



UNIVERSITÉ DE RENNES 1

MASTER MATHÉMATIQUES APPLIQUÉES, STATISTIQUE

Analyse des réseaux sociaux

PÂQUARSE MAHOUVI
22100770

Enseignante : Noemi Navarro

Novembre 2022 — Janvier 2023

Contents

1 Chargement des packages	2
2 Partie I - Représentation et description d'un réseau	2
2.1 Description	2
2.2 Représentation graphique du réseau avec les différentes communautés coloriées par noeud (Force Atlas 2)	2
2.3 Quelques mesures	3
2.4 Graphique et tableau de la distribution des degrées	4
2.4.1 Graphique de la distribution des degrées	4
2.4.2 Tableau de la distribution des dégrés	4
3 Partie II - Régularités sur la structure des réseaux sociaux	5
3.1 Diamètre et longueur moyenne des chaînes	5
3.2 Coefficient de clustering	5
3.3 Distribution des dégrées (ln fréquence vs ln dégré)	6
3.4 Assortativité	6
3.5 Relation entre clustering et dégré	7
4 Partie III. Génération d'un réseau	7
4.1 Génération d'un réseau avec le modèle de Poisson ()	7
4.1.1 Représentation du réseau Erdos-Renyi	8
4.1.2 Caractéristique du réseau Erdos-Renyi	8
4.1.3 Tableau	8
4.2 Génération d'un réseau avec le modèle de Watts et Strogatz (1998)	9
4.2.1 Représentation du réseau de Watts et Strogatz	9
4.2.2 Caractéristique du réseau de Watts et Strogatz	9
4.2.3 Tableau	9
4.3 Génération d'un réseau avec le modèle "caveman"	10
4.3.1 Représentation du réseau caveman	10
4.3.2 Caractéristique du réseau de caveman	10
4.3.3 Tableau	10
4.4 Génération d'un réseau avec le modèle "connected caveman"	11
4.4.1 Représentation du réseau connected caveman	11
4.4.2 Caractéristique du réseau connected caveman	11
4.4.3 Tableau	12
4.5 Génération d'un réseau avec le modèle relaxed caveman"	12
4.5.1 Représentation du réseau "relaxed caveman"	12
4.5.2 Caractéristique du réseau	13
4.5.3 Tableau	13
5 Partie IV - Simulation de diffusion sur un réseau	13
5.1 Recherche des premiers infectés	13
5.2 Pour le modèle SI	14
5.3 Modèle SIR	15
5.4 Modèle Threshold	17

1 Chargement des packages

C'est la toute première partie de notre code. Ici, nous avons importés librairies indispensables à la réalisation de notre projet. Il s'agit entre autre, de networkx, seaborn, pandas, numpy ou encore nbconvert pour avoir le rendu pdf

2 Partie I - Représentation et description d'un réseau

2.1 Description

La base de données utilisée pour le présent projet est composé d'un ensemble de langage de programmation (noeud). La présence d'un lien entre deux noeuds quelconque symbolisant l'existence d'une affinité entre ces deux langages. La variable `value` contenu dans le fichier `links.csv` permet de quantifier cette affinité. Cette base de données provient de kaggle.

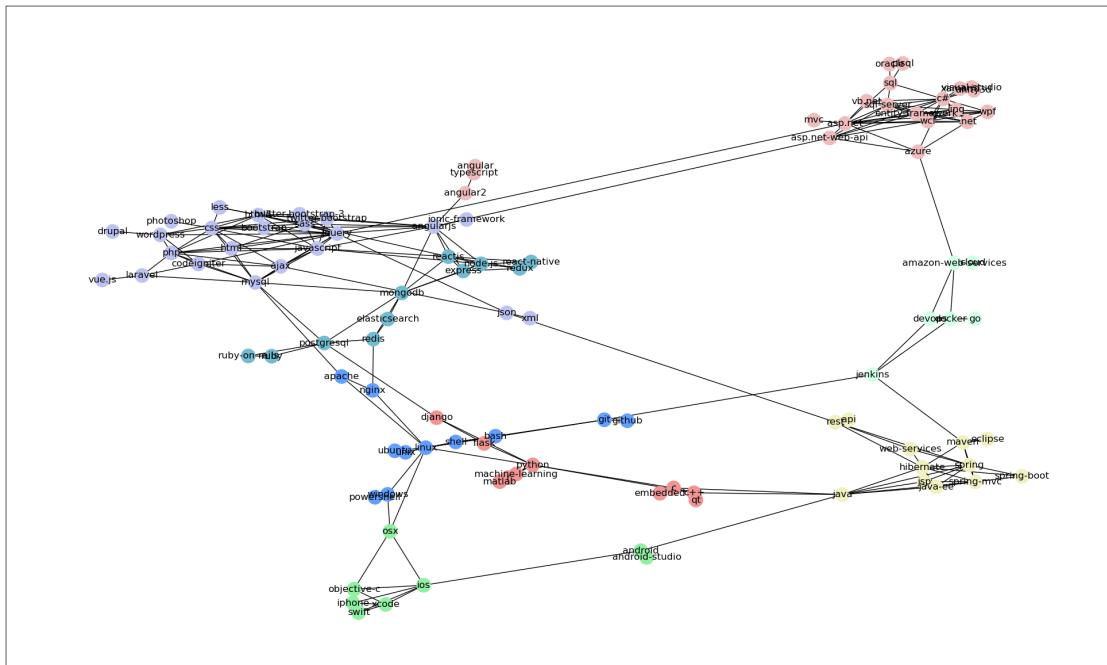
Le code ci-dessous permet d'avoir une visulation dynamique de notre réseau. On peut donc remarquer que notre graphe n'est pas connexe. En effet, elle contient six (06) composantes connexes. Le nombre de sommet contenu dans chaque composante est donné par :

[102, 4, 3, 2, 2, 2]

Nous conviendront, qu'il ne présente aucun ou très peu d'intérêt à étudier le mécanisme et le fonctionnement des graphes à 02, 03 ou 4 sommets. Pour la suite, on travaillera donc uniquement sur la plus importante composante connexe. Soit celui contenant 102 sommets ; il sera nommé g

2.2 Représentation graphique du réseau avec les différentes communautés colorées par noeud (Force Atlas 2)

Warning: uncompiled fa2util module. Compile with cython for a 10-100x speed boost.



On remarque la présence de huit (08) différentes communautés. Au sein de chaque communauté, se retrouve des langages très proches dans la littérature. A titre d'illustration, nous avons une communauté composée des langages osx, android-studio, ios, iphone, objective-c, code et swift. Que ce soit, osx, android-studio, ios, ou iphone, nous avons une idée plus ou moins précise desdits langages. Swift quant à lui, est un langage de programmation open-source développé par Apple pour iOS, macOS, watchOS et tvOS. Objective-c est un langage de programmation orienté objet utilisé pour le développement d'applications pour macOS et iOS. Android-studio est un environnement de développement intégré (IDE) pour le développement d'applications Android. On peut donc remarquer que chaque communauté est composée de langages ayant à peu près le même but ou des objectifs complémentaires. Cela témoigne donc de la capacité de ForceAtlas2 à regrouper des noeuds proches.

2.3 Quelques mesures...

Le graphe est un graphe orienté : False . C'est donc un graphe non orienté

Nombre de noeuds : 102

Le nombre de liens : 235

Le nombre de composante connexe : 1

La densité du graphe G : 0.045622209279751504

Le diamètre du graphe g est : 10 Ainsi, la plus grande distance possible entre deux langages est de : 10

- Le degré moyen d'un graphe indique le nombre de liens connectés à chaque noeud en moyenne.
Dans notre cas :

Le degré moyen est : 4.607843137254902

Cela indique donc que les langages informatiques ont beaucoup d'affinité avec quatre (04) autres langages informatiques en moyenne

Longueur Moyenne des chaînes

La longueur moyenne des chaînes est : 4.463667820069205

En d'autres termes, pour se connecter d'un noeud à n'importe quel autre, il faudrait traverser au moins 4 arêtes. Cela voudra donc dire que pour passer d'un langage informatique à n'importe quel autre, il faudra passer en moyenne par trois (03) autres langages

Le coefficient de clustering Le coefficient de regroupement d'un noeud v est défini comme la probabilité que deux amis de v choisis au hasard soient amis entre eux. Par conséquent, le coefficient de clustering moyen est la moyenne des coefficients de clustering de tous les noeuds. Plus le coefficient de clustering moyen est proche de 1, plus le graphe sera complet car il n'y a qu'une seule composante géante. Enfin, c'est un signe de fermeture triadique car plus le graphe est complet, plus il y aura de triangles.

Le coefficient de clustering est 0.46661551955669606

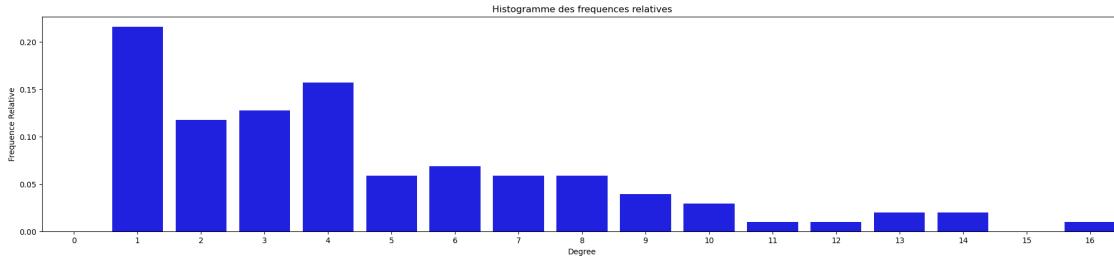
Cela signifie donc que pour un langage v donné, deux langages aux hasards connectés à v ont 46% de chance d'être connectés entre eux

Dans le but de comparer les statistiques obtenues sur le présent réseau avec ceux qui seront générés par les différentes lois de probabilité, lesdites statistiques seront stockées dans une table qui sera étendue par la suite

Graphe	Noeud	liens	densite	diamètre	degré moyen	Longueur Moyenne Chaine	Clustering Coef
g	102	235	0,045	10	4,60	4, 46	0,46

2.4 Graphique et tableau de la distribution des degrées

2.4.1 Graphique de la distribution des degrées



```

css   : 14
asp.net : 13
c#    : 14
javascript : 12
jquery : 16
angularjs : 13

```

On a également les logiciels python et r qui sont les plus utilisés. Il serait également intéressant de regarder leur degré

```

python : 7
r : 3
mysql : 11

```

- Remarque : On remarque qu'un certain nombre de sommet ont des degrés assez élevés. Il serait important d'y jeter un coup d'œil

Commentaire : On remarque que le langage avec le plus grand nombre de liaison est jquery. Alors, pour ceux qui ne connaissent pas jquery, c'est une bibliothèque JavaScript gratuite, libre et multiplateforme compatible avec l'ensemble des navigateurs Web (Internet Explorer, Safari, Chrome, Firefox, etc.), elle a été conçue et développée en 2006 pour faciliter l'écriture de scripts. Il s'agit du framework JavaScript le plus connu et le plus utilisé. Il permet d'agir sur les codes HTML, CSS, JavaScript et AJAX et s'exécute essentiellement côté client. On remarque donc son importance. MySQL, le langage qu'on apprend depuis un certains nombre d'année n'est pas en laisse en matière de relation, contrairement à "r" qui n'en a que trois (03)

2.4.2 Tableau de la distribution des degrés

Tableau de la distribution des degrés

	Degree	Nombre	ln Frequency	ln degree
0	0	0	-inf	-inf
1	1	22	-1.533930	0.000000

2	2	12	-2.140066	0.693147
3	3	13	-2.060023	1.098612
4	4	16	-1.852384	1.386294
5	5	6	-2.833213	1.609438
6	6	7	-2.679063	1.791759
7	7	6	-2.833213	1.945910
8	8	6	-2.833213	2.079442
9	9	4	-3.238678	2.197225
10	10	3	-3.526361	2.302585
11	11	1	-4.624973	2.397895
12	12	1	-4.624973	2.484907
13	13	2	-3.931826	2.564949
14	14	2	-3.931826	2.639057
15	15	0	-inf	2.708050
16	16	1	-4.624973	2.772589

3 Partie II - Régularités sur la structure des réseaux sociaux

Vérification dans la base de données des régularités empiriques des réseaux sociaux

3.1 Diamètre et longueur moyenne des chaînes

Le diamètre du graphe g est : 10

La longueur moyenne des chaines est : 4.463667820069205

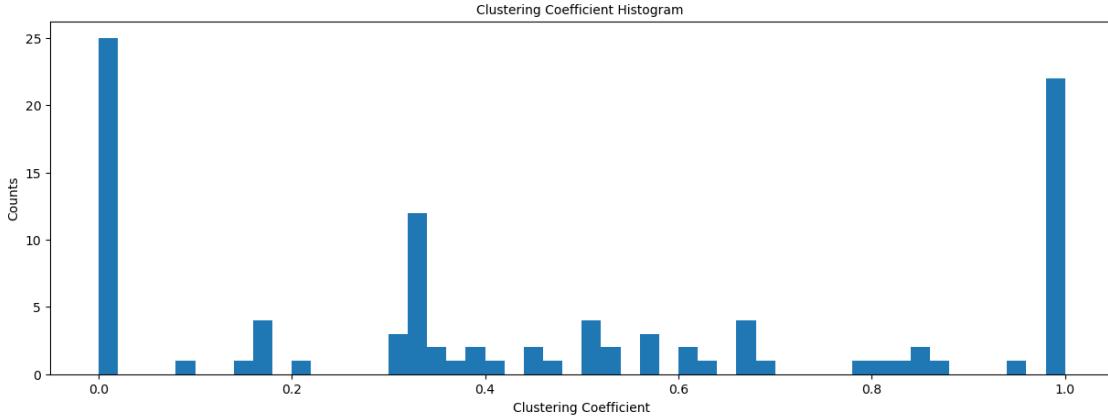
On remarque donc une confirmation de l'hypothèse du petit monde. En effet, d'après l'expérience de Milgram (1933-1984) sur l'envoi des lettres, on avait remarqué que les différents chemins pris par les lettres avaient cinq (05) intermédiaire en moyenne. Dans notre cas, la moyenne des chaines est de 4.46, une valeur très proche des cinq obtenu par Milgram. Ceci témoigne donc de la confirmation de cette hypothèse.

3.2 Coefficient de clustering

le coefficient de clustering est : 0.46661551955669606

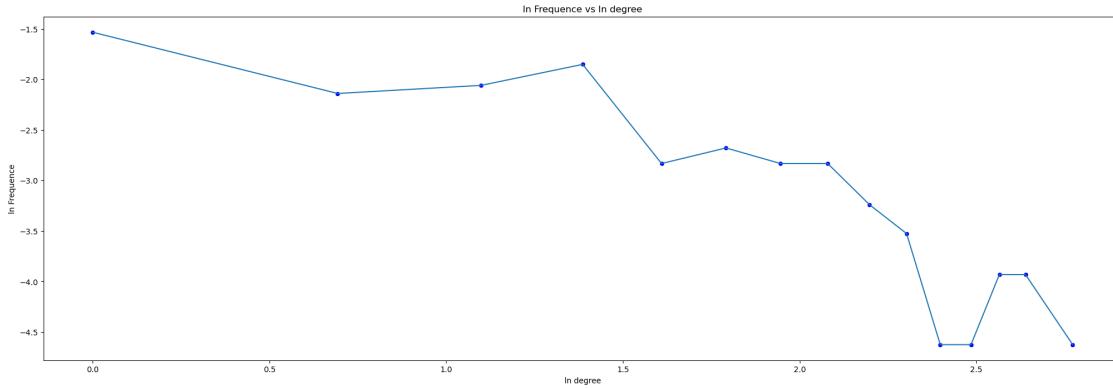
Maintenant, affichons la distribution du coefficient de clustering...

Text(0, 0.5, 'Counts')



En effet, la littérature voudrait que le coefficient de clustering soit élevé par rapport à un grand réseau généré aléatoirement de caractéristiques similaires. En se basant sur les résultats obtenus sur les graphes construits aléatoirement, on remarque la confirmation de cette hypothèse sur les graphes aléatoires ayant des caractéristiques proche de notre graphe g contrairement au graphe

3.3 Distribution des degrées (ln fréquence vs ln degré)



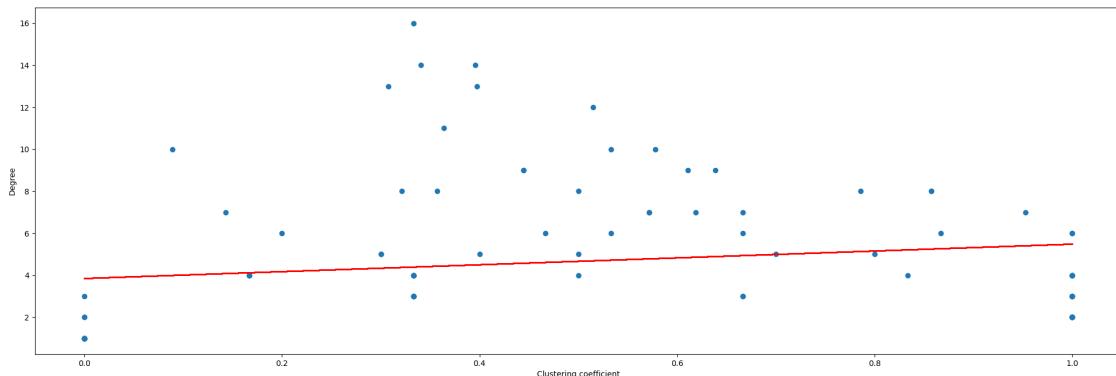
3.4 Assortativité

L'assortativité décrit la préférence des noeuds d'un réseau à s'attacher à d'autres qui sont similaires d'une certaine manière. L'assortativité de degré mesure la tendance des noeuds de degré élevé à être connectés à d'autres noeuds de degré élevé, et des noeuds de degré faible à être connectés à d'autres noeuds de degré faible. Une valeur d'assortativité positive indique que les noeuds de degré élevé ont tendance à être connectés entre eux, et une valeur négative indique que les noeuds de degré élevé ont tendance à être connectés à des noeuds de degré faible. Une valeur de 0 indique un manque de tendance, les noeuds de degré élevé peuvent être connectés avec des noeuds de degré faible. Il est important de noter que l'assortativité d'un réseau peut varier en fonction de la mesure utilisée et de la méthode utilisée pour calculer la valeur d'assortativité. Il peut donc être utile de comparer les résultats à des réseaux similaires.

le coefficient d'assortativité est : 0.12558932726566102

Ainsi, les langages de degré élevé (resp. faible) ont tendance à être connectés à des langages de degré faible (resp. faible)

3.5 Relation entre clustering et dégré



- Commentaire

Il est important de remarquer une corrélation positive, très faible entre le degré et le coefficient de clustering. Cela pourrait signifier que les noeuds ayant un degré élevé ne sont pas nécessairement connectés à d'autres noeuds ayant un grand nombre de lien entre eux.

En résumant, les résultats obtenus sur le graphique g, on a le tableau suivant :

Graphe	Noeud liens	densite	diamètredégré moyen	Longueur Moyenne Chaine	Clustering Coef	Assortativity coef
g	102	235	0,045	10	4,46	0,46

4 Partie III. Génération d'un réseau

4.1 Génération d'un réseau avec le modèle de Poisson ()

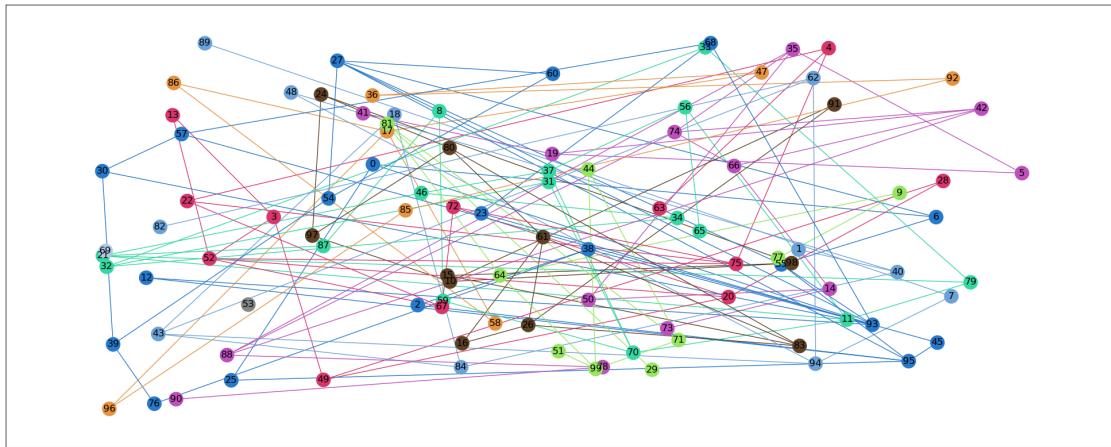
Génération d'un réseau avec le modèle Poisson (ou Erdos-Renyi) avec degré moyen égal au degré moyen du réseau pris dans la partie I, soit le réseau g

Pour se faire, il faudra trouver la bonne probabilité qui permet d'avoir le même degré moyen. Cette probabilité est donnée par :

$$prob = \frac{\text{Nombre de lien}}{\text{Nombre total arrte possible}}$$

avec $\text{Nombre total de sommet} = |V|(|V| - 1)/2$

4.1.1 Représentation du réseau Erdos-Renyi



4.1.2 Caractéristique du réseau Erdos-Renyi

Le nombre de composante connexe : 3

Nombre de noeuds : 100

Le nombre de liens : 232

La densité du graphe d'Erdos-Renyi : 0.04686868686868687

le degré moyen est : 4.64

La longueur moyenne des chaines est : 3.0310204081632657

Le coefficient de clustering est 0.08493867243867242

Le coefficient d'assortativité est : 0.05484417243530656

4.1.3 Tableau

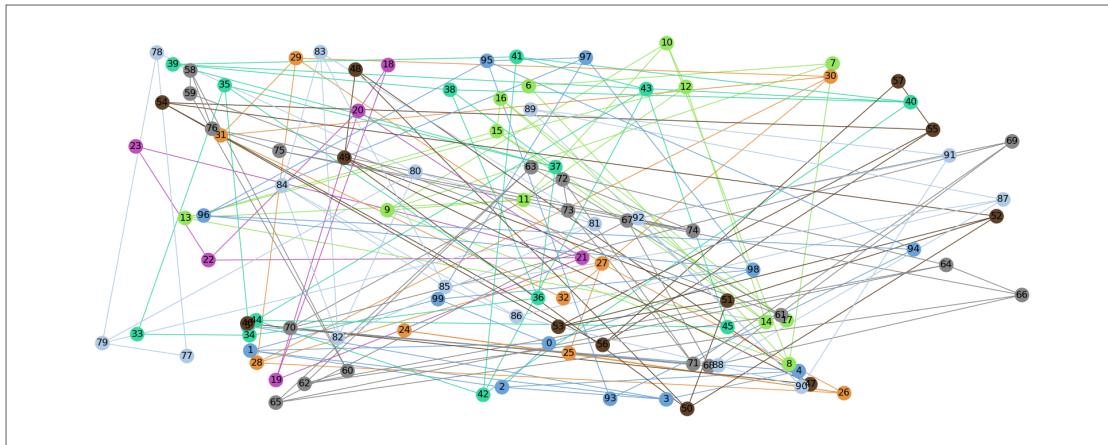
Graphe	componen	Noeuds	Liens	densite	diamèt	régré	Longueu	Clustering	Assortativity
					moyen	moyen	Moyenne	Coef	coef
g	1	102	235	0,045	10	4,60	4,46	0,46	0.125
Erdos-	3	100	232	0,046	-	4,64	4,03	0,08	0.05
Renyi									

Commentaire :

Tout d'abord, il est important de remarquer le caractère non assortatif du réseau d'Erdos-Renyi (assortativity coef faible), ce qui confirme l'une des propriétés de ce réseau ; les noeuds de degré élevé ont autant de chances de se connecter à des noeuds de degré faible qu'à des noeuds de degré élevé. Toutefois, on observe de fortes similitude entre le réseau d'Erdos-Renyi et le graphe utilisé pour cette étude. Cela peut donc indiquer que notre réseau constitué d'ensemble de langage informatique est assez aléatoire et ne suit pas forcément des règles particulières pour la formation des liens.

4.2 Génération d'un réseau avec le modèle de Watts et Strogatz (1998)

4.2.1 Représentation du réseau de Watts et Strogatz



4.2.2 Caractéristique du réseau de Watts et Strogatz

Le nombre de composante connexe : 1

Nombre de noeuds : 100

Le nombre de liens : 200

La densité du graphe de Watts_et_Strogatz est : 0.04040404040404041

Le diamètre du graphe Watts_et_Strogatz est : 11

le degré moyen est : 4.0

La longueur moyenne des chaines est : 5.376599999999999

Le coefficient de clustering est 0.4016666666666667

Le coefficient d'assortativité est : 0.035294117647056644

4.2.3 Tableau

Graphe	componen	Noeuds	Liens	densite	diamètrelégré	Longueur	Clustering	Assortativity
					moyen	Moyenne	Coef	coef
g	1	102	235	0,045	10	4,60	4,46	0,46
Erdos-	3	100	232	0,046	-	4,64	4,03	0,08
Renyi								0,05
Watts	1	100	200	0,040	11	4,00	5,37	0,40
et stro-								0,03
gatz								

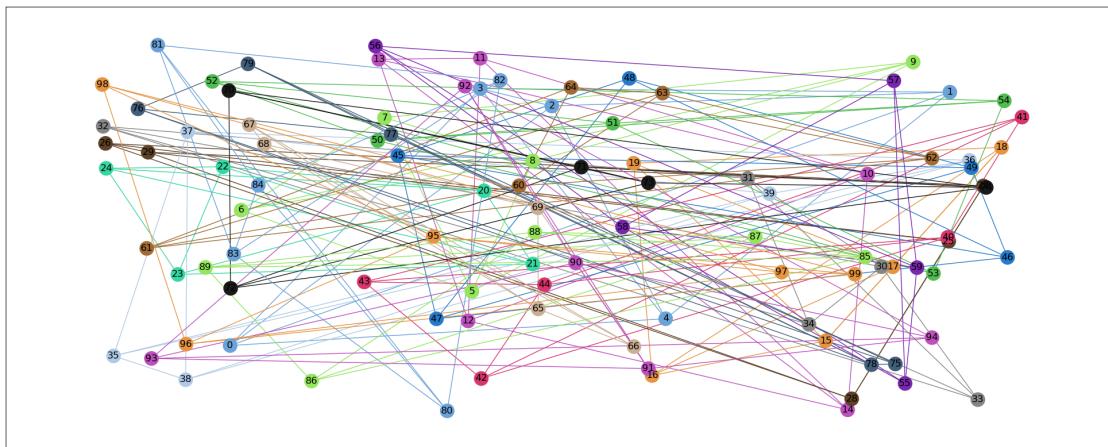
Commentaire :

En premier lieu, il est important de remarquer les énormes similitudes entre notre de réseau de base et le réseau aléatoire généré par la technique de Watts et strogatz. Celui ci confirme l'hypothèse du

petit monde de notre réseau g . On remarque également la positivité du coefficient d'assortativité (certes petit, mais positif). Cela confirme donc la tendance des noeuds de degré élevé (resp. faible) à se connecter à d'autres noeuds de degré élevé (resp. faible). Par ailleurs, tout comme le graphe g , le graphe aléatoire de Watts et strogatz contient une seule composante connexe et à un diamètre de 11 (10 pour le graphe g). Le *dégré moyen* du graphe de watts et strogatz est 4 (4,6 pour le graphe g). Quant à la *longueur moyenne des chaînes*, elle est de 5,37 pour le graphe de watts et strogatz et de 4,46 pour le graphe g . Contrairement au graphe d'Erdos-Renyi, le graphe de Watts et strogatz a un coefficient de clustering très proche du graphe g . Pour rappelle, le coefficient de clustering moyen (coefficient de regroupement) est la moyenne des coefficients de clustering de tous les noeuds. Plus le coefficient de clustering moyen est proche de 1, plus le graphe sera complet car il n'y a qu'une seule composante géante. Ceci témoigne des fortes similitudes qui existent entre le graphe g et le graphque de watts et strogatz.

4.3 Génération d'un réseau avec le modèle "caveman"

4.3.1 Représentation du réseau caveman



4.3.2 Caractéristique du réseau de caveman

Le nombre de composante connexe : 1

Nombre de noeuds : 100

Le nombre de liens : 200

La densité du graphe G : 0.04040404040404041

le degré moyen est : 4.0

La longueur moyenne des chaînes est : 0.7999999999999998

Le coefficient de clustering est 1.0

Le coefficient d'assortativité est : nan

4.3.3 Tableau

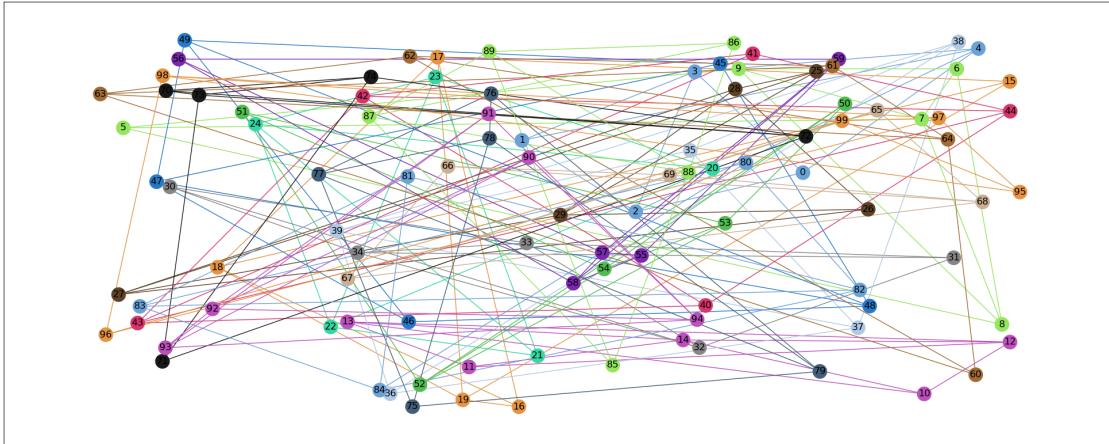
Graphe	composante connexe	Nombre de noeuds	densité	diamètre moyen	dégré moyen	Longueur Moyenne Chaîne	Clustering Coef	Assortativity coef
g	1	102	235	0,045	10	4,60	4,46	0,46
Erdos-Renyi	3	100	232	0,046	-	4,64	4,03	0,08
Watts et strogatz	1	100	200	0,040	11	4,00	5,37	0,40
Caveman	1	100	200	0,040	-	4,00	0,79	1,00

Commentaire :

Contrairement aux deux précédent graphe, le graphe de **Caveman** ne présente pas de lien particulier avec le graphe d'origine g. Certes, elle présente le même nombre de composante connexe, un degré moyen et une densité proche du graphe g, mais présente des caractéristiques très éloignées de g. En effet, la longueur moyenne, qui représente le nombre moyen de chaînes à traverser pour passer d'un noeud à un autre est très éloigné de la longueur moyennes des chaînes du graphe g (0,79 pour Caveman contre 4,46 pour g). Par ailleurs, le coefficient de clustering du graphe **caveman** (1,00) reste quand même éloigné du coefficient de clustering de g quand on prend en compte le fait que le graphe de watts et strogatz à un coefficient de clustering à 0,06 près du coefficient de clustering de g.

4.4 Génération d'un réseau avec le modèle “connected caveman”

4.4.1 Représentation du réseau connected caveman



4.4.2 Caractéristique du réseau connected caveman

Le nombre de composante connexe : 1

Nombre de noeuds : 100

Le nombre de liens : 200

La densité du graphe G : 0.0404040404040404
 Le diamètre du graphe g est : 22
 le degré moyen est : 4.0
 La longueur moyenne des chaînes est : 0.7999999999999998
 Le coefficient de clustering est 0.7333333333333334
 Le coefficient d'assortativité est : -0.2820512820512913

4.4.3 Tableau

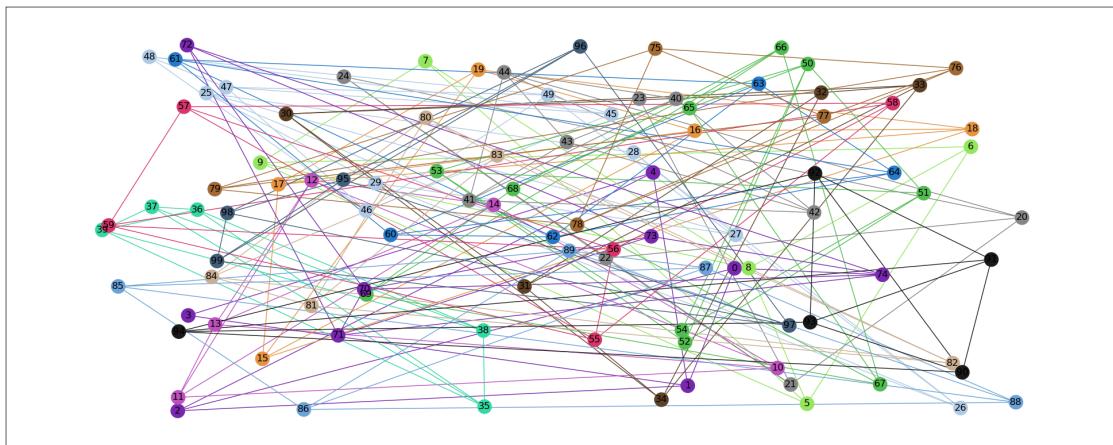
Graphe	composante connexe	Nombre de noeuds	Nombre de liens	densité	diamètre	degré moyen	Longueur Moyenne Chaîne	Clustering Coef	Assortativity coef
g	1	102	235	0,045	10	4,60	4,46	0,46	0.125
Erdos-Renyi	3	100	232	0,046	-	4,64	4,03	0,08	0.05
Watts et strogatz	1	100	200	0,040	11	4,00	5,37	0,40	0.03
Caveman	1	100	200	0,040	-	4,00	0,79	1,00	-
Connected Caveman	1	100	200	0,040	22	4,00	0,79	0,73	-2,28

Commentaire :

Le graphe généré avec le modèle **connected caveman** présente des caractéristiques très proche du modèle de **caveman** qui est lui même assez éloigné du réseau de base g. Par ailleurs, il faut remarquer que ce réseau présente un coefficient d'assortativité négatif. Un coefficient d'assortativité négatif indique que les noeuds de degré élevé ont tendance à être connectés à des noeuds de degré faible, ce qui n'est pas le cas du graphe g.

4.5 Génération d'un réseau avec le modèle "relaxed caveman"

4.5.1 Représentation du réseau "relaxed caveman"



4.5.2 Caractéristique du réseau

Le nombre de composante connexe : 11
 Nombre de noeuds : 100
 Le nombre de liens : 200
 La densité du graphe G : 0.04040404040404041
 le degré moyen est : 4.0
 La longueur moyenne des chaines est : 0.7999999999999998
 Le coefficient de clustering est 0.8756666666666669
 Le coefficient d'assortativité est : 0.02049834294832798

4.5.3 Tableau

Graphe	componen	Noeuds	liens	densite	diamèt	régré	Longueu	Clustering	Assortativity
					moyen	moyen	Moyenne	Coef	coef
g	1	102	235	0,045	10	4,60	4,46	0,46	0.125
Erdos-	3	100	232	0,046	-	4,64	4,03	0,08	0.05
Renyi									
Watts	1	100	200	0,040	11	4,00	5,37	0,40	0.03
et stro-									
gatz									
Caveman	1	100	200	0,040	-	4,00	0,79	1,00	-
Connected	<u>l</u>	Caveman	100	200	0,040	22	4,00	0,79	0,73
Relaxed	11	100	200	0,040	-	4,00	0,79	0,87	-2,28
Cave-									
man									0,02

Commentaire :

Tout comme le modèle de Caveman, le modèle Relaxed caveman présente exactement les mêmes caractéristiques.

Pour conclure, le modèle qui ressemble le plus à notre modèle g est le modèle de Watts et Strogatz. Ce dernier sera donc utilisé pour la suite

5 Partie IV - Simulation de diffusion sur un réseau

5.1 Recherche des premiers infectés

Pour la suite, c'est le réseau généré avec le modèle de Watts et Strogatz qui sera utilisé. La première étape sera de détecter les individus ayant la mesure la plus élevée pour les mesures de centralités (degré, proximité-closness, intermédiaire et vecteur propre). Calculons, ces statistiques.

A cette étape, on arrive malheureusement pas à calculer les centralités de vecteur propre. La solution aura été d'augmenter le nombre d'intégration

```

Noeud Degree closeness_centrality betweenness_centrality \
68      68       6           0.240876          0.260603

```

66	66	5	0.234043	0.184557
45	45	5	0.222472	0.154161
64	64	5	0.220000	0.088690
96	96	5	0.213823	0.103873

Vecteur_Propre	Clustering_coef
68	0.275232
66	0.247061
45	0.126784
64	0.201105
96	0.120690

A présent, nous rechercherons les individus présentant la mesure la plus élevé pour chaque variable...

```
le noeud avec le plus grand degré est le noeud : 68
le noeud avec le plus grand proximité closness est le noeud : 68
le noeud avec le plus grand intermédiarité est le noeud : 68
le noeud avec le plus grand vecteur propre est le noeud : 68
le noeud avec le plus grand coefficient de clustering est le noeud : 34
```

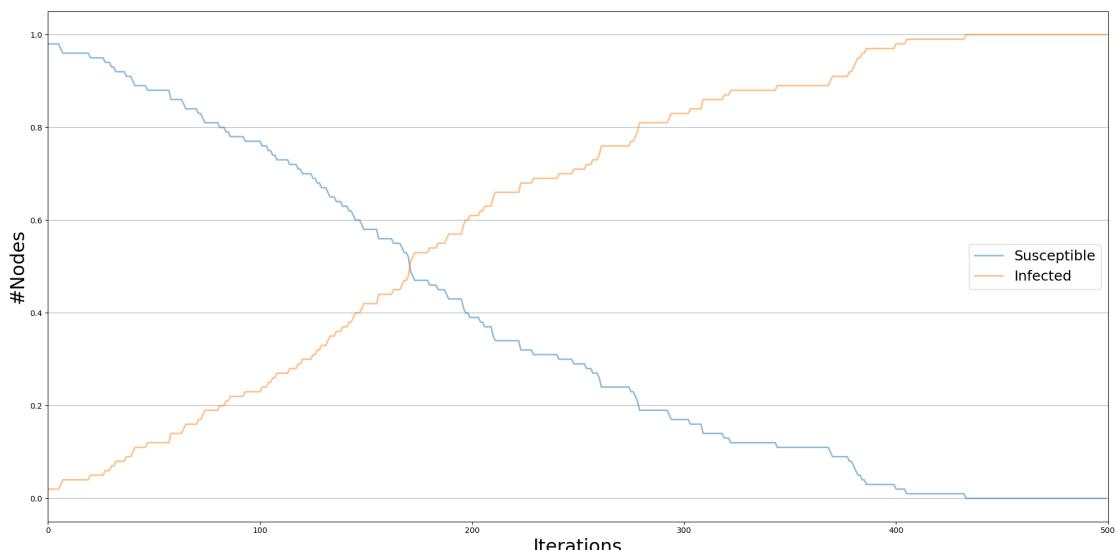
Pour la suite, seront considérés comme des premiers infectés les noeuds : 68 et 34

```
no display found. Using non-interactive Agg backend
```

Avant toute étude, il parait intéressant de s'intéresser à la diffusion du graphe intiale pour

5.2 Pour le modèle SI

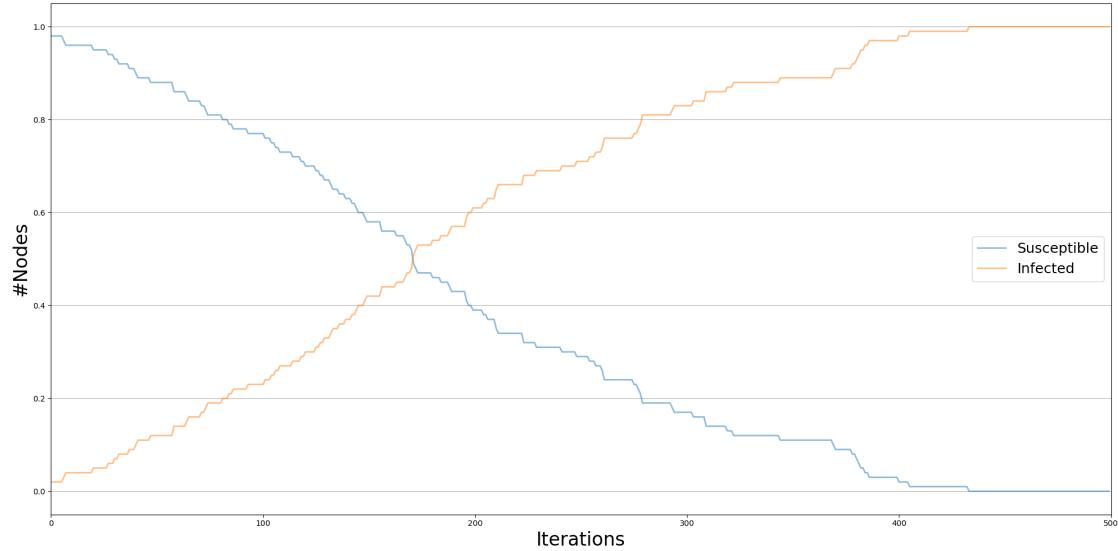
Le modèle SI est un modèle mathématique utilisé pour décrire l'épidémiologie des maladies contagieuses qui ne prend pas en compte le fait que les individus peuvent devenir résistants ou immunisés après avoir été infectés. Les nœuds infectés sont définis comme des nœuds qui ont la maladie dans cette phase initiale. Le paramètre beta est utilisé pour décrire la probabilité qu'un individu infecté transmette la maladie à un individu susceptible par un lien. Les gens peuvent avoir deux états : Susceptible ou infecté



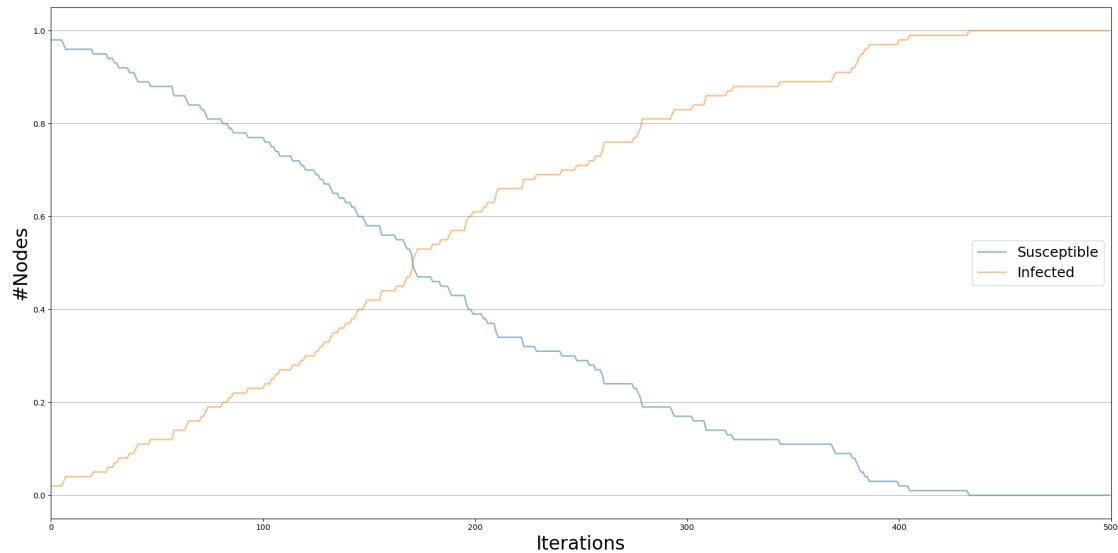
Commentaire : Au bout de 500 itérations, tous les membres de notre réseau adoptent la nouvelle habitude. La valeur minimum de β pour que tous le monde adopte la nouvelle habitude est 0.01

Justification avec d'autres représentation.

Pour $\beta = 0.0001$



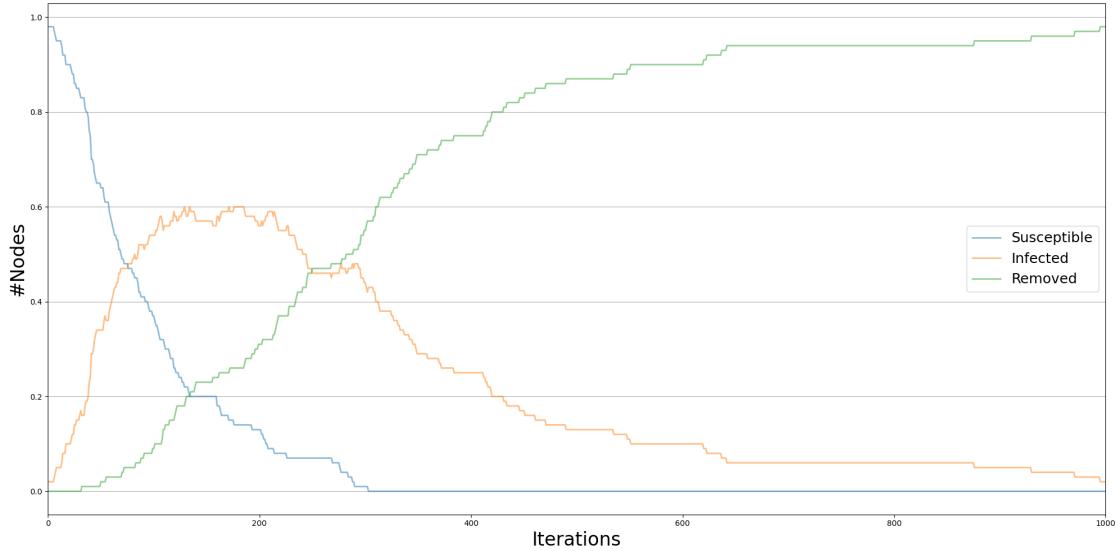
Pour $\beta = 0.5$



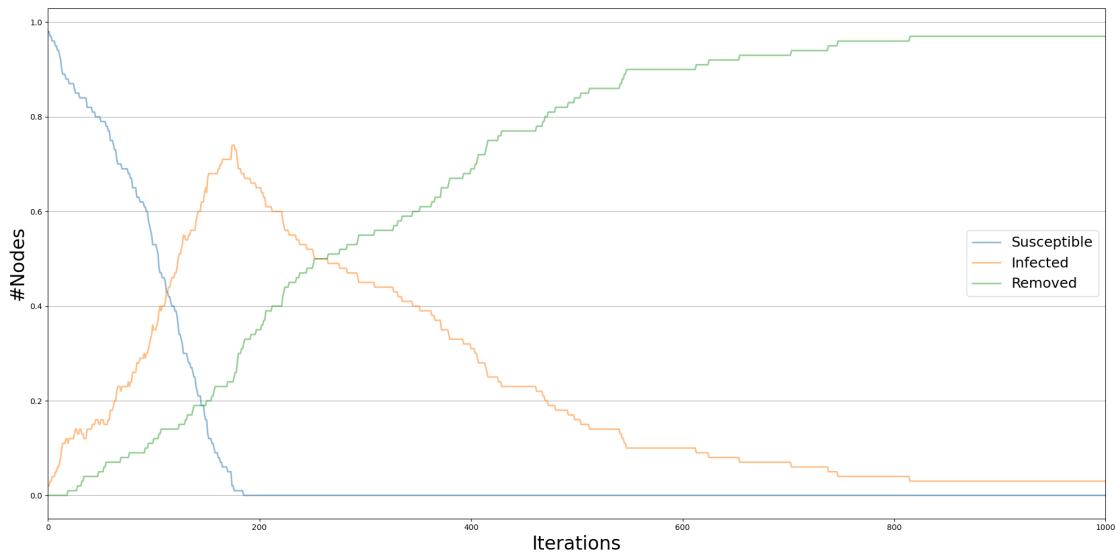
On peut donc remarquer que la meilleure valeur de β est bien 0.01

5.3 Modèle SIR

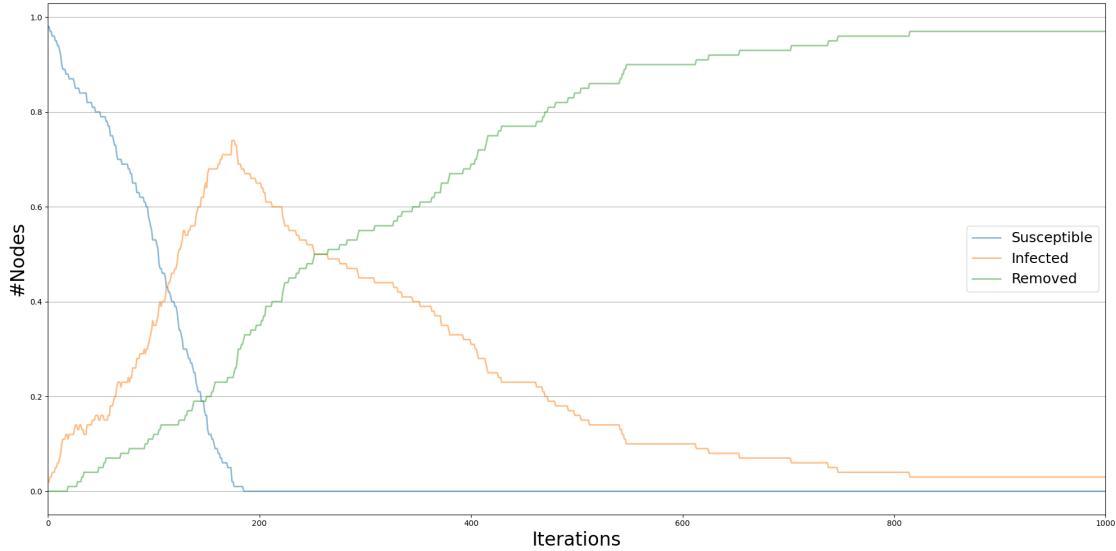
Le rapport $\frac{\beta}{\gamma} = 6$. Soit ($\beta = 0.03$ et $\gamma = 0.005$)



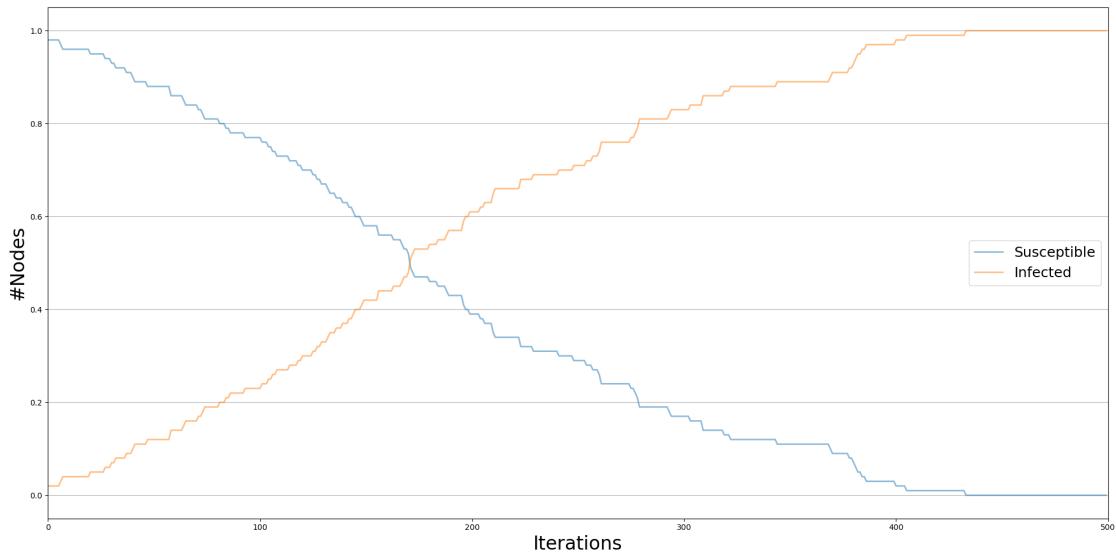
Pour $\beta = 0.04$ et $\gamma = 0.005$



Pour $\beta = 0.02$ et $\gamma = 0.005$



5.4 Modèle Threshold



Pour cibler les nœuds du réseau, si l'objectif est d'obtenir une cascade complète d'adoptions, le meilleur critère est selon nous, la centralité ou encore le clustering. En effet, ayant généralement de grand degré, ils permettent une diffusion beaucoup plus rapide.