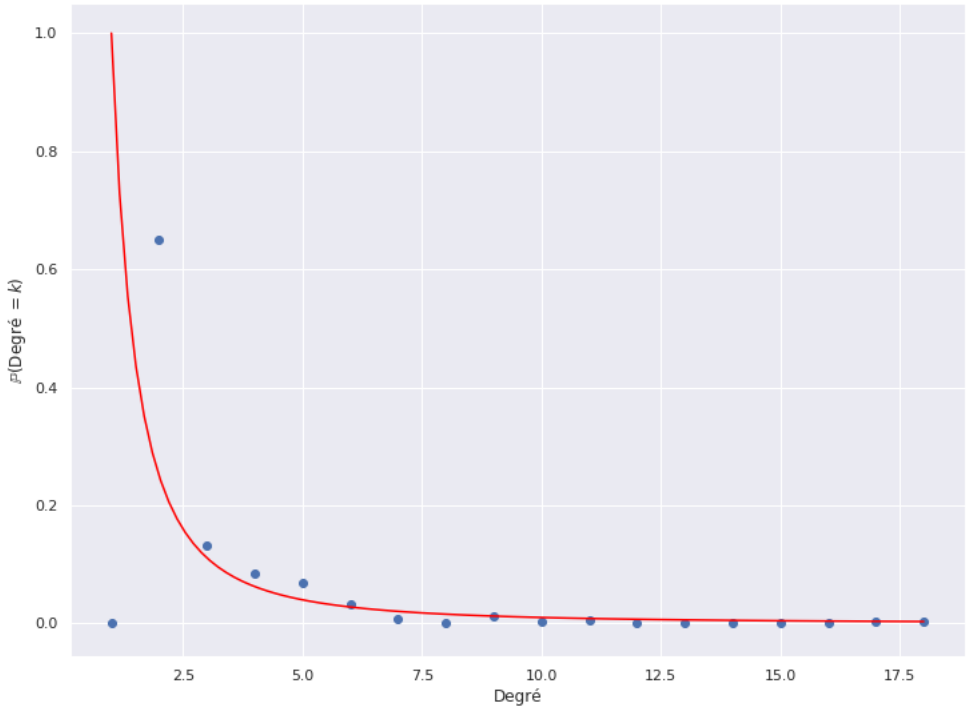

FIGURE 2.8 – Ajustement de la courbe de la loi puissance

La queue de la courbe semble bien s'ajuster au nuage de point, pour vérifier formellement notre impression on va alors appliquer le test d'adéquation du χ^2 à une loi.

Théorème 2.2.1 (Test du χ^2 pour l'adéquation à une loi $\mathbb{P}_0$)**.**
Le test du χ^2 consiste à découper l'espace des n observations en k classes, et à comparer les fréquences empiriques de chaque classe i : $\dfrac{n_i}{n}$ avec la probabilité théorique $\mathbb{P}_0$ donnée par $\mathcal{H}_0$, p_i. La statistique de test est :

$$\xi_n^{\chi^2} = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j} = \sum_{j=1}^{k} n \frac{(\frac{n_j}{n} - p_j)^2}{p_j}$$

Elle converge en loi vers un $\chi^2(k-1)$ sous $\mathcal{H}_0$ et vers $+\infty$ sous $\mathcal{H}_1$. La région de rejet est donc :

$$R_n = \left\{ \xi^{\chi^2} \geq q_{1-\alpha} \left(\chi^2(k-1) \right) \right\}$$

En testant l'adéquation avec le paramètre $\hat{\alpha}$, on obtient une p-valeur de 0.2937 et en faisant le test avec le paramètre $\tilde{\alpha}$ on obtient une p-valeur de 0.13. Cela indique qu'on ne rejette pas l'adéquation à la loi puissance pour aucun des deux paramètres. Cependant n'oublions pas que l'on a fixé arbitrairement le paramètre k_{min}, trouvons en alors une estimation.

Estimation du paramètre k_{min}

Pour estimer k_{min}, Clauset, Young et Gleditsch[7] proposent une méthode assez simple : on choisit une valeur de $\hat{k}_{min}$ qui permet le meilleur ajustement de la distribution observé au modèle de la loi puissance. En général on le choisit assez élevé pour avoir que $\hat{k}_{min} > k_{min}$ puis on diminue progressivement sa valeur pour avoir le meilleur ajustement possible. Afin de mesurer la qualité de l'ajustement on utilise la statistique de Kolmogorov-Smirnov :

$$D = \max_{k \geq k_{min}} |S(k) - P(k)|$$

Avec S la fontion de survie empirique de nos données et P celle de la loi puissance avec le meilleur paramètre α pour $k \geq k_{min}$. Notre estimateur $\hat{k}_{min}$ est alors la valeur qui minimise D.

Dans notre cas, on trouve $k_{min} = 1$ avec $D = 0.39$.

Adéquation à la famille de loi

Théorème 2.2.2 (Test du χ^2 pour l'adéquation à une famille de lois paramétriques $(\mathbb{P}_\theta)_{\theta \in \Theta}$)**.**
On suppose $\theta \in \Theta$ où Θ est une partie de $\mathbb{R}^d$. On remplace les p_j précédents par $pj(\hat{\theta})$ où $\hat{\theta}$ est l'estimateur du maximum de vraissemblance de θ. Ceci revient à tester l'adéquation à la loi la plus vraissemblable de la famille :

$$\xi_n^{\chi^2} = \sum_{j=1}^{k} \frac{(n_j - np_j(\hat{\theta}))^2}{np_j(\hat{\theta})} = \sum_{j=1}^{k} n \frac{(\frac{n_j}{n} - p_j(\hat{\theta}))^2}{p_j(\hat{\theta})}$$

Elle converge en loi vers un $\chi^2(k-d-1)$ sous $\mathcal{H}_0$ et vers $+\infty$ sous $\mathcal{H}_1$. La région de rejet est donc :

$$R_n = \left\{ \xi^{\chi^2} \geq q_{1-\alpha} \left(\chi^2(k-d-1) \right) \right\}$$

Dans notre cas, on considère comme unique paramètre de la loi puissance le coefficient α et on fixe $k_{min} = 1$. On obtient une p-valeur de 0.29 donc on ne rejette pas l'adéquation à une la famille de loi puissance.

Chapitre 3

Analyse de réseau

3.1 Théorie des réseaux

La *théorie des réseaux* est l'étude de graphes en tant que représentation d'une relation symétrique ou asymétrique entre des objets discrets. Elle s'inscrit naturellement dans la théorie des graphes : un réseau peut être défini comme un graphe où les nœuds (sommets) et/ou les arêtes (ou « arcs », lorsque le graphe est orienté) ont des attributs, comme une étiquette.

3.1.1 Mesures et propriétés

Les réseaux ont généralement des attributs qui peuvent être mesurés afin d'analyser leurs propriétés et caractéristiques. Le comportement de ces propriétés de réseau définit souvent des modèles de réseau et peut être utilisé pour analyser le contraste de certains modèles. Le lexique de la théorie des graphes contient de nombreuses définitions d'autres termes utilisés en science des réseaux.

Une propriété qui distingue les «vrais réseaux» des graphes aléatoires est la *transitivité* qu'on peut interpréter par l'expression «l'ami de mon ami est mon ami», on définit ainsi une mesure de ce phénomène appelé *coefficient de clustering*.

Définition 3.1.1.
Il existe deux définitions légèrement différentes du coefficient de clustering :

— *Le coefficient de clustering global qui est défini comme :*

$$C = \frac{6 \times \textit{nombre de triangles du graphe}}{\textit{nombre de chemins de longueur 2}}$$

où un triangle est un graphe possédant 3 sommets et 3 arêtes.

— *Le coefficient de clustering local qui est défini comme :*

$$C_i = \frac{\textit{nombre de triangles connectés au sommet } i}{\textit{nombre de triplés centrés en } i}$$

En prenant la moyenne de ces coefficients locaux, on obtient le coefficient local moyen :

$$\widetilde{C} = \frac{\sum C_i}{\textit{nombre de sommets du graphe}}$$

Dans la version locale les nœuds de petits degrés ont plus d'influence que ceux de grands degrés.

En d'autres termes, on étudie les liens au niveau des *triades* (relations entre trois sommets) et l'on vérifie que si $a \sim b$ et $b \sim c$ alors $a \sim c$.
Pour la formule du coefficient de clustering global : l'indicateur varie entre 0 et 1. Il vaut 0 dans le cas d'*arbres* (graphe non orienté, connexe, et sans cycle) par exemple et 1 quand chaque sommet appartient à une *clique* (ensemble de sommets deux-à-deux adjacents), elle s'interprète comme une probabilité.

3.1.2 Les modèles en théorie des réseaux

Les modèles en théorie des réseaux servent de fondement à la compréhension des interactions au sein de réseaux complexes empiriques. Divers modèles de génération de graphes aléatoires produisent des structures en réseaux qui peuvent être utilisées afin de les comparer aux réseaux complexes du monde réel.

Définition 3.1.2 (Modèle d'Erdős–Rényi)**.**
On fixe un entier $n \geq 1$ et un paramètre $p \in [0,1]$. Le graphe aléatoire d'Erdős–Rényi est le graphe $G = G(n,p)$ dont les sommets sont les entiers $i \in \{1,\ldots,n\}$ et tel que les variables aléatoires :

$$X_{ij} = \begin{cases} 1 \textit{ si } \{i,j\} \textit{ est une arête de } G \\ 0 \textit{ sinon} \end{cases}$$

sont indépendantes et suivent une même loi de Bernoulli de paramètre p. Autrement dit, chaque arête de G apparaît avec probabilité p indépendamment des autres arêtes.

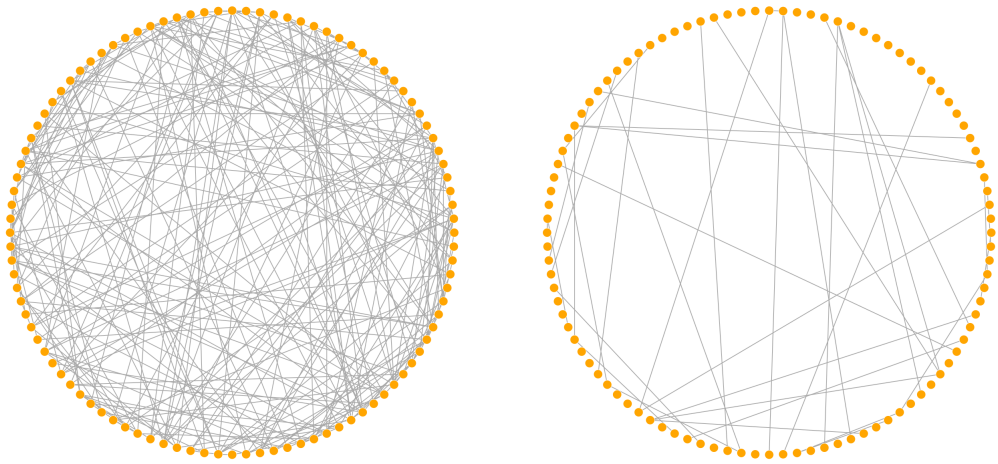

FIGURE 3.1 – Exemples d'Erdös-Rényi avec $n = 100$ pour $p = 0.05$ et $p = 0.001$

Définition 3.1.3 (Modèle de configuration)**.**
Le modèle de configuration est plus complexe que celui d'Erdös-Rényi. On commence en effet par choisir le degré de chaque sommet : on se donne $p = (p_k)_{k\in\mathbb{N}}$ *une distribution de probabilité sur* $\mathbb{N}$*, et pour chaque sommet, son degré est tiré au sort selon* p*, indépendamment de tous les autres. On peut représenter cela par des "demi-arêtes" sortant de chacun des sommets. Ces demi-arêtes sont ensuite reliées deux par deux, uniformément au hasard. S'il y en avait un nombre impair, la dernière est supprimée.*

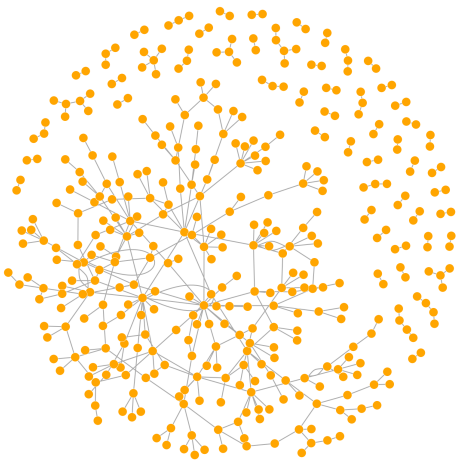

FIGURE 3.2 – Exemples d'un modèle de configuration avec p : la distribution de degré de notre réseau

3.2 Analyse du graphe en tant que réseau

3.2.1 Transitivité du réseau

Les points d'articulation que l'on a étudié plus haut peuvent être vu comme étant les personnes dont le rôle est central dans la transmission de maladies infectieuses. En effet, si un individus est porteur d'une maladie telle que le VIH, l'hépatite C ou autre, la maladie se propagera de proche en proche dans le réseau social via ces points d'articulations.
Pour notre graphe complet on trouve : $C = 0.147$ et pour les deux plus grandes composantes on trouve $C = 0.279$ pour la plus grande et $C = 0.034$ pour la plus petite. Pour savoir si ces coefficients sont assez petit pour démontrer l'indépendance ou non de former un lien entre deux sommets nous allons générer 10000 graphes suivant le modèle de configuration avec pour entrée la distribution de degré de notre graphe et calculer à chaque itération le coefficient de clustering associé. On obtient alors le résultat suivant :

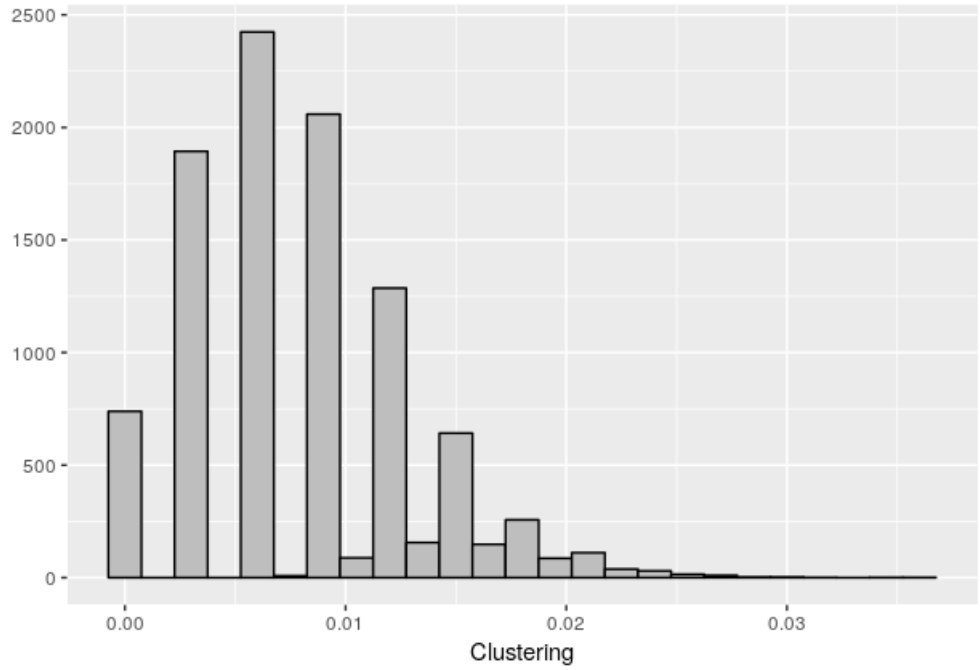

FIGURE 3.3 – Distribution du coefficient de clustering dans nos simulations

Et on constate que plus 99% des valeurs obtenues sont inférieures à 0.034, on peut donc conclure que sachant que dans le modèle de configuration on a indépendance entre les événements $v_i \sim v_j$, $v_j \sim v_k$ et $v_i \sim v_k$, on a pas indépendance dans notre graphe au vu des coefficients de clustering que l'on a obtenue pour ce dernier.

3.2.2 Influence des covariables dans la structure du réseau

Dans plusieurs réseaux, on peut caractériser les sommets par différentes variables :

— L'âge, la race, la classe sociale d'une personne.

— Le type d'alimentation dans la chaîne alimentaire.

— La taille d'un serveur dans le réseau Internet.

Dans presque tous les réseaux sociaux, on observe une influence de ces types dans la structure du réseau. Dans les réseaux technologiques, biologiques et d'information, on observe plutôt le contraire.

On peut aussi mettre en évidence différentes caractéristiques des individus dans le graphe comme par exemple l'âge ou le sexe afin de distinguer certaines tendances sociologiques :

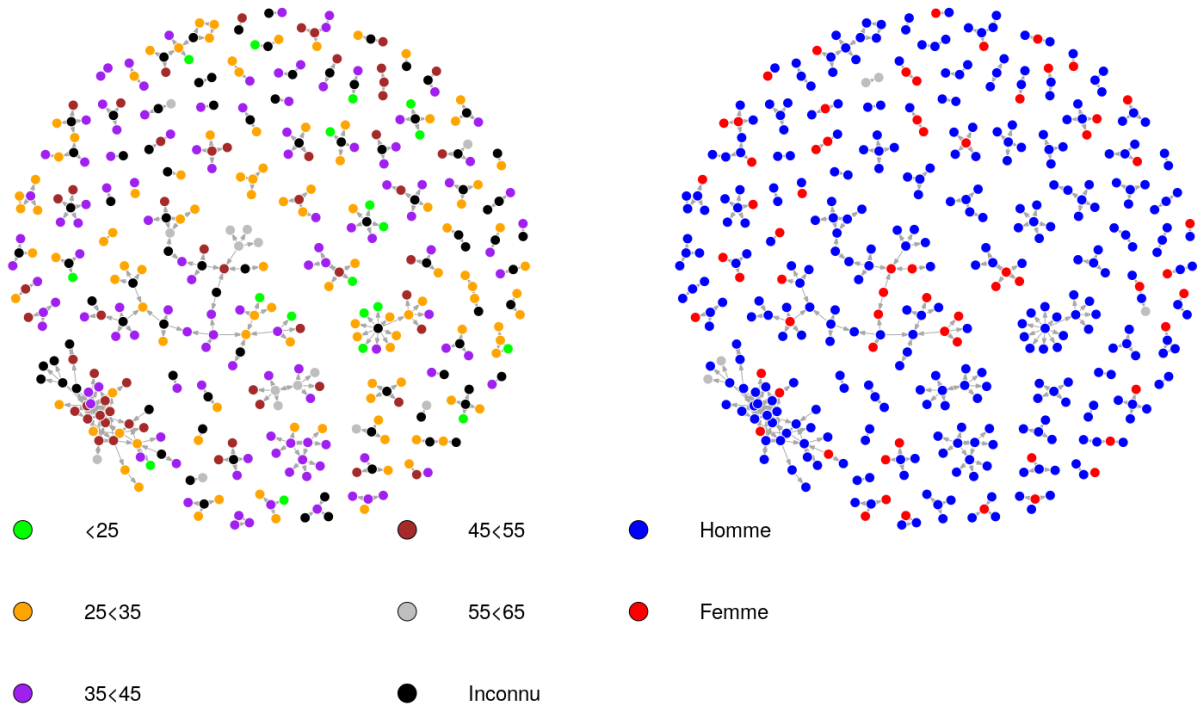

FIGURE 3.4 – Visualisation selon l'âge et selon le sexe

On constate directement que les hommes sont très majoritaire dans notre échantillon en observant le graphe de droite, pour celui de gauche la répartition de l'âge au sein de notre graphe est assez homogène bien qu'on remarque tout de même certaines affinités.

Regardons alors si les personnes du même sexe ont tendance à plus se connecter entre eux plutôt qu'avec les autres :

	Homme	Femme	Non renseigné
Homme	277	49	3
Femme	39	11	0
Non renseigné	0	0	1

En faisant un test d'indépendance du χ^2, on obtient alors une p-valeur de $6.4e^{-20}$ et on rejette alors l'hypothèse d'indépendance.

Regardons si la répartition des sexes diffère selon l'âge :

Âge \ Sexe	F	H
≤ 25	6	13
26-35	22	85
36-45	15	102
46-55	8	47
56-65	0	15

En faisant un test d'indépendance du χ^2, on obtient une p-valeur inférieure à 0.06736 et on ne rejette alors pas l'hypothèse d'indépendance.

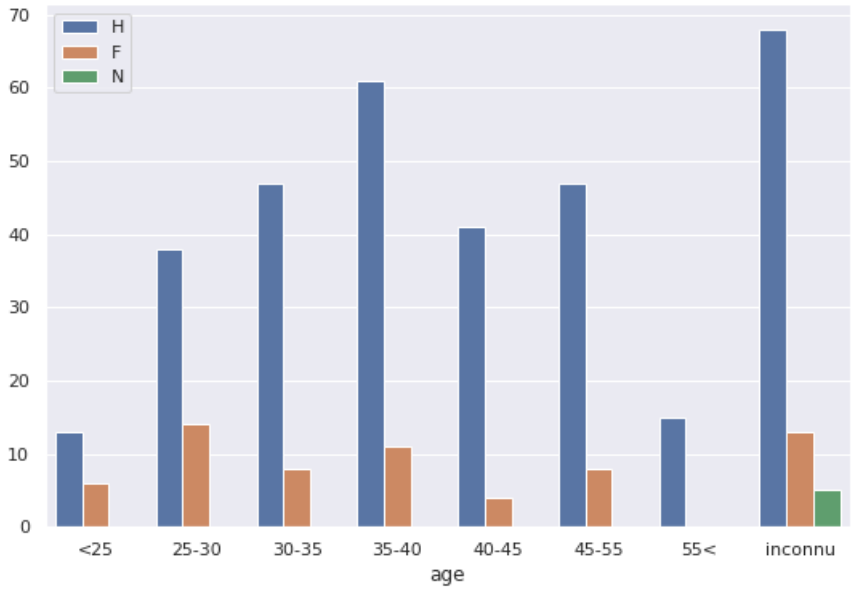

FIGURE 3.5 – Nombre d'hommes et de femmes selon différentes tranches d'âge

Toute proportion gardée, la répartition de l'âge est donc bien la même pour les deux sexes.

3.3 Adéquation à un modèle Erdös-Rényi

Soient $\{1, \cdots, n\}$ un ensemble de sommets et notons $v_i \sim v_j$ si il existe un arc entre v_i et v_j.
Soient (Y_{ij}) une suite de variables aléatoires qu'on suppose i.i.d. avec $i, j \in 1, \ldots, n$ et $i \neq j$ telles que

$$\begin{cases} Y_{ij} = 1 \text{ si } v_i \sim v_j \\ Y_{ij} = 0 \text{ sinon} \end{cases}$$

On a alors $Y_{ij} \sim \mathcal{B}(p)$ pour tout i, j avec p inconnu qu'on cherche à estimer.

Proposition 3.3.1. *Un estimateur sans biais et fortement consistant de p est :*

$$\hat{p} = \frac{2}{n(n-1)} \sum_{1 \leq i \neq j \leq n} Y_{ij}$$

Démonstration.
On a

$$\begin{aligned} \mathbb{E}[\hat{p}] &= \mathbb{E}\left[\frac{2}{n(n-1)} \sum_{1 \leq i \neq j \leq n} Y_{ij}\right] \\ &= \sum_{1 \leq i \neq j \leq n} \mathbb{E}\left[Y_{ij}\right] \\ &= p \end{aligned}$$

l'estimateur est donc sans biais et de plus

$$\begin{aligned} \mathbb{V}ar[\hat{p}] &= \mathbb{V}ar\left[\frac{2}{n(n-1)} \sum_{1 \leq i \neq j \leq n} Y_{ij}\right] \\ &= \frac{4}{n^2(n-1)^2} \sum_{1 \leq i \neq j \leq n} \mathbb{V}ar\left[Y_{ij}\right] \\ &= \frac{2}{n(n-1)} p(1-p) \underset{n \to \infty}{\longrightarrow} 0 \end{aligned}$$

D'après la loi forte des grands nombres, $\hat{p}$ converge presque-sûrement vers p. L'estimateur est donc fortement consistant. □

Dans notre cas, on trouve alors : $\hat{p}(\omega) \simeq 0.002$

Essayons de générer un graphe à 399 sommets suivant ce modèle :

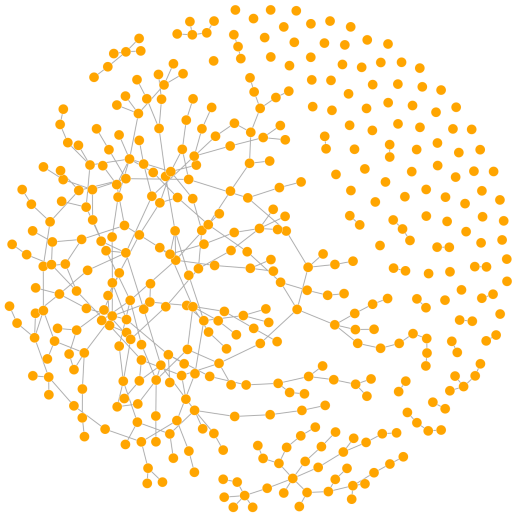

FIGURE 3.6 – Simulation d'un graphe selon le modèle Erdös-Rényi

On constate que le graphe ainsi générer ne ressemble pas vraiment au nôtre. Cependant, on peut se demander si notre réseau réel n'a pas véritablement cette structure. En effet, il est possible que beaucoup de nos groupes(composantes connexes) qui ne sont pas connectés dans l'étude le sont en réalité via un lien caché.

Afin de vérifier la similarité du modèle ainsi obtenu, simulons un nombre important (ici 10000) de graphes aléatoires et comparons le avec notre réseau selon divers critères de similarité.

Si on s'intéresse au coefficient de clustering, on obtient :

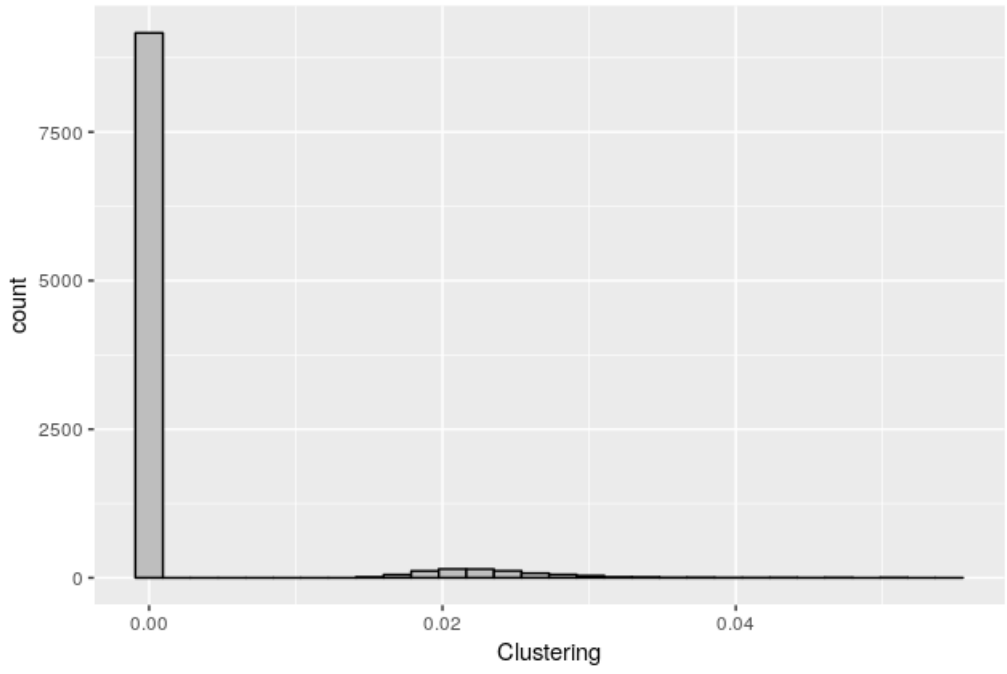

FIGURE 3.7 – Coefficient de clustering pour les simulations du modèle Erdös-Rényi

On trouve alors 9159 graphes simulés ayant un coefficient de clustering nul, cela s'explique simplement du fait qu'il y a indépendance dans la distribution des liens entre les sommets. Si on omet les graphes ayant un coefficient de clustering nul, on a :

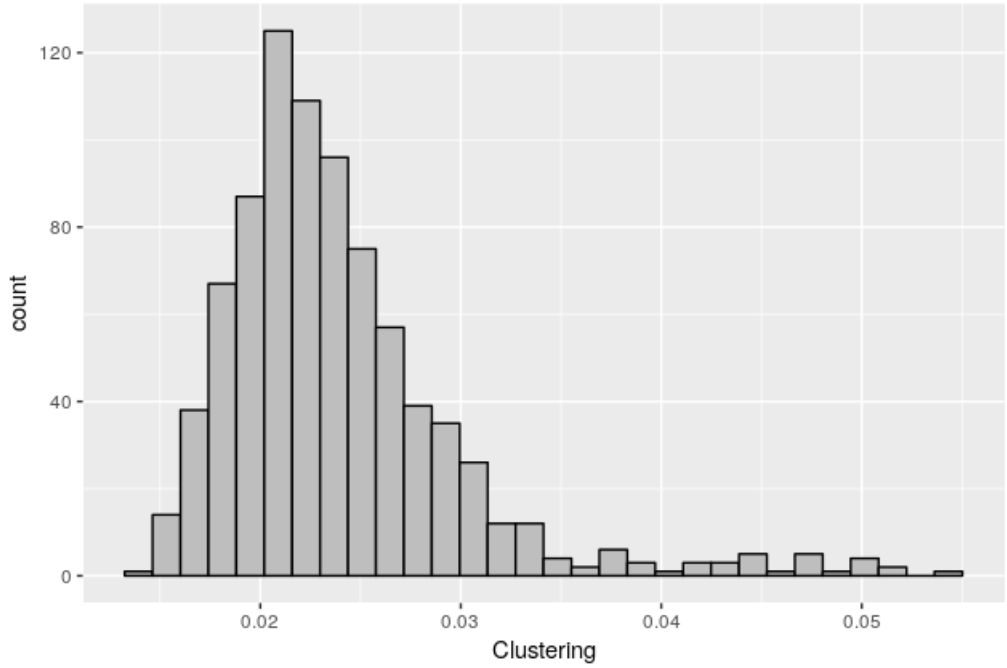

FIGURE 3.8 – Coefficient de clustering pour les simulations du modèle Erdös-Rényi

On retrouve alors le fait que le coefficient de clustering de notre réseau que l'on a calculé plus haut est suffisamment significatif pour indiquer la dépendance entre les événements $v_i \sim v_j$, $v_j \sim v_k$ et $v_i \sim v_k$ pour l'ensemble des sommets de notre réseau.

Si on s'intéresse maintenant au nombre de composantes connexes, on obtient :

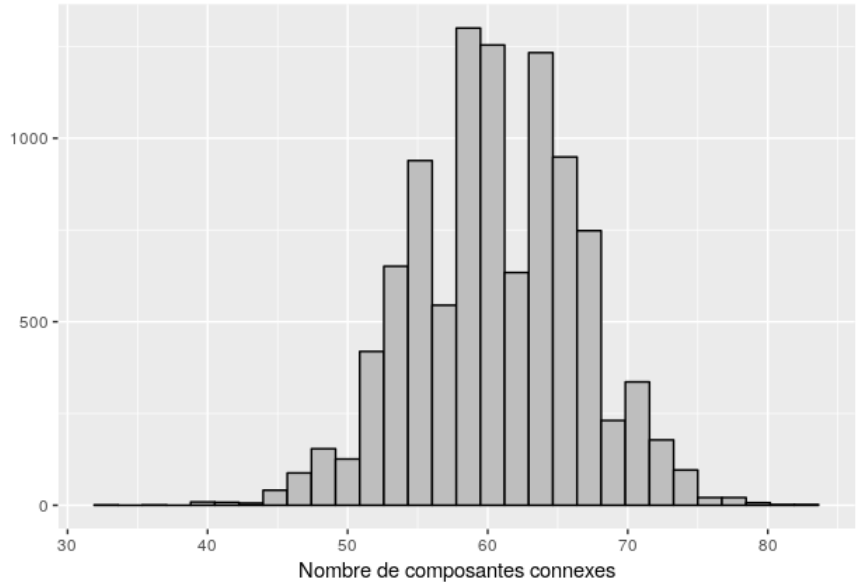

FIGURE 3.9 – Nombre de composantes connexes pour les simulations du modèle Erdös-Rényi

On trouve ici que 99% des graphes simulés ont un nombre de composantes connexes inférieur à 82 (nombre de composantes connexes de notre réseau).

Intéressons nous enfin à la longueur géodésique de nos simulations :

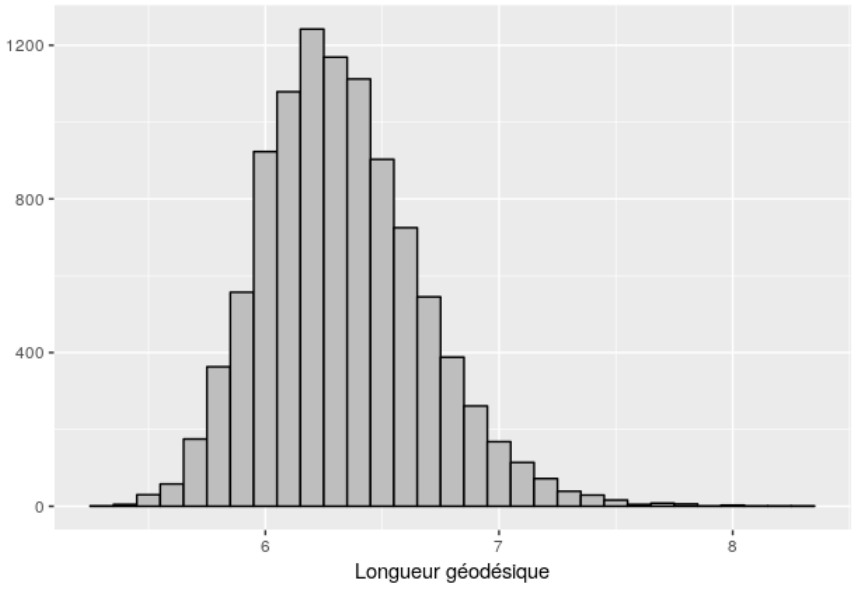

FIGURE 3.10 – Longueur géodésique pour les simulations du modèle Erdös-Rényi

On trouve alors que 95% des observations se situent entre 5.74 et 7.08, la longueur géodésique de notre réseau valant 4.4 on conclut une fois encore que ce modèle diffère du notre selon ce critère.

En conclusion, on constate que :

— La distribution des degrés est très différente d'une loi de puissance.

— Le coefficient de clustering est proche de $C = p$.

— Ils ne présentent aucun attachement préférentiel.

Bref, ils ne sont pas particulièrement représentatifs des réseaux réels.

Chapitre 4

Clustering spectral

Le but de cette section est de déterminer si il y a des communautés cachés dans notre graphe, c'est-à-dire déterminer si certains individus ont tendance à communiquer entre eux en formant des groupes. Il s'agit donc ici de créer une partition de notre ensemble d'individus et donc de réaliser un clustering.

4.1 Théorie spectrale des graphes

La théorie spectrale des graphes s'intéresse aux rapports entre les spectres des différentes matrices que l'on peut associer à un graphe et ses propriétés.

4.1.1 Graphes et algèbre linéaire

Tout graphe $\mathcal{G} = (V, E)$ peut être représenté sous forme de matrice. Les relations entre arêtes et sommets, appelées les relations d'incidence, sont toutes représentées par la matrice d'incidence du graphe. Les relations d'adjacences (si deux sommets sont reliés par une arête ils sont adjacents) sont représentés par sa matrice d'adjacence. Elle est définit par :

$$a_{ij} = \begin{cases} 1 \text{ si } i \sim j \\ 0 \text{ sinon} \end{cases}$$

La matrice des degrés D est une matrice diagonale où les éléments D_{ii} correspondent au nombre de liens du sommet i, c'est-à-dire à son degré. En utilisant cette matrice et la précédente, on peut également définir la matrice laplacienne non normalisée $L = D - A$.

On obtient sa forme normalisée L' par $L' = D^{-1/2}LD^{-1/2} = I - D^{-1/2}AD^{-1/2}$, où I est la matrice

identité. On obtient aussi L' directement par chacun de ses éléments :

$$\ell_{i,j} := \begin{cases} 1 & \text{si } i = j \text{ et } \deg(v_i) \neq 0 \\ -\frac{1}{\sqrt{\deg(v_i)\deg(v_j)}} & \text{si } i \neq j \text{ et } v_i \text{ est adjacent à } v_j \\ 0 & \text{sinon.} \end{cases}$$

On a les propriétés suivantes pour le laplacien :

Propriétés 4.1.1.

1. *Soit e, le vecteur rempli de 1, e est un vecteur propre associé à 0 pour L*
2. *Toutes les valeurs propres du laplacien sont réels et positives ou nulles.*
3. *Le nombre de composantes connexes d'un graphe non orienté et non pondéré est donné par la valeur de la multiplicité k de 0 dans L.*

Démonstration.

1. On a $Le = De - Ae$ or la somme des colonnes de A correspond au degré du sommet de la ligne correspondante. On a donc $Le = 0$.
2. Comme notre graphe est non orienté la matrice d'adjacence est symétrique,on a la matrice L qui est symétrique réel et donc le théorème spectral nous assure que ses valeurs propres sont réel.
   D'autre part, montrons qu'elle est définie positive ce qui nous assure la deuxième partie de la proposition :
   Soit x un vecteur de R^n, on a :
$$\begin{aligned} {}^t xLx &= \sum_{i=1}^{n} d(i)x_i^2 - \sum_{i,j=1}^{n} \mathbb{1}_{i\sim j} x_i xj \\ &= \frac{1}{2}\left(\sum_{i=1}^{n} d(i)x_i^2 - 2 * \sum_{i,j=1}^{n} \mathbb{1}_{i\sim j} x_i xj + \sum_{j=1}^{n} d(i)x_j^2\right) \\ &= \frac{1}{2}\sum_{i,j=1}^{n} \mathbb{1}_{i\sim j}(x_i - x_j)^2 \geq 0 \end{aligned}$$
3. Supposons dans un premier temps que l'on a $k = 1$. C'est-à-dire qu'on aurait une seule composante connexe soit donc tout le graphe.
   Soit v la valeur propre associé à la valeur 0, on a d'après le calcul précédent que
$$ {}^t vLv = \frac{1}{2}\sum_{i,j=1}^{n} \mathbb{1}_{i\sim j}(v_i - v_j)^2 = 0$$
   et donc comme tous les termes sont positifs cela signifie que $v_i = v_j$ c'est-à-dire que le vecteur propre associé à 0 est e. Ce vecteur est unique, donc la multiplicité de 0 est bien 1 .

Supposons que $k \geq 2$, on a donc k composantes connexe.

Sans perte de généralité, on suppose que les sommets sont ordonnés de manière à ce que la matrice d'adjacence soit diagonale par blocs. La matrice L est alors aussi diagonale par blocs et L s'écrit :

$$\left[\begin{array}{c|c|c} L_1 & & 0 \\ \hline & \ddots & \\ \hline 0 & & L_k \end{array}\right]$$

Les L_i sont aussi les laplaciens du graphe de la i-ème composante connexe donc chaque L_i étant un graphe connexe admet d'après le cas précèdent 0 comme valeur propre et le spectre de L étant la réunion des spectre des L_i on a 0 qui est bien de multiplicité k ce qui conclut la preuve.

□

4.2 Analyse spectrale

On introduit une matrice de dissimilarité A, où les coefficients A_{ij} représentent la distance/dissimilarité entre les points i et j c'est-à-dire ici entre les sommets correspondants. L'approche général du clustering spectral est d'utiliser une méthode de clustering usuel (k-means, regroupement hiérarchique..) sur des vecteurs propres de la matrice du laplacien L de A.

Le but d'un clustering est de séparer des points en différents groupes selon leur similarités. On veut pouvoir trouver une partition de notre graphe de telle sorte que les sommets dans les clusters soit «similaires» et que les sommets dans des clusters différents soit «dissimilaires».

Soit $A, B \subset V$ on définit $Cut(A,B) = \sum\limits_{i \in A, j \in B} \mathbb{1}_{i \sim j}$.
Le but du clustering spectral consiste à choisir la partition $\{A_1, \cdots, A_k\}$ qui minimise la valeur de $Cut(A_1, \cdots, A_k) = \sum\limits_{i=1}^{k} Cut(A_i, \overline{A_i})$.

Considérons le cas où on veut séparer notre échantillon en 2 groupes :
On peut réécrire le problème de minimisation par le problème équivalent :

$$\min_{A \subset V} {}^t v L v \text{ où } v \text{ vérifie : } \begin{cases} v_i = \sqrt{\overline{A}/A} \text{ si } v_i \in A \\ v_i = -\sqrt{(A/\overline{A})} \text{ si } v_i \in \overline{A} \\ v \perp e \\ \|v\| = \sqrt{n} \end{cases}$$

On considère alors le problème relaxé suivant :

$$\min_{v} {}^t v L v \text{ où } v \text{ vérifie : } \begin{cases} v \perp e \\ \|v\| = \sqrt{n} \end{cases}$$

On sait que e est un vecteur propre de L, donc en reconnaissant une forme quadratique et par les propriétés du Laplacien, le vecteur v qui résout ce problème est le vecteur propre lié à la deuxième valeur propre la plus petite λ_2 du laplacien normalisé.
Ensuite pour partitionner, on transforme ce vecteur en indicatrice selon le signe des coordonnées du deuxième vecteur propre. On peut aussi appliquer un algorithme de clustering type k-means pour séparer nos variables en 2 groupes.

Regardons maintenant le cas ou $k > 2$.
Le problème devient matriciel et dans un premier temps on est amené à résoudre le problème équivalent :

$$\min_{A_1,A_2,.....,A_k} Tr({}^tHLH) \text{ où } H \text{ vérifie : } \begin{cases} H_{ij} = \dfrac{1}{\sqrt{\#A_i}} \text{ si } i \in A_j \\ 0 \text{ sinon} \\ {}^tHH = I \end{cases}$$

On considère alors le problème relaxé suivant :

$$\min_H Tr({}^tHLH) \text{ avec } {}^tHH = I$$

dont la solution est donnée par le choix de H comme la matrice qui contient les k premiers vecteurs propres de L comme colonnes et comme précédemment on réalise la partition en utilisant un k-means sur nos vecteurs propres.

On pourrait se questionner sur la validité du problème relaxé. En effet,il n'y a pas de garantie sur la qualité de la solution du problème relaxé par rapport à la solution réelle. On peut prendre l'exemple développé dans [8] avec les graphes dit de cafards dans lequel la séparation naturelle en 2 groupes, conduit avec le problème relaxé à un ratio $\dfrac{k}{2}$ fois plus important, où k désigne le nombre d'arètes coupées.

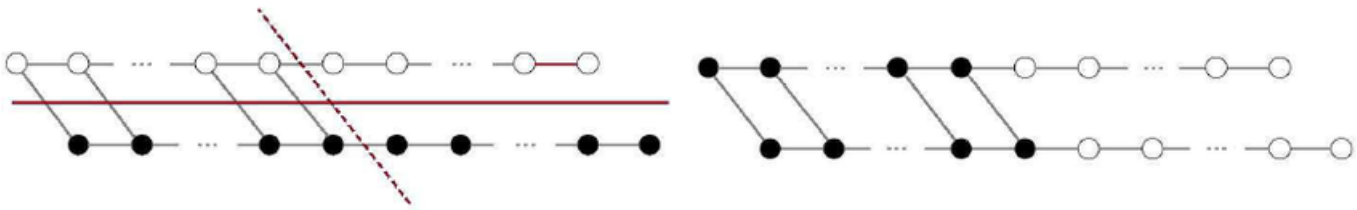

FIGURE 4.1 – Exemple des cafards

Application à notre graphe

Ici on a déjà des clusters qui correspondent à chaque composante connexe, on va donc se concentrer sur la recherche de réseau caché à l'intérieur de nos composantes principales. On s'occupera du cas de la plus grande composante qui contient 38 individus.

Le nombre de clusters à l'intérieur des grandes composantes étant ici inconnu, on va utiliser l'approche clustering top down qui consiste à recouper les sous graphes obtenus en coupant en deux clusters à

chaque fois par clustering spectral. On est donc ramené au cas ou $k = 2$ à chaque itération et il nous faut donc regarde λ_2 la seconde plus petite valeur propre de notre graphe/sous-graphe.

On peut quand même déterminer de manière heuristique de nombre de clusters en observant des trous entre nos valeurs propres :

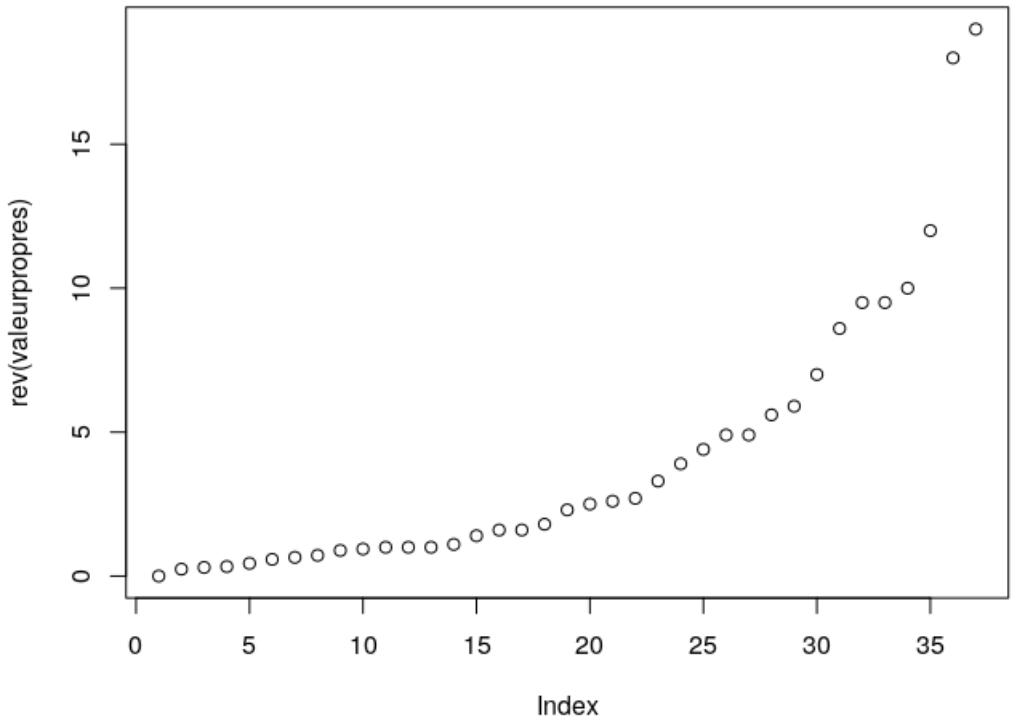

FIGURE 4.2 – Valeurs propres

Ici il semble donc y avoir un trou donc possiblement 2 clusters. On retrouve aussi le fait que 0 est valeur propre simple de notre composante puisque elle est connexe.

Effectuons la première itération de l'approche top down :

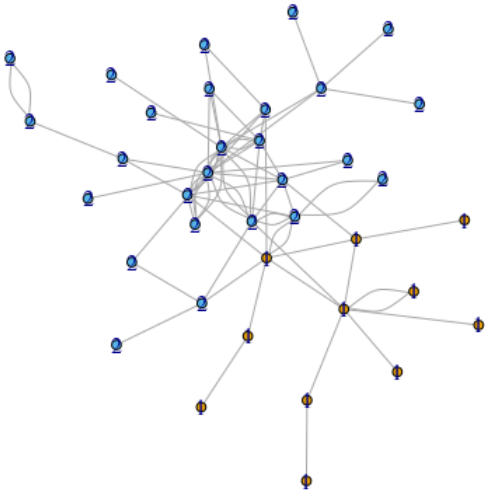

FIGURE 4.3 – Communautés cachés au sein de la composante principale

On voit que dans le deuxième cluster on a 11 personnes donc on va juste appliquer le clustering spectral sur la deuxième cluster.

On obtient alors :

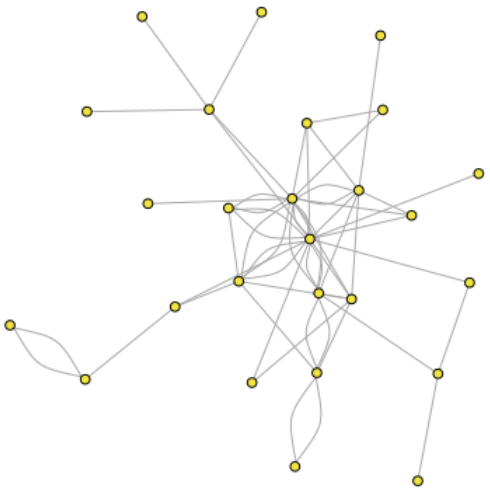

FIGURE 4.4 – Isolation d'une de communautés cachés au sein de la composante principale

et on voit qu'on n'obtient pas de nouveaux clusters donc l'hypothèse d'avoir 2 clusters semble vraisemblable.

Regardons le degré moyen par cluster : on observe un degré moyen de 2.81 pour le cluster en jaune 2.81 et de 4.65 pour le cluster en bleu donc les individus sont plus connectés à l'intérieur du cluster en bleu.

On peut donc supposer qu'à l'intérieur même des composantes principales, on a un cercle qui se connaît et qui interagit beaucoup avec ses membres et de l'autre coté un groupe d'utilisateurs plutôt isolés.

Chapitre 5

Inférence

Le but de cette partie est d'estimer l'effectif total de la population d'usager de drogue injectable à Paris.

5.1 Hypothèses et modélisation du problème

Notations :
U est l'ensemble de tous les individus.
S est l'ensemble de tout les individus dans l'échantillon.
A et B sont des ensembles disjoints des individus.
N_X est le nombre d'éléments dans l'ensemble X.
En particulier on notera n le nombre d'individus dans l'échantillon et N le nombre total d'individus de la population.
d_i est le degré de l'individu i
d_X est le degré moyen des individus de l'ensemble X.

Dans la suite on supposera que chaque personne interrogée dans notre étude désigne un seul voisin et que l'échantillon est petit de telle sorte que l'on puisse utiliser l'hypothèse de remise.
Basons nous d'abord sur les hypothèses suivantes issues de Volz-Heckathorn[9], pour construire notre modèle :

— Les sondés donnent précisément leurs nombre de liens

— On considère qu'il y a réciprocité entre les sondés et leurs cibles. Cela se traduit par la non orientation de notre graphe.

— Le recrutement peut être modélisé comme une chaîne de Markov où $X_n \in [1, \ldots, 399]$ désigne le numéro du dernier individu recruté au temps n : il s'agit donc d'une marche aléatoire sur notre graphe. Elle sera supposée irréductible ce qui signifie que l'on peut rejoindre n'importe quels

individus par un chemin et que chaque état a un temps de retour fini, i.e la chaîne est apériodique.
D'après Volz-Heckathorn l'hypothèse d'irréductibilité a du sens car la plus grande composante connexe contient quasiment toute l'information, ici c'est un peu plus compliqué car on a de nombreuses composantes connexes. De plus, on ne distribue pas seulement un seul coupon mais au plus 3.
À noter qu'il ne s'agit pas *stricto sensu* d'une marche aléatoire puisque la connaissance des voisins autour d'un sommet finit par orienter la marche. On pourrait plutôt dire que c'est une chaîne de Markov conditionnellement à notre graphe.

5.1.1 Autour de notre chaîne de Markov

Si au temps t, on est dans notre chaîne de Markov à l'état i, au temps $t+1$ on a une probabilité $\frac{1}{d(i)}$ d'arriver à l'état j, ce qui correspond bien à choisir l'individu j parmi les $d(i)$ voisins de i de manière uniforme. La matrice de transition de notre chaîne de Markov est alors $P = (P_{i,j}) = \frac{\mathbb{1}_{i\sim j}}{d(i)}$ où $i \sim j$ si i est voisin avec j.

Proposition 5.1.1.
Sous les hypothèses que la chaîne de Markov définie précédemment soit irréductible, apériodique et récurrente positive alors le vecteur π défini pour $i \in S$ par $\pi(i) = \frac{d_i}{\sum_j d_j}$ est l'unique mesure de probabilité de la chaîne.

Démonstration.
En effet on a $(\pi P)_i = \sum_{j=1}^{n} \pi(j)P(i,j) = \sum_{j=1}^{n} \frac{d_j}{\sum_k d_k} \times \frac{\mathbb{1}_{i\sim j}}{d_j} = \sum_{j=1}^{n} \frac{\mathbb{1}_{i\sim j}}{\sum_k d_k}$ et on voit qu'il reste dans la somme des indicatrices que i et j soient voisins c'est-à-dire que $d(i)$ normalisé par la somme des degré ce qui est bien notre vecteur π.
L'unicité provient des hypothèses faites sur notre chaîne de Markov. □

D'autre part les hypothèses faites sur notre chaîne de Markov précédemment nous permettent de déduire l'existence d'une probabilité stationnaire d'après de le théorème suivant :

Théorème 5.1.1. *Si la chaîne de Markov est irréductible, apériodique et récurrente positive alors la loi de probabilité de la chaîne de Markov tend vers une unique mesure de probabilité*

5.2 Estimateurs de la population

Le but de cette section est de retrouver le nombre total d'individus de notre population étudié ainsi que les proportions de certains types.

Montrons comment on peut estimer N, la taille de la population totale à partir de la mesure de probabilité π définie par la chaîne de Markov précédente :

Définition 5.2.1.
La probabilité de sélection de l'individu i est égale à $p_i = \frac{d_i}{Nd_U}$.Cette probabilité de sélection correspond aussi à choisir l'individu i donc correspond à π_i.

Définition 5.2.2.
L'estimateur du degré moyen d_U des individus donnés dans le graphe par la formule $\widehat{d_U} = \frac{n}{\sum\limits_{i\in S}\frac{1}{d_i}}$.

La probabilité d'inclusion peut alors être estimée par $\widehat{p_i} = \frac{d_i}{N\widehat{d_U}}$ et N est alors donnée par la formule suivante : $N = \frac{d_i}{\pi_i\widehat{d_U}}$ Si on s'intéresse maintenant à des sous-populations (nombre de femmes, nombre par tranche d'âges, etc..) de notre échantillon alors on a les estimateurs suivants qui ne reposent pas sur cette mesure stationnaire :

Notons y_k la valeur prise par l'individu k, y peut être par exemple l'âge ou une variable dichotomique du genre homme ou femme, T_y représente le total de y dans la population.

Définition 5.2.3.
L'estimateur du nombre total y dans la population T_y est donné par $\widehat{T_y} = \frac{1}{n}\sum\limits_{i\in S}\frac{y_i}{p_i}$ qui est aussi égal à $\frac{N}{\sum\limits_{i\in S} d_i^{-1}}\sum\limits_{i\in S} d_i^{-1}y_i$

Proposition 5.2.1. *L'estimateur $\widehat{T_y}$ est sans biais.*

Démonstration.
On a $\sum\limits_{i\in S}\frac{y_i}{p_i}$ qui peut être réécrit $\sum\limits_{i=1}^{n}\frac{y_i}{p_i\mathbb{1}_{i\in S}}$ et donc comme $\mathbb{E}(\mathbb{1}_{i\in S}) = \mathbb{P}(i \in S) = p_i$ par définition de la probabilité de sélection,en utilisant la linéarité de la somme on trouve $\mathbb{E}(\widehat{T_y}) = \sum\limits_{i=1}^{N}\frac{y_i\mathbb{E}(\mathbb{1}_{i\in S})}{p_i} = \sum\limits_{i=1}^{n}\frac{y_ip_i}{p_i} = \sum\limits_{i=1}^{N} y_i = T_y$ □

Cependant de manière générale le N n'est pas connu (c'est ce que l'on veut estimer) de fait on privilégiera l'estimateur suivant :

Définition 5.2.4.
L'estimateur de la proportion de la population de type A, P_A, est donné par

$$\widehat{P_A} = \frac{\sum_{i \in S \bigcap A} d_i^{-1}}{\sum_{i \in S} d_i^{-1}}$$

.

Adaptons maintenant ces estimateurs à notre cas : dans un premier temps, vu le nombre important de composantes connexes présentes dans notre graphe, il est assez dangereux de supposer que notre chaîne de Markov est irréductible. On va donc utiliser une méthode qu'on peut rencontrer dans l'algorithme PageRank[10] de Google qui consiste à rééchantillonner notre chaîne de Markov de matrice P. Dans un second temps, on cherchera à comprendre comment passer à une distribution à plusieurs coupons.

5.2.1 Correction de l'irréductibilité

On introduit un facteur $\alpha \in [0, 1]$ appelé amortissement de sorte que la chaîne de Markov est soit parcouru avec une probabilité α, soit on choisit un individu au hasard uniformément avec une probabilité $1 - \alpha$. On obtient de fait une nouvelle chaîne de Markov associée à la matrice de transition P' qui s'écrit $P' = \alpha P + (1 - \alpha)\frac{U}{n}$ où U est la matrice de $M_n(\mathbb{R})$ remplit de 1.

Proposition 5.2.2. *La chaîne de Markov associée à cette matrice de transition est irréductible,apériodique et récurrente positive.*

Démonstration.
La matrice de transition P' a toutes ses entrées strictement positives, on a donc l'existence d'un chemin quelque soit les individus qu'on veut rejoindre. Elle est donc irréductible et pour la même raison, elle est apériodique. De plus, notre chaîne de Markov est à états finis, l'irréductibilité fourni alors aussi le fait qu'elle est récurrente positive. □

D'après le théorème de la partie précédente on a donc existence d'une mesure de probabilité pour la chaîne de Markov P' qu'on note π

Proposition 5.2.3. *La mesure stationnaire π associé à la matrice de transition P' vérifie $\pi = \alpha\pi P + (1 - \alpha)\frac{u}{n}$ ou u est le vecteur colonne rempli de 1.*

Démonstration.
L'égalité $\pi P' = \pi$ donne $\pi = \alpha\pi P + (1-\alpha)\pi\dfrac{U}{n}$ or on a $\pi_i U = \sum\limits_{j=1}^{n} \pi_j U_{ij} = \sum\limits_{j=1}^{n} \pi_j$ et π étant une mesure de probabilité, $\sum\limits_{j=1}^{n} \pi_j = 1$ d'où le résultat. □

On obtient alors π soit :

— en résolvant le système $\pi(I-\alpha P) = (1-\alpha)\dfrac{u}{n}$, on peut utiliser une méthode de Gauss ou multiplier par l'inverse à droite selon la stabilité de l'opération.

— en itérant plusieurs fois de telle sorte que la suite définit par $\pi_{k+1} = P'\pi_k$ en partant d'une distribution π_0 arbitraire converge vers π

Vu la faible taille de notre système, on utilisera l'inverse à supposer que $I - \alpha P$ est inversible.
On supposera que cette modification n'entraîne pas de changement pour la forme de notre estimateur cependant il faut de ce fait vérifier que notre vecteur propre a toutes ses composantes positives car la probabilité d'inclusion est toujours strictement positive dans notre cas (chaque individu a au moins un voisin donc $d_i \geq 1$). On utilise alors le théorème suivant :

Théorème 5.2.1 (Perron-Froebnius)**.**
Soit A une matrice de $M_n(\mathbb{R})$, irréductible alors : le rayon spectral ρ de A est une valeur propre simple de A et le sous espace vectoriel est engendrée par un vecteur strictement positif de norme 1.

Montrons que l'on peut appliquer ce théorème à notre matrice de transition P' pour obtenir une probabilité π qui est bien chargée en tout ses points.

Proposition 5.2.4. *1 est valeur propre de P'*

Démonstration.
En multipliant la matrice P' par notre vecteur e, on retrouve d'une part que $Pe = e$ car on a d_i fois $\dfrac{1}{d_i}$ sur une ligne. D'autre part, le produit matrice vecteur Ue est égale à ne donc on a finalement P' qui est une combinaison convexe de deux matrices stochastiques et finalement P' admet 1 comme valeur propre. □

Proposition 5.2.5. *Toutes les valeurs propres de P' sont $\leq$ 1*

Démonstration.
Soit λ valeur propre de P' et v son vecteur propre. On a $\forall i \in [1,n] : \sum_{j=1}^{n} P'_{ij} v_j = \lambda v_i$.
Soit i_0 tel $v_{i_0} = \max |v_i|$
On a $\forall i \in [1,n] : \sum_{j=1}^{n} P'_{i,j} v_j = \lambda v_i$ et on divise des deux cotés par v_{i_0}, et on se place à l'indice $i = i_0$.
On obtient alors $\sum_{j=1}^{n} P'_{i_0 j} \frac{v_j}{v_{i_0}} = \lambda$ et on aussi $|\frac{v_j}{v_{i_0}}| \leq 1$, de plus la matrice est stochastique d'où en appliquant la valeur absolue à la somme, on a $|\lambda| \leq 1$. □

Les hypothèses de notre théorème sont vérifiées et on aura donc notre mesure de probabilité qui aura ses entrées strictement positives.
On pourrait alors s'interroger sur l'influence de α i.e du ré-échantillonage : Dans le modèle PageRank la valeur de référence utilisée est 0.85, celle ci est déterminé empiriquement. PageRank utilise cet α pour garantir une bonne convergence lors de l'utilisation de la méthode de la puissance (la vitesse de convergence est d'ordre α^k). Si on choisit α trop grand alors la vitesse de convergence vers la mesure stationnaire n'est pas optimale. Si α est trop petit alors on saute d'individus en individus sans qu'il y ait nécessairement de liens entre eux, cela correspond à une trop grande perturbation de notre modèle.

Par définition du coefficient d'amortissement, α est aussi la probabilité de sélectionner un voisin partant du sommet où on est. De cela, on peut en déduire que la longueur des chemins suivis entre deux rééchantillonages suit une loi géométrique de paramètre α. La longueur moyenne vaut donc $\frac{\alpha}{1-\alpha}$. Ici on va considérer que cette longueur moyenne correspond à la géodésique moyenne. Comme $\ell = 4.4$, le α correspondant serait 0.81. Cette valeur étant proche de 0.85, on considérera 0.85 comme notre valeur de préférence.

5.2.2 Plusieurs coupons

Il nous faut ensuite déterminer comment passer d'une distribution à 1 coupon à celles de 0 à 3 coupons. On passe de la représentation par chaîne à une représentation qui consiste à devoir indexer notre chaîne de Markov par des arbres.
Soient T un arbre avec n noeuds et σ un noeud de T, on note parent(σ) un parent du noeud σ. Le processus de Markov indexé par T est une suite de variables aléatoire ($X_\sigma \in V : \sigma \in T$) $X_{racine(T)}$ est initialisé par π_0.
La matrice de transition est alors défini par $P_{ij} = P(X_\sigma = i | X_{parent(\sigma)} = j)$. En fait, cela revient à numéroter les individus suivants en fonction de la racine de laquelle on part. Finalement on a pas de changement fondamental entre notre cas de départ et ce nouveau cas donc on peut considérer une marché aléatoire simple.

5.3 Résultats et discussion

On va maintenant estimer la taille totale de la population N. Avant d'appliquer à notre cas, on va vérifier que nos estimateurs fonctionnent bien en testant sur un jeu de données connu. On simule alors 1000 graphes qui suivent chacun un modèle d'Erdös-Rényi de paramètres $n = 1000$ et de probabilité $p = 0.002$. Ici on supposera que si on a un degré nul pour un sommet, il est son propre voisin, quitte à augmenter les degrés de tout le monde. On part donc d'une population de 333 personnes et on veut pouvoir retrouver les 1000 à la fin par application des estimateurs obtenus par RDS. On trace alors l'histogramme qui représente l'estimation trouvée pour les deux méthodes :

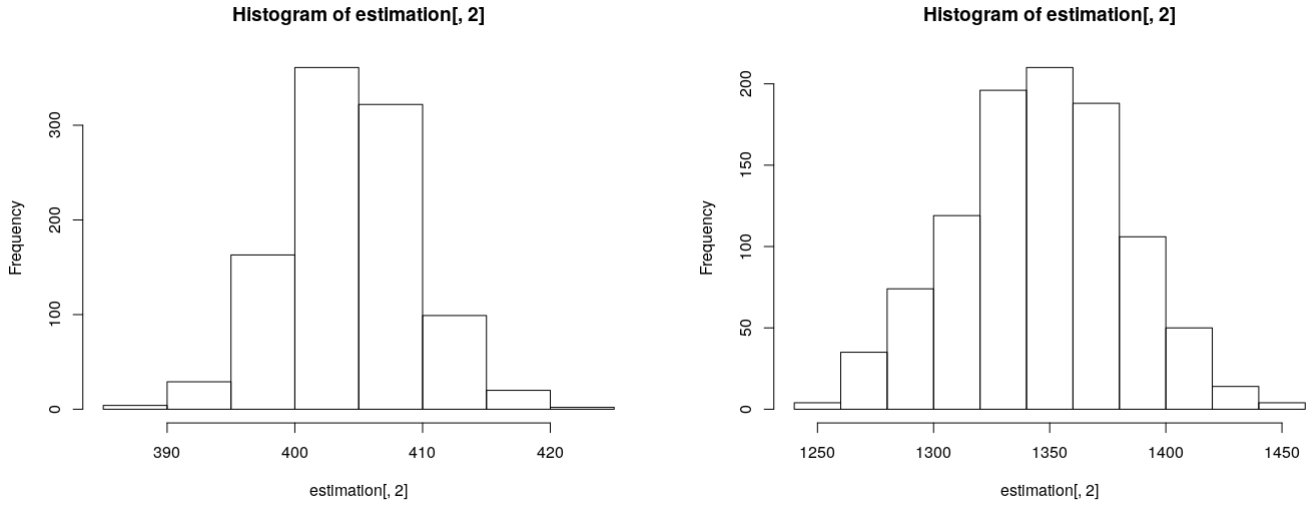

FIGURE 5.1 – Estimation selon la première méthode à gauche et la deuxième à droite

On voit que pour le premier estimateur la médiane se situe autour de 400 ce qui fait que notre estimateur donne une valeur assez sous-estimée de notre population totale. Pour le second estimateur cette fois on a surestimé la valeur de notre population mais la médiane est plus proche de 1000.D'un point de vue pratique surestimer la population est moins risqué que la sous-estimer dans le cas ou on veut faire de la prévention.

Appliquons maintenant ces deux estimateurs à notre cas :

On se placera au point possédant le plus de voisins dans notre graphe, il s'agit de l'individu n°7 avec un degré de 17 pour les calculs.

Estimations sans prendre en compte l'irréductibilité.

On trouve $N = 587$ personnes. Étant donné qu'on part d'une petite population $n = 399$ cette estimation semble peut vraisemblable au vu de la taille de la population de Paris qui est de 2.1 millions de personnes. On peut donc supposer que l'hypothèse d'irréductibilité de Volz-Heckathorn n'avait pas de sens ici.

Estimation avec correction.

α	N
0.1	5058
0.2	5003
0.3	5069
0.4	5275
0.5	5673
0.6	6376
0.7	7612
0.8	10372
0.9	18732
0.85	13142

On observe une nette différence entre la population estimée en supposant l'hypothèse d'irréductibilité vérifiée et celle obtenue après correction. Pour la valeur de référence $\alpha = 0.85$ on obtient une population de $N = 13142$ personnes. Il serait assez difficile de dire si cette estimation semble correcte étant donnée qu'elle est inconnue mais elle semble intuitivement plus réaliste que précédemment.

On a aussi les proportions de femmes et de personnes de 18-30 ans qui sont données par $Pf = 0.157$ et $P18 = 0.209$ soit en multipliant par le N précèdent un nombre de 2071 femmes et un nombre de 2748 personnes entre 18-30 ans.

Conclusion

À partir de la donnée des liens dans le réseau récupérée par la méthode RDS d'échantillonnage des population cachée on a pu montrer que la distribution des degrés de notre réseau suivait une loi de puissance i.e beaucoup de personnes très peu connectées et une petite partie hyper-connectées, ce qui est aussi réaffirmé et nuancé par certaines statistiques (beaucoup de points d'articulations, coefficient de clustering faible, importance des graines). Le clustering révèle l'existence d'un réseau organisé à l'intérieur de nos composantes autour duquel gravite quelques utilisateurs occasionnels. Un objectif important de notre TER concerne l'estimation de la taille totale de la population cachée ; la spécificité des hypothèses concernant les estimateurs liés à la méthode RDS nous a obligé à réechantilloner nos données pour nous adapter au mieux à celles-ci. On obtient alors une estimation de la population totale d'environ 13000 personnes, intuitivement cela peut sembler être une bonne estimation.

Bibliographie

[1] Douglas D.Heckathorn. Respondent-driven sampling : A new approach to the study of hidden populations. *University of Connecticut*, 1997.

[2] Sylvie Deuffic-Burban Marie Jauffret-Roustide Jean-Stephane Dhersin Anthony Cousien, Viet Chi Tran and Yazdan Yazdanpanah. Hepatitis c treatment as prevention of viral transmission and liver-related morbidity in persons who inject drugs. *HEPATOLOGY, Vol. 63, No. 4*, 2016.

[3] J. Travers Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 1969.

[4] *https ://cran.r-project.org/web/packages/igraph/igraph.pdf.*

[5] Alexandre Blondin Massé. Réseaux et graphes. *Université du Québe*, 2014.

[6] C.R. Shalizi A. Clauset and M.E.J. Newman. Power-law distributions in empirical data. *SIAMReview*, 2009.

[7] M. Young A. Clauset and K. S. Gleditsch. On the frequency of severe terrorist events. *J. Conflict Resolution*, 2007.

[8] Gary L Miller Stephen Guattery. On the performance of spectral graph partitioning methods.

[9] Erik Volz and Douglas Heckathorn. *Journal of Official Statistics, Vol. 24, No. 1.*

[10] *https ://web.stanford.edu/group/SOL/dissertations/pagerank-sensitivity-thesis-online.pdf.*

[11] Marc Perrenoud et Karen Brändle Pierre Bataille. « Échantillonner des populations rares ». *Sociologie*, 01 octobre 2018.

[12] M. E. J. Newman. The structure and function of complex networks. *University of Michigan*, 2003.

[13] Olivier Cogis et Claudine Schwartz. Théorie des graphes. *Cassini*, 2018.