

Simulation de la propagation d'épidémie par l'approche PageRank

HPC & Simulation
IATIC 5 2019/2020

Nicolas BLOT
Clément LEFEVRE



Table des matières

TABLE DES MATIERES	2
TABLE DES FIGURES	3
INTRODUCTION	4
DESCRIPTION DU PROBLEME	5
PARAMETRES DE SIMULATIONS	6
A. GRAPHES UTILISES.....	6
B. MATRICE DE TRANSITION ET VECTEUR STATIONNAIRE.....	6
C. AUTRES PARAMETRES DE SIMULATION	7
D. CONTEXTE D'EXECUTION	7
E. TECHNOLOGIES UTILISEES	8
F. EXECUTION DU PROGRAMME	8
ANALYSE DES RESULTATS.....	9
A. IMPACT DE LA POPULATION INFECTEE INITIALEMENT.....	9
B. IMPACT DE LA POPULATION VACCINEE INITIALEMENT	11
CONCLUSION.....	13

Table des figures

FIGURE 1 : ANALOGIE DU PROBLEME	5
FIGURE 2 : DESCRIPTION DES GRAPHS UTILISES	6
FIGURE 3 : CALCUL DE L'ETAT STATIONNAIRE.....	7
FIGURE 4 : SPECIFICATIONS DU PROCESSEUR	7
FIGURE 5 : AUTRES SPECIFICATIONS	8
FIGURE 6 : UTILISATION DU PROGRAMME	8
FIGURE 7 : SIMULATION POUR LE GRAPHE P2P-GNUTELLA09 (5% DE LA POPULATION INITIALEMENT INFECTEE)	9
FIGURE 8 : SIMULATION POUR LE GRAPHE P2P-GNUTELLA08 (5% DE LA POPULATION INITIALEMENT INFECTEE)	9
FIGURE 9 : SIMULATION POUR LE GRAPHE P2P-GNUTELLA05 (5% DE LA POPULATION INITIALEMENT INFECTEE)	9
FIGURE 10 : SIMULATION POUR LE GRAPHE P2P-GNUTELLA04 (5% DE LA POPULATION INITIALEMENT INFECTEE)	9
FIGURE 11 : SIMULATION POUR LE GRAPHE WIKI-VOTE (5% DE LA POPULATION INITIALEMENT INFECTEE)	9
FIGURE 12 : SIMULATION POUR LE GRAPHE P2P-GNUTELLA04 (25% DE LA POPULATION INITIALEMENT INFECTEE)	10
FIGURE 13 : SIMULATION POUR LE GRAPHE P2P-GNUTELLA05 (25% DE LA POPULATION INITIALEMENT INFECTEE)	10
FIGURE 14 : SIMULATION POUR LE GRAPHE P2P-GNUTELLA08 (25% DE LA POPULATION INITIALEMENT INFECTEE)	10
FIGURE 15 : SIMULATION POUR LE GRAPHE P2P-GNUTELLA09 (25% DE LA POPULATION INITIALEMENT INFECTEE)	10
FIGURE 16 : SIMULATION POUR LE GRAPHE WIKI-VOTE (25% DE LA POPULATION INITIALEMENT INFECTEE)	10
FIGURE 17 : SIMULATION POUR LE GRAPHE P2P-GNUTELLA04 (5% DE LA POPULATION INITIALE VACCINEE)	11
FIGURE 18 : SIMULATION POUR LE GRAPHE P2P-GNUTELLA05 (5% DE LA POPULATION INITIALE VACCINEE)	11
FIGURE 19 : SIMULATION POUR LE GRAPHE P2P-GNUTELLA08 (5% DE LA POPULATION INITIALE VACCINEE)	11
FIGURE 20 : SIMULATION POUR LE GRAPHE P2P-GNUTELLA09 (5% DE LA POPULATION INITIALE VACCINEE)	11
FIGURE 21 : SIMULATION POUR LE GRAPHE WIKI-VOTE (5% DE LA POPULATION INITIALE VACCINEE)	11
FIGURE 22 : SIMULATION POUR LE GRAPHE P2P-GNUTELLA04 (5% DE LA POPULATION INITIALE VACCINEE - 25% CONTAMINEE)	12
FIGURE 23 : SIMULATION POUR LE GRAPHE P2P-GNUTELLA05 (5% DE LA POPULATION INITIALE VACCINEE - 25% CONTAMINEE)	12
FIGURE 24 : SIMULATION POUR LE GRAPHE P2P-GNUTELLA09 (5% DE LA POPULATION INITIALE VACCINEE - 25% CONTAMINEE)	12
FIGURE 25 : SIMULATION POUR LE GRAPHE P2P-GNUTELLA08 (5% DE LA POPULATION INITIALE VACCINEE - 25% CONTAMINEE)	12
FIGURE 26 : SIMULATION POUR LE GRAPHE WIKI-VOTE (5% DE LA POPULATION INITIALE VACCINEE - 25% CONTAMINEE)	12

Introduction

Ce projet nous a été confié dans le cadre du module *HPC et Simulation*, encadré par Mme EMAD. Il consiste en la modélisation de la propagation d'une épidémie au sein d'une population, à l'aide de l'algorithme *PageRank*.

L'algorithme *PageRank* est l'algorithme d'analyse des liens concourants au système de classement des pages internet utilisé par le moteur de recherche *Google*. L'idée derrière ce projet est d'utiliser l'approche *PageRank* pour étudier la propagation d'une épidémie.

Dans le contexte actuel, notamment avec l'émergence du virus *COVID-19* (communément appelé *coronavirus*), ce type d'approche trouve toute son importance, et pourrait permettre d'estimer sa propagation, et ainsi de mettre en place différents plans d'actions.

En effet, offrir une réponse rapide et un contrôle efficace face des crises sanitaires de grande ampleur demeurent aujourd'hui un défi majeur pour les scientifiques et les responsables de la santé publique.

L'idée derrière ce projet est donc de vérifier l'avantage et la pertinence de l'approche *PageRank* face à un tel problème, en réalisant différentes simulations.

Après avoir décrit le problème, nous étudierons nos différentes simulations et leurs paramètres, avant d'analyser les résultats obtenus puis de conclure sur la pertinence de cette approche.

Description du problème

Le but de ce problème est de prédire quels individus, ou groupes d'individus sont les plus susceptibles de propager une épidémie.

Ainsi, cela permettrait une réaction rapide et efficace pour contrôler et limiter la propagation de l'épidémie, dans le but d'assister les campagnes de vaccinations menées par les différentes organisations de santé.

De manière générale, la propagation d'un virus est formalisée par trois variables :

- λ_c : infectiosité minimale d'un virus pour envahir un réseau
- ν : taux d'infectiosité d'un individu dans un réseau
- δ : taux de guérison d'un individu contaminé

On calcule ensuite $\lambda = \frac{\nu}{\delta}$, le taux de propagation du virus. Si $\lambda \geq \lambda_c$, alors l'infection se propage et devient persistante, sinon, elle est stoppée rapidement, de manière exponentielle.

Pour modéliser ce problème à l'aide de l'approche PageRank, il convient de faire l'analogie suivante :

Internet	Population (communauté)
Page sur internet	Individu au sein d'une population
Promeneur	Virus
Promenade du marcheur	Propagation du virus
Rang d'une page (probabilité de présence du promeneur sur cette page)	Probabilité d'être infecté par le virus durant une épidémie

Figure 1 : Analogie du problème

D'un point de vue mathématique, le formalisme est le suivant :

$G = (V, E)$: Graphe dirigé

V : Ensemble d'individus

E : Ensemble de liens sortant entre individus

n : Nombre d'individus dans le graphe G

d_i : Nombre de liens de l'individu i vers d'autres individus

$d = (d_1, \dots, d_n)$: degré du graphe

$P_{ji} = P[s_{t+1} = j \mid s_t = i] = \frac{1}{d_i}$, si $i \rightarrow j$: Probabilité que le virus passe d' i à j à un instant t

A l'issue de l'algorithme PageRank, on obtient une distribution stationnaire pour l'état final de la propagation de l'épidémie : $PX = X$, avec $X = (x_1, x_2, \dots, x_n)$ la distribution stationnaire pour l'ensemble de la population (x_i représente ici la probabilité que l'individu i soit infecté au cours de l'épidémie).

De la même manière que pour l'algorithme PageRank, il convient de définir un vecteur de saut, puisqu'il peut arriver que le virus infecte un individu d'une autre manière que par un autre individu.

Afin d'évaluer les performances de l'approche PageRank face à la propagation d'une épidémie, nous allons implémenter trois scénarios :

- Scénario 1 : Propagation du virus sans vaccination.
- Scénario 2 : Propagation du virus avec vaccination d'individus aléatoirement
- Scénario 3 : Propagation du virus avec vaccination des individus les plus susceptibles de le transmettre (obtenus via l'algorithme PageRank).

Paramètres de simulations

A. Graphes utilisés

Pour réaliser les différentes simulations, nous avons utilisé différents graphes proposés par [SNAP](#) (Stanford Large Network Dataset Collection), un projet à l'initiative de l'université de Stanford.

Dans le cadre de cette modélisation, on considère uniquement les graphes dirigés, afin de pouvoir exécuter l'algorithme PageRank de manière optimale.

De manière générale, le format des graphes est le suivant : $X \rightarrow Y$, représentant un arc de X vers Y

Utiliser plusieurs graphes et comparer leurs résultats permet d'affiner et de rendre plus pertinentes nos simulations. Nous avons donc choisi les graphes suivants :

Graphe	Nbr. de sommets	Nbr. d'arêtes	Degré moyen	Densité
p2p-Gnutella04	10876	39994	3.6773	0.0007
p2p-Gnutella05	8846	31839	3.5993	0.0008
p2p-Gnutella08	6301	20777	3.2974	0.0010
p2p-Gnutella09	8114	26013	3.2059	0.0007
wiki-Vote	7115	103689	14.5733	0.0041

Figure 2 : Description des graphes utilisés

Les graphes de la forme *p2p-GnutellaXX* correspondent à des snapshots réalisés en août 2002, issus du réseau de partage peer-to-peer (pair à pair) *Gnutella*.

Le graphe *wiki-Vote* correspond lui à l'historique des votes pour les élections d'administrateurs de la plateforme d'encyclopédie en ligne *Wikipedia*.

La densité d'un graphe est un paramètre particulièrement intéressant pour nos simulations, puisqu'elle représente si le graphe a peu, ou beaucoup d'arêtes. Un graphe dense est un graphe dont le nombre d'arête est proche du nombre maximal. A l'inverse, un graphe creux possède lui peu d'arêtes. La distinction entre graphe creux et dense est relativement vague, et dépend surtout du contexte.

Dans notre cas, le graphe le plus dense est *wiki-Vote*, ce qui signifie que le virus aura tendance à se propager plus facilement au sein de ce graphe.

Les graphes les plus creux sont *p2p-Gnutella04* et *p2p-Gnutella09*, le virus se propagera donc moins vite.

B. Matrice de transition et vecteur stationnaire

La matrice de transition est obtenue à partir de la formule suivante :

$$A = \alpha P + (1 - \alpha) v z^T$$

P : matrice de probabilité

$\alpha \in [0,1]$: facteur d'amortissement

v : vecteur initial

$z = (1, \dots, 1)$

Pour la calculer, nous procédons de la manière suivante :

$$A_{i,j} = \begin{cases} \frac{\alpha R_{i,j}^T}{\sum_{j=1}^n R_{i,j}^T} + (1 - \alpha) \frac{1}{n}, & \text{si } \sum_{j=1}^n R_{i,j}^T \neq 0 \\ \frac{1}{n}, & \text{sinon} \end{cases}$$

Avec $R_{i,j}$ matrice d'adjacence du graphe étudié.

Pour calculer la distribution stationnaire à l'issue de l'algorithme PageRank amélioré, nous procédons de manière itérative, en comparant la différence entre le vecteur obtenu à l'étape k et celui obtenu à l'étape $k-1$ avec un seuil d'erreur (ϵ) : $\|v_k - v_{k-1}\| > \epsilon$

L'algorithme suivi est donc le suivant :

```

Iterate
while ( $\|x^k - x^{k-1}\| > \epsilon$ ) do
     $x^{k+1} = Ax^k$ 
     $x^{k+1} = \frac{x^{k+1}}{\|x^{k+1}\|}$ 
     $k = k + 1$ 
end while
    
```

Figure 3 : Calcul de l'état stationnaire

C. Autres paramètres de simulation

Paramètres pour l'algorithme PageRank :

- Facteur de dumping (α) = 0.85
- Taux d'erreur(ϵ) = 10^{-5}

Paramètres des simulations :

- Nombre d'itérations : **100**
- Probabilité de contaminer un voisin (v) : **0.2**
- Probabilité de contaminer un non-voisin : **1 - α**
- Probabilité de guérir (δ) : **0.1**
- Pourcentage de la population initialement infectée (X) : **10%**
- Pourcentage de la population initialement vaccinée (Y) : **5%**
- Nombre d'itérations : **200**

Afin d'évaluer l'impact du pourcentage de la population initialement infectée (X), et du pourcentage de la population initialement vaccinée (Y), nous ferons varier ces paramètres.

D. Contexte d'exécution

Les différentes simulations présentées au sein de ce rapport ont été effectuées sous le système d'exploitation *MacOS Mojave 10.14.5*, sur un MacBook Pro 13" (mi-2012).

Pour de meilleures performances, les simulations ont été réalisées avec la machine branchée sur secteur.

La configuration matérielle et logicielle est détaillée ci-dessous :

Processeur	Intel Core i5 3210M
Nombre de cœurs	2
Nombre de threads	4
Fréquence	2.5 Ghz
Fréquence Max (Turbo)	3.1 Ghz
Caches	L1: 4x32 Ko, L2: 2x256 Ko, L3: 3 Mo

Figure 4 : Spécifications du processeur

Mémoire (RAM)	16 Go (DDR3 1600Mhz)
Version de Python	3.6

Figure 5 : Autres spécifications

E. Technologies utilisées

Nous avons choisi de réaliser ce projet en *Python*, car il s'agit d'un langage de programmation que nous apprécions, et qui se prête tout particulièrement au calcul matriciel.

De plus, nous utilisons les bibliothèques suivantes :

- *networkx* : permettant la lecture et l'étude de graphes
- *numpy* : permettant de manipuler des tableaux bidimensionnels
- *matplotlib* : permettant de tracer les résultats de nos simulations
- *argparse* : permettant de créer un parseur pour utiliser le programme.

F. Exécution du programme

Le fonctionnement du programme est détaillé au sein des fichiers *README.md* et *README.html*. Afin de faciliter l'installation et l'exécution du programme, un *Makefile* est fourni.

Installation :

make install

Exécution :

make run : Exécute le programme sur l'ensemble des graphes présents dans le dossier graphs.

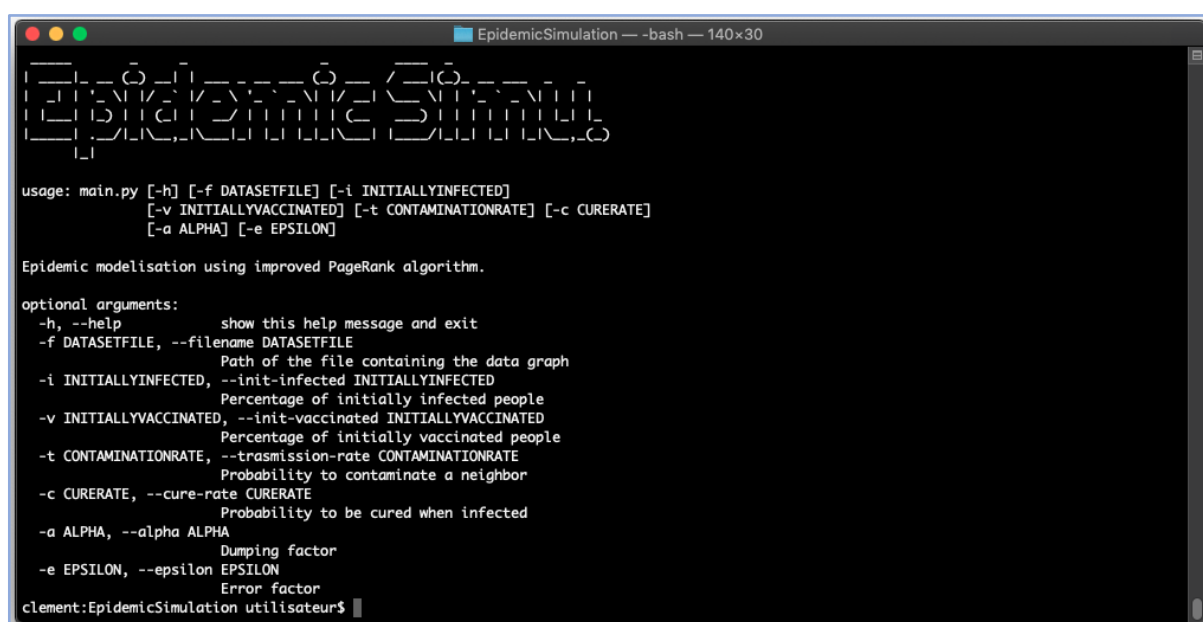
python3 src/main.py -f [filename]

Suppression des graphiques :

make clean

Aide :

make help



```

usage: main.py [-h] [-f DATASETFILE] [-i INITIALLYINFECTED]
              [-v INITIALLYVACCINATED] [-t CONTAMINATIONRATE] [-c CURERATE]
              [-a ALPHA] [-e EPSILON]

Epidemic modelisation using improved PageRank algorithm.

optional arguments:
  -h, --help            show this help message and exit
  -f DATASETFILE, --filename DATASETFILE
                        Path of the file containing the data graph
  -i INITIALLYINFECTED, --init-infected INITIALLYINFECTED
                        Percentage of initially infected people
  -v INITIALLYVACCINATED, --init-vaccinated INITIALLYVACCINATED
                        Percentage of initially vaccinated people
  -t CONTAMINATIONRATE, --transmission-rate CONTAMINATIONRATE
                        Probability to contaminate a neighbor
  -c CURERATE, --cure-rate CURERATE
                        Probability to be cured when infected
  -a ALPHA, --alpha ALPHA
                        Dumping factor
  -e EPSILON, --epsilon EPSILON
                        Error factor
clement:EpidemicSimulation utilisateur$

```

Figure 6 : Utilisation du programme

Analyse des résultats

A. Impact de la population infectée initialement

1. 5% de la population infectée initialement (15% vaccinée)

Dans un premier temps, on considère que 5% de la population est infectée par le virus initialement.

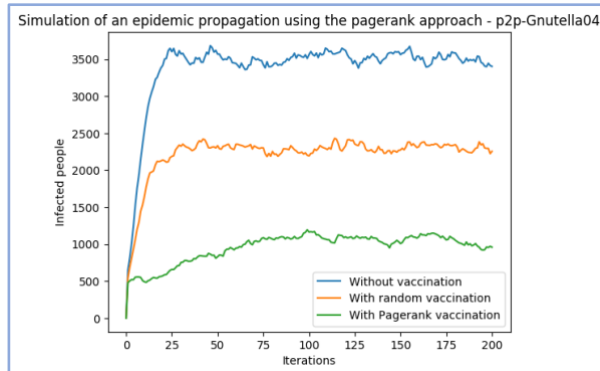


Figure 10 : Simulation pour le graphe p2p-Gnutella04 (5% de la population initialement infectée)

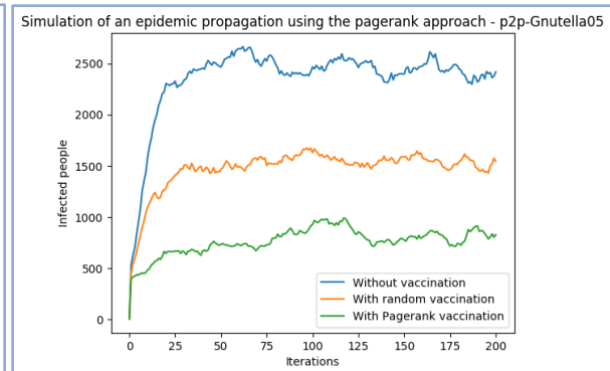


Figure 9 : Simulation pour le graphe p2p-Gnutella05 (5% de la population initialement infectée)

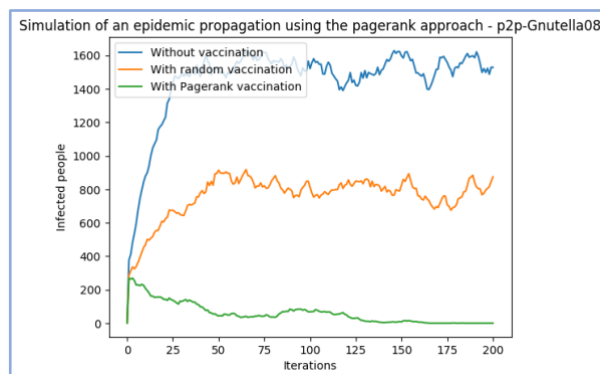


Figure 8 : Simulation pour le graphe p2p-Gnutella08 (5% de la population initialement infectée)

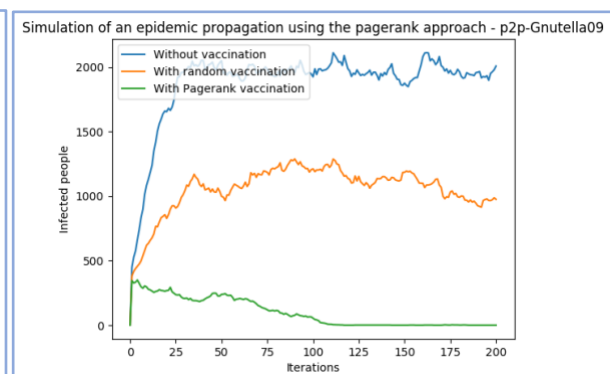


Figure 7 : Simulation pour le graphe p2p-Gnutella09 (5% de la population initialement infectée)

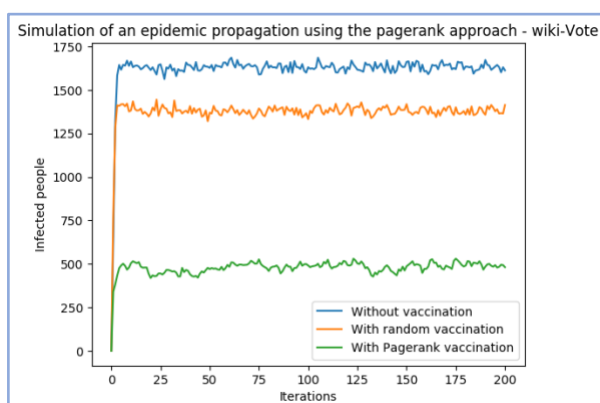


Figure 11 : Simulation pour le graphe wiki-Vote (5% de la population initialement infectée)

On observe assez distinctement les différences de résultats entre les trois approches. Dans le cas d'absence de vaccination, on remarque que le virus s'étend très rapidement, puis sa propagation stagne.

De manière logique la vaccination aléatoire réduit légèrement la propagation du virus, puis à également tendance à stagner.

La vaccination selon l'approche PageRank offre les meilleurs résultats. En effet, en comparaison aux deux autres approches, considérablement moins de personnes sont contaminées.

Dans certains cas de figures (p2p-Gnutella04, p2p-Gnutella05, wiki-Vote), l'épidémie tend à stagner, à une valeur relativement faible. En revanche, si on considère les graphes p2p-Gnutella08 et p2p-Gnutella09, on remarque que l'épidémie disparaît en un temps fini relativement rapide (environ 100-150 itérations). Ces deux graphes sont ceux qui possèdent le moins d'arêtes, la propagation du virus est donc réduite.

2. 25% de la population infectée initialement (15% vaccinée)

Au sein de cette partie, on considère que 25% de la population est infectée initialement par le virus. Ainsi, nous souhaitons observer la réaction des différentes approches de vaccinations, principalement l'approche PageRank.

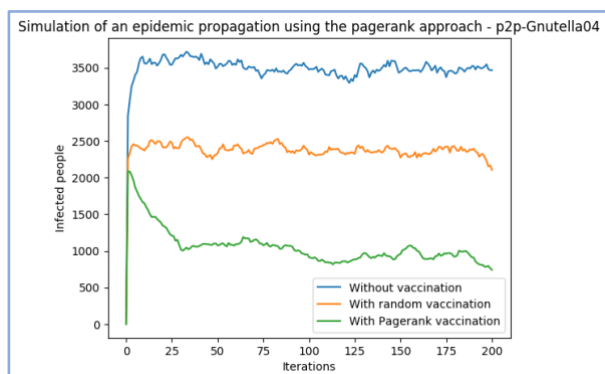


Figure 12 : Simulation pour le graphe p2p-Gnutella04 (25% de la population initialement infectée)

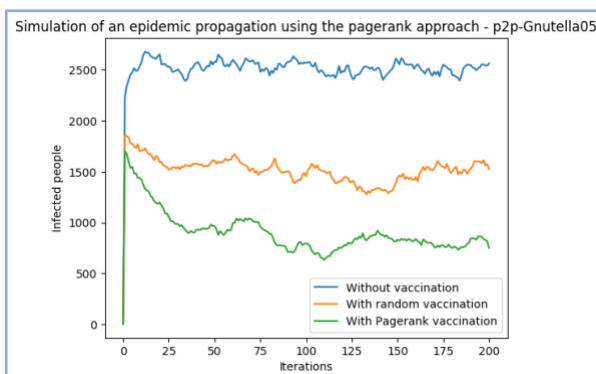


Figure 13 : Simulation pour le graphe p2p-Gnutella05 (25% de la population initialement infectée)

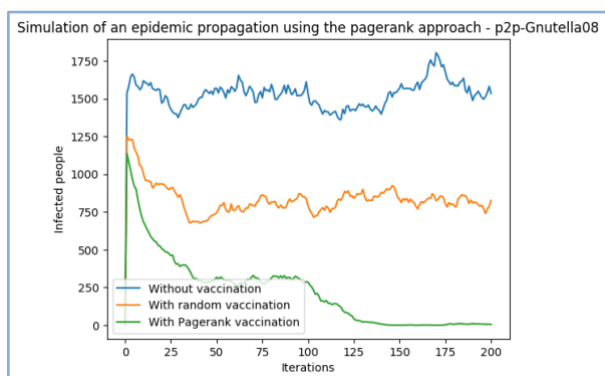


Figure 14 : Simulation pour le graphe p2p-Gnutella08 (25% de la population initialement infectée)

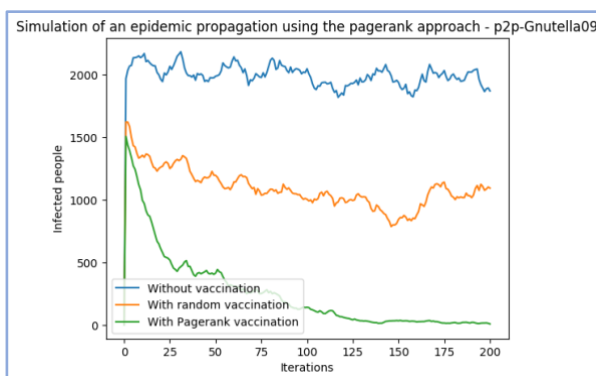


Figure 15 : Simulation pour le graphe p2p-Gnutella09 (25% de la population initialement infectée)

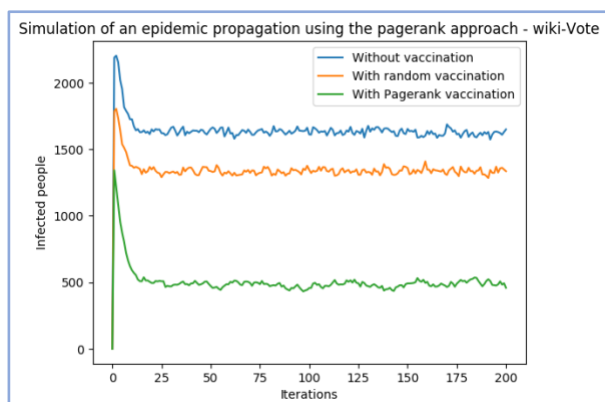


Figure 16 : Simulation pour le graphe wiki-Vote (25% de la population initialement infectée)

Pour ces simulations, un quart (25%) de la population initiale est infectée. On pourrait donc s'attendre à ce que l'épidémie prenne plus d'ampleur. Pourtant, les résultats sont relativement similaires aux précédents. En cas de non-vaccination, la propagation du virus croît rapidement, puis tend à stagner.

Concernant la vaccination selon l'approche PageRank, les résultats sont également similaires aux précédents.

De manière assez étonnante, pour les graphes p2p-Gnutella08 et p2p-Gnutella09, les épidémies sont résorbées au même niveau que lorsque 5%

de la population initiale était infectée (au bout de 100 itérations environ).

L'approche PageRank semble donc assez peu sensible à l'augmentation du pourcentage de la population infectée.

B. Impact de la population vaccinée initialement

On va désormais s'intéresser à évaluer l'impact du taux de population vaccinée initialement. Les résultats présentés en partie A. prenaient en compte un taux de vaccination de 15%.

Au sein de cette partie, on fixera le taux de vaccination à 5%. Ainsi, on pourra évaluer l'impact de ce paramètre.

1. 5% de la population vaccinée initialement (5% contaminée)

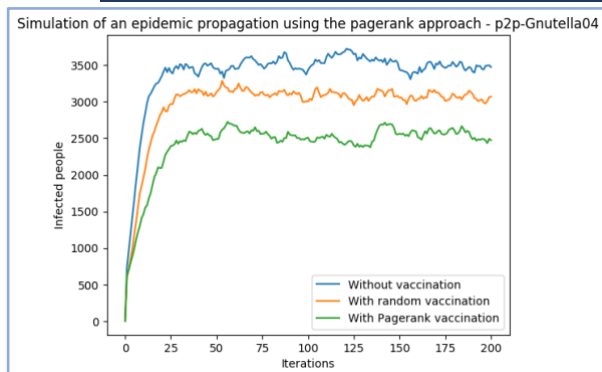


Figure 17 : Simulation pour le graphe p2p-Gnutella04 (5% de la population initiale vaccinée)

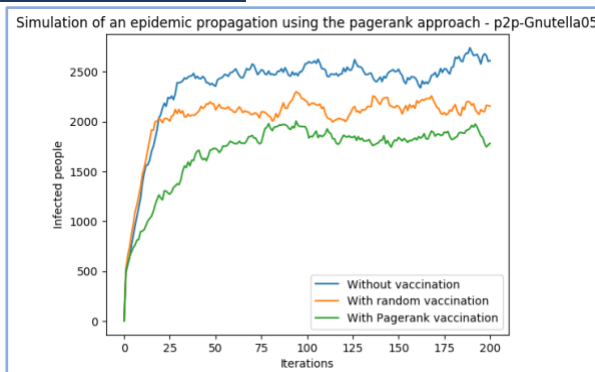


Figure 18 : Simulation pour le graphe p2p-Gnutella05 (5% de la population initiale vaccinée)

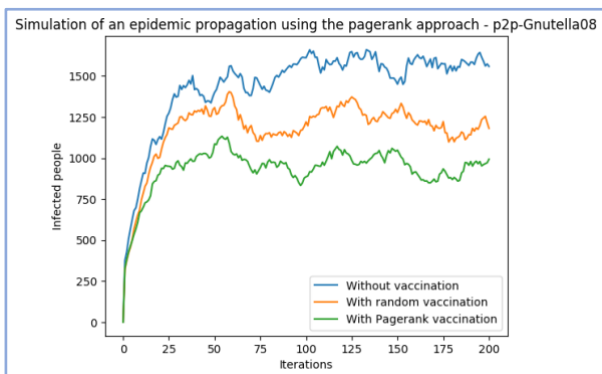


Figure 19 : Simulation pour le graphe p2p-Gnutella08 (5% de la population initiale vaccinée)

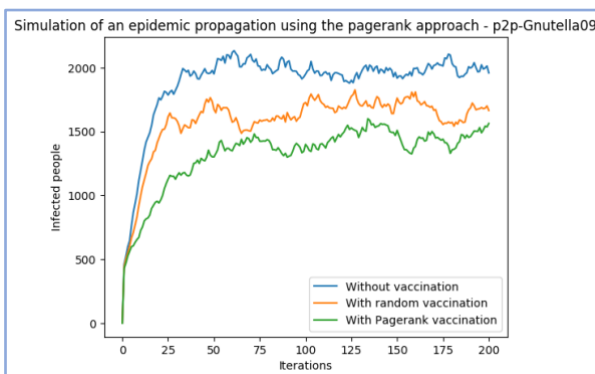


Figure 20 : Simulation pour le graphe p2p-Gnutella09 (5% de la population initiale vaccinée)

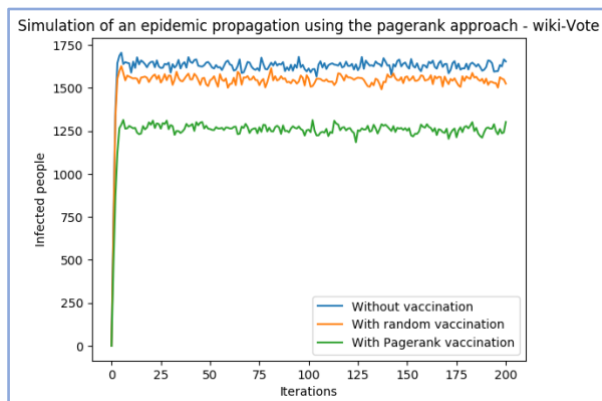


Figure 21 : Simulation pour le graphe wiki-Vote (5% de la population initiale vaccinée)

En premier lieu, on remarque que les trois approches offrent cette fois des résultats relativement similaires. En effet, puisque l'on a réduit le nombre de personnes vaccinées initialement, la vaccination aléatoire et la vaccination PageRank ont un impact plus faible.

Néanmoins, la hiérarchie constatée au cours des observations précédentes est respectée, et l'approche PageRank demeure la plus performante.

Du fait de la réduction de l'influence de la vaccination, l'épidémie n'est plus résorbée pour les graphes p2p-Gnutella08 et p2p-Gnutella09, comme c'était le cas précédemment.

L'approche PageRank a donc été, de manière logique, relativement impactée par la réduction du taux de vaccination initial.

2. 5% de la population vaccinée initialement (25% contaminée)

Pour cette dernière partie, nous allons augmenter le taux de contamination initial à 25%, tout en conservant un taux de vaccination relativement faible (5%).

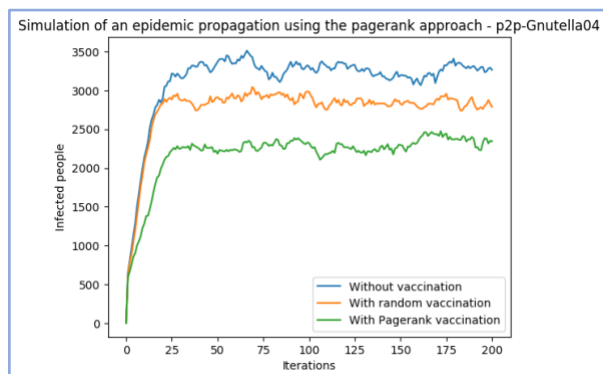


Figure 22 : Simulation pour le graphe p2p-Gnutella04 (5% de la population initiale vaccinée - 25% contaminée)

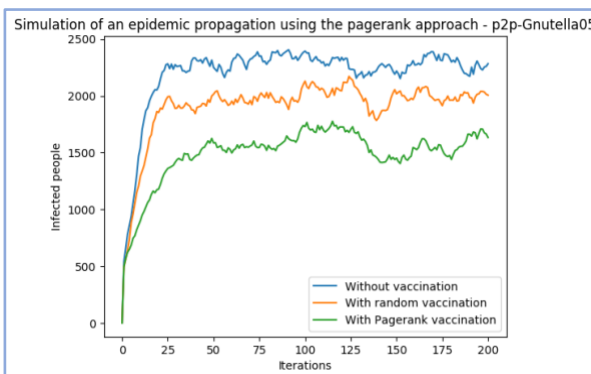


Figure 23 : Simulation pour le graphe p2p-Gnutella05 (5% de la population initiale vaccinée - 25% contaminée)

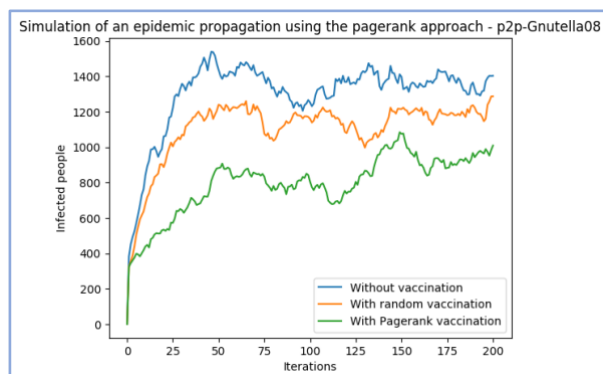


Figure 25 : Simulation pour le graphe p2p-Gnutella08 (5% de la population initiale vaccinée - 25% contaminée)

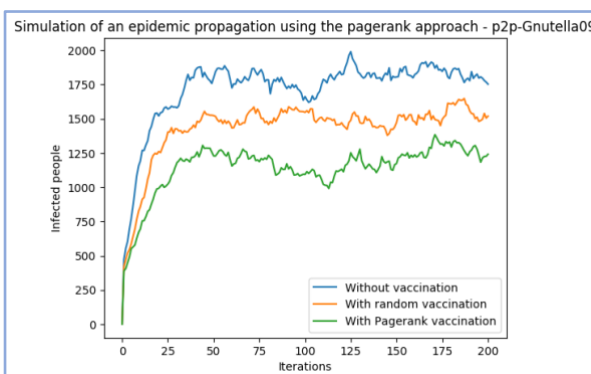


Figure 24 : Simulation pour le graphe p2p-Gnutella09 (5% de la population initiale vaccinée - 25% contaminée)

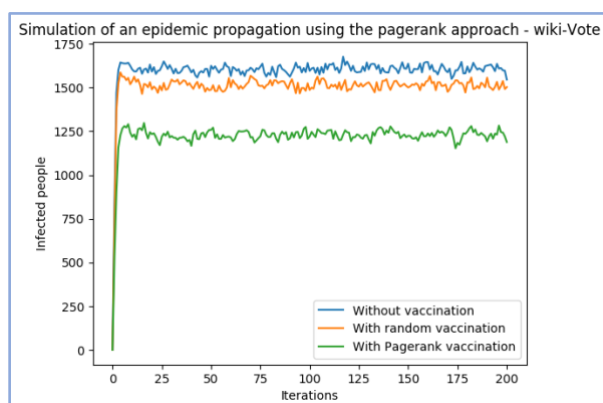


Figure 26 : Simulation pour le graphe wiki-Vote (5% de la population initiale vaccinée - 25% contaminée)

Les résultats sont similaires aux précédents, l'augmentation du taux de contamination initiale ne semble pas impacter énormément la propagation du virus.

On observe que le virus se propage très rapidement dans le dernier graphe, ce qui peut s'expliquer par sa très forte densité (le degré moyen de chaque nœud est supérieur à 14).

Ces deux observations nous permettent donc d'affirmer que l'efficacité de l'approche PageRank dépend assez fortement du taux de vaccination initiale. Ainsi, pour une approche PageRank plus efficace, et donc une résorption de l'épidémie, il peut être judicieux de vacciner au minimum 10% des personnes les plus importantes (susceptibles de propager le plus la maladie).

Conclusion

Au vu des résultats observés, l'approche PageRank semble réellement pertinente face à la propagation d'une épidémie.

En effet, si la vaccination d'individus aléatoirement permet de contenir l'épidémie, celle-ci perdure dans le temps. En revanche, grâce à l'approche PageRank, en vaccinant les individus les plus susceptibles de transmettre le virus, on constate que l'épidémie est rapidement réduite, jusqu'à être totalement résorbée.

Ainsi, nous avons pu constater que, malgré une augmentation du taux de contamination initial, l'approche PageRank permet de réduire, puis de stopper relativement rapidement la propagation d'une épidémie. A l'inverse, le taux de vaccination initial impacte particulièrement la propagation.

Bien que les résultats observés soient particulièrement intéressants, et offrent une première estimation de la propagation d'une épidémie, ils souffrent d'un manque de précision, et d'adéquation avec ce que l'on pourrait observer dans des conditions réelles.

En effet, en réalité, certains individus sont plus vulnérables que d'autres face au virus, et il est donc plus envisageable que ces individus soient facilement contaminables.

A l'inverse, certains individus en bonne santé et possédant une bonne condition physique seraient peut-être moins exposés à une éventuelle contamination. Dans notre cas, le taux de contamination est fixe.

De la même manière, certains facteurs sociétaux, tels que la classe sociale des individus, leur rythme de vie, etc., sont susceptibles d'influencer indirectement la propagation de l'épidémie, mais ne sont pas pris en compte dans cette modélisation.

En conclusion, nous pouvons affirmer que l'approche PageRank permet de lutter de manière plus efficace contre les épidémies, mais il est possible d'optimiser encore cette modélisation en prenant en compte d'autres facteurs. Ainsi, en alliant le Calcul Haute Performance au Big Data, il serait possible de mettre en place une approche encore plus complète, réaliste et pertinente.