

*Rapport d'étude d'un très grand réseau social*



**Maya Besma Le Corre**

Juillet 2018

CNAM – PROBTP - Certification Analyste de données massives

1 -	INTRODUCTION.....	2
2 -	COLLECTE DES DONNÉES.....	2
3 -	PLATEFORME .....	2
4 -	DÉMARCHE .....	2
5 -	ANALYSE.....	3
	Objet.....	3
	Commentaires .....	3
	Résultat .....	3
6 -	VISUALISATION.....	7
	Analyse statistiques.....	7
	Découverte du phénomène cercle d’ami .....	8
	Choix des paramètres et algorithmes de spatialisation .....	9
	Visualisation du réseau social .....	10
7 -	CONCLUSION .....	11
8 -	RÉFÉRENCES.....	11

## 1 - INTRODUCTION

Le but de l'étude consiste à analyser la structure du réseau social. Les questions topologiques de communautés d'utilisateurs seront abordées. Une visualisation partielle du graphe en utilisant les techniques et principes abordés en cours est proposée.

## 2 - COLLECTE DES DONNÉES

Friendster a été un réseau social sur lequel des utilisateurs pouvaient nouer des contacts avec d'autres et partager des contenus (essentiellement orienté vers les jeux). Considéré comme un ancêtre des réseaux sociaux comme Facebook et Twitter, il a disparu en 2015. <http://socialcomputing.asu.edu/datasets/Friendster>,

## 3 - PLATEFORME

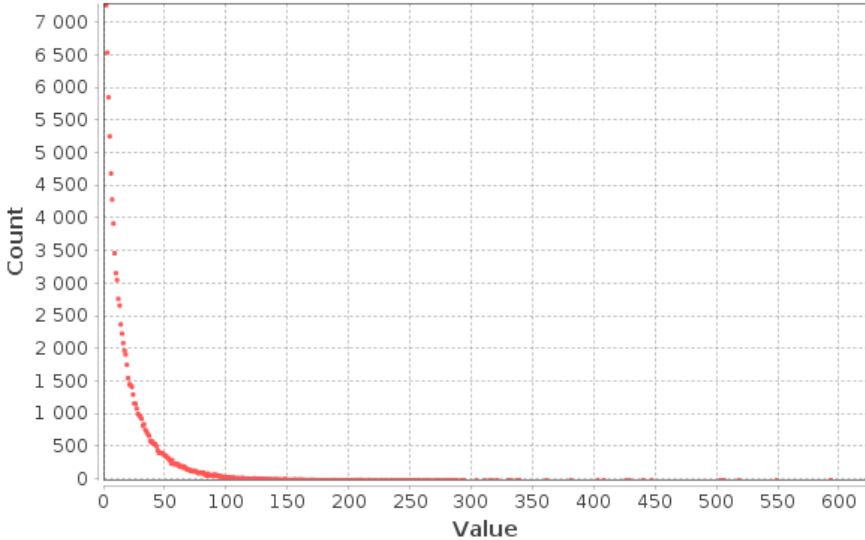
1. Plateforme Spark –version 2.2.0 (version java : « 1.8.0\_171 »)
2. Notebook Jupyter --version 4.3.0
3. Kernels : Scala –version 2.11.8

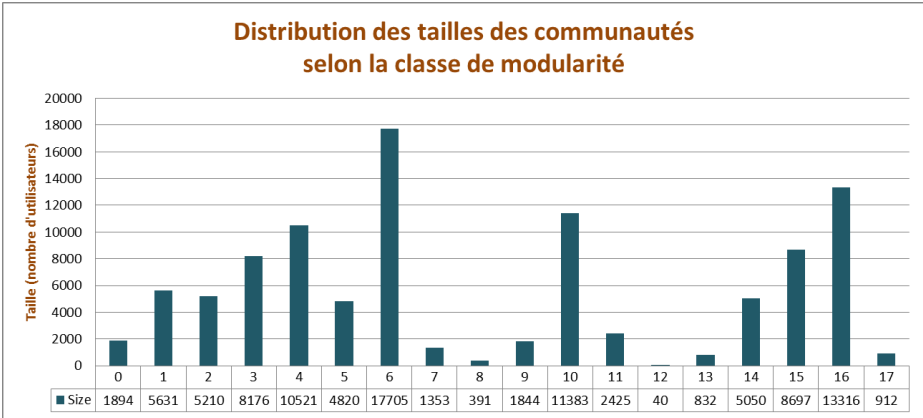
## 4 - DÉMARCHE

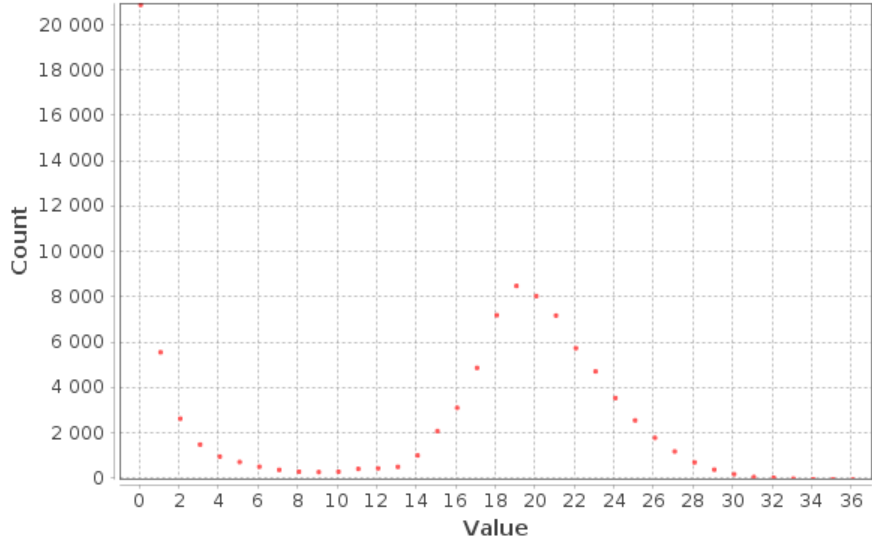
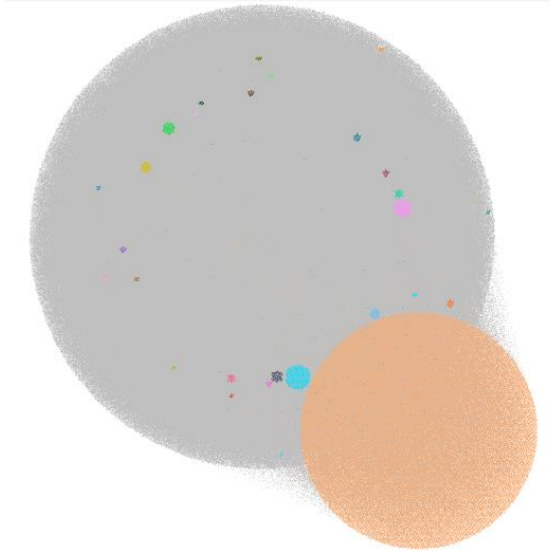
1. Création du sous graphe à partir des données
2. Analyse topologique de la structure globale du réseau et du sous graphe
3. Analyse spectrale de la structure globale du réseau et du sous graphe
4. Visualisation du sous graphe

## 5 - ANALYSE

Objet	Commentaires	Résultat
Informations générales	<p>Le réseau est simplement décrit par sa structure. Les nœuds ne possèdent pas d'attribut. Le graphe est défini comme non orienté et non pondéré. Le couple d'entier (1,2) dans le fichier edges.csv signifie que 1 est ami avec 2.</p> <p>Le fichier nodes.csv contient tous les utilisateurs et doit être considéré en tant que dictionnaire. Le site stipule qu'il n'y a aucune donnée manquante.</p> <p>Pourtant, après analyse du fichier edges.csv, il apparaît des nœuds non présents dans le dictionnaire.</p> <p>Une première tentative infructueuse a consisté à analyser le graphe en tenant compte de tous les nœuds y compris les absents. Après avoir supprimé les quelques centaines d'arêtes orientées, On obtient un graphe quelconque sans boucle complètement connexe. Le résultat est resté infructueux (tous les calculs n'aboutissent pas). Les calculs aboutissent sur chaque sous-graphe pris séparément.</p> <p>L'absence d'indication oblige à travailler sur hypothèse. Je fais l'hypothèse suivante pour expliquer les nœuds «absents» du dictionnaire : il existe deux sous graphes représentant des relations différentes entre les utilisateurs.</p> <p>Le sous graphe étudié est celui qui tient compte uniquement des nœuds présents dans le dictionnaire. Il sera considéré comme un graphe orienté étant donné la présence d'arêtes orientées (on trouve quelques liens 1,2 et 2,1).</p> <p>Le réseau Friendster un graphe en deux parties. Il est ni eulérien, ni hamiltonien, ni cordal et il contient des isthmes.</p>	<ul style="list-style-type: none"> <li>Le graphe complet :  Type: Graph  Number of nodes: 5 689 498  Number of edges: 14 067 887  Average degree: 4.9452  Densité : 0  C'est graphe creux donc</li> </ul> <p>Nombre de cliques : 12 889 424  Test Erdős-Gallan =&gt; True  Test method="hh" =&gt; True</p> <ul style="list-style-type: none"> <li>Le sous-graphe étudié :  Type: DiGraph  Number of nodes: 100 199  Number of edges: 981 920  Average degree: 9,8  Densité : 0  Diamètre : 36  Rayon : 0</li> </ul>
Connexité	Le sous graphe étudié est connexe. Il n'existe pas de nœud isolé et contient 13462 ponts. C'est une composante connexe géante.	Nombre de composante connexe : 1 seule composante faiblement connectée

Objet	Commentaires	Résultat
Distribution des degrés	<p>Les résultats sont beaucoup plus rapides à obtenir et sont surtout plus lisibles en considérant un graphe orienté. En moyenne un utilisateur a entre 9 et 10 amis. La distribution de degrés montre une très grande variance entre les utilisateurs.</p> <p><b>Degree Distribution</b></p>  <p>La courbe est caractéristique des réseaux de petit monde. On trouve de nombreux utilisateurs avec peu d'ami et quelques utilisateurs (en général appelé des Hubs) avec un grand nombre d'ami. Cette distribution suit une loi de puissance. Cela met en évidence le phénomène d'attractivité plus important des liens ayant un degré plus important. Les plus gros grossissent plus vite. On voit ci-contre le top 10 des Hubs.</p>	<p><b>Classements des utilisateurs par ordre décroissant de degrés :</b></p> <p>(3687,621) ; (3552,592) ; (3570,592) ;  (3553,548) ; (82,517) ; (3542,505) ;  (212,502) ; (3674,446) ; (2680,439) ;  (3651,428)</p>

Objet	Commentaires	Résultat																		
Recherche de communauté	<p>La modularité calculée est cohérente avec la présence de communauté vu sa valeur significative (supérieure à 0.3).</p> <p>Il existe des communautés car on trouve beaucoup d'ensembles de sommets très liés entre eux et peu liés vers le reste du graphe.</p>	<p>Modularité: 0.691</p> <p><b>Nombre de communauté : 18</b></p> <p>(Paramétrage effectué pour une chercher ce niveau de précision. En augmentant la précision on peut détecter plus de 26 000 communautés)</p> <p><b>Coefficient de clustering :</b></p> <ul style="list-style-type: none"><li>- résultat avec Spark : 26,75 %</li><li>- moyenne des coefficients de clustering : 0,134</li></ul>																		
	 <table><tr><th>Size</th><td>1894</td><td>5631</td><td>5210</td><td>8176</td><td>10521</td><td>4820</td><td>17705</td><td>1353</td><td>391</td><td>1844</td><td>11383</td><td>2425</td><td>40</td><td>832</td><td>5050</td><td>8697</td><td>13316</td><td>912</td></tr></table>	Size	1894	5631	5210	8176	10521	4820	17705	1353	391	1844	11383	2425	40	832	5050	8697	13316	912
Size	1894	5631	5210	8176	10521	4820	17705	1353	391	1844	11383	2425	40	832	5050	8697	13316	912		

Objet	Commentaires	Résultat
Vue générale des liens	<p style="text-align: center;"><b>Eccentricity Distribution</b></p>  <p style="text-align: center;">Figure 1 : Distribution de l'excentricité</p> <p>L'éloignement moyen entre les utilisateurs de ce réseau social est assez petit. C'est une mesure topologique qui abonde dans le sens de la structure de petit monde. La distribution suit une loi normale décentrée autour d'une excentricité de 20.</p> <p>La distribution de la centralité est répartie selon une loi de Pareto : 20% des utilisateurs sont très centraux contre une faible centralité de 80% des utilisateurs.</p>	<p>Distance du parcours moyen : 7.3</p>  <p style="text-align: center;">Figure 2 : Représentation circulaire de la distribution de la centralité</p> <p>En orange claire se trouve plus de 23000 utilisateurs avec la plus forte centralité.</p>

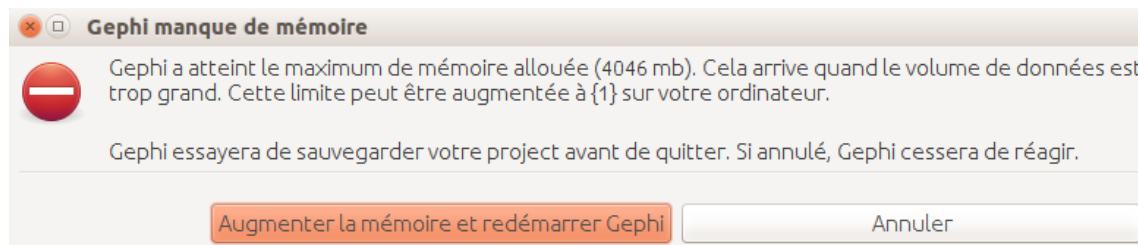
Objet	Commentaires	Résultat																						
PageRank	Cet algorithme de parcours permet de retrouver ainsi les sommets les plus « populaires ». Difficiles de se prononcer sans informations sur les utilisateurs mais on constate malgré tout qu’il s’agit des plus petits identifiants.	<table><tr><th>Utilisateur</th><th>Score pageRank</th></tr><tr><td>1</td><td>8020</td></tr><tr><td>2</td><td>3474</td></tr><tr><td>82</td><td>1692</td></tr><tr><td>11</td><td>1039</td></tr><tr><td>3</td><td>1034</td></tr><tr><td>13</td><td>921</td></tr><tr><td>4</td><td>671</td></tr><tr><td>54</td><td>648</td></tr><tr><td>5</td><td>547</td></tr><tr><td>20</td><td>541</td></tr></table>	Utilisateur	Score pageRank	1	8020	2	3474	82	1692	11	1039	3	1034	13	921	4	671	54	648	5	547	20	541
Utilisateur	Score pageRank																							
1	8020																							
2	3474																							
82	1692																							
11	1039																							
3	1034																							
13	921																							
4	671																							
54	648																							
5	547																							
20	541																							

## 6 - VISUALISATION

### Analyse statistiques

Les rapports statistiques ont été obtenus à l'aide de GEPHI. Plusieurs tentatives :

- Première tentative avec le fichier original. Utilisation de Gephi : la taille du réseau est trop importante. Malgré un découpage en fichier plus petit et en modifiant la taille de la mémoire utilisable au lancement de l'application. Un fichier de moins de 200 méga octets n'est pas visualisable sur mon poste de travail (mémoire attribuée 4Go)





- Seconde tentative réussit avec le graphe analysé plus haut.

L'ensemble des mesures topologiques sont caractéristiques des réseaux de terrain. Les différents rapports ont disponible sur le repository GitHub dédié à ce rapport projet. La modularité a été utilisée pour la coloration des différentes communautés principalement. Le nombre d'utilisateur par classe de modularité donne la taille de la communauté.

Découverte du phénomène cercle d'ami

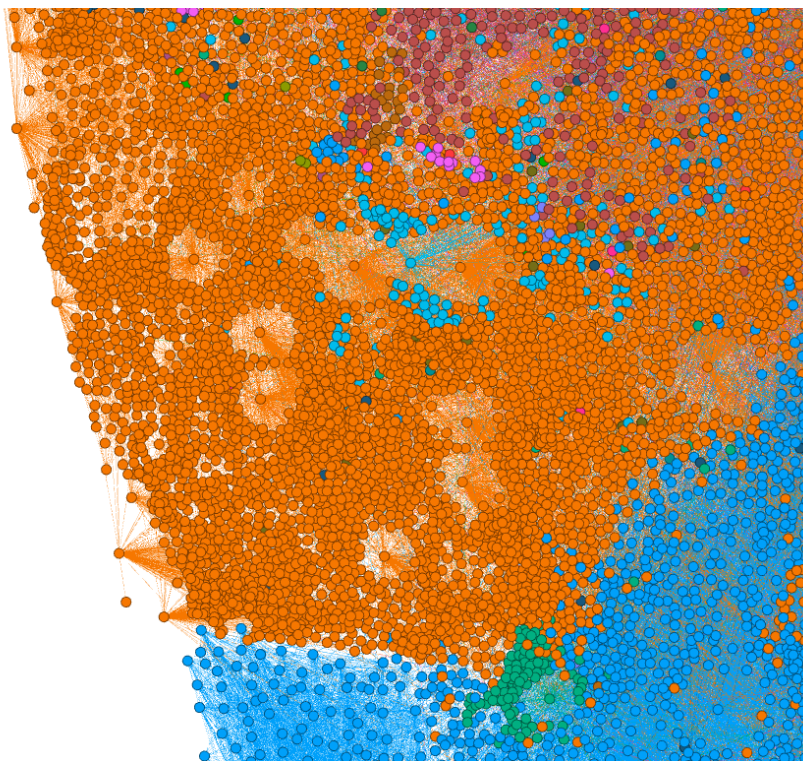


Figure 3 : Zoom Cercle d'ami

Avec la spatialisation OpenOrd et en supprimant les utilisateurs n'ayant qu'un seul ami, on découvre que cette particularité concerne que certains nœuds. Selon la description du réseau par Wikipédia, une fonctionnalité du site permettait de définir un cercle d'ami.

### Choix des paramètres et algorithmes de spatialisation

Une tentative de représentation sur la base des valeurs propres pour effectuer une analyse spectrale n'a pas donné satisfaction en ce qui concerne la recherche de communauté pour ce réseau social (voir le résultat en annexe dans le repository GitHub dédié à ce rapport).

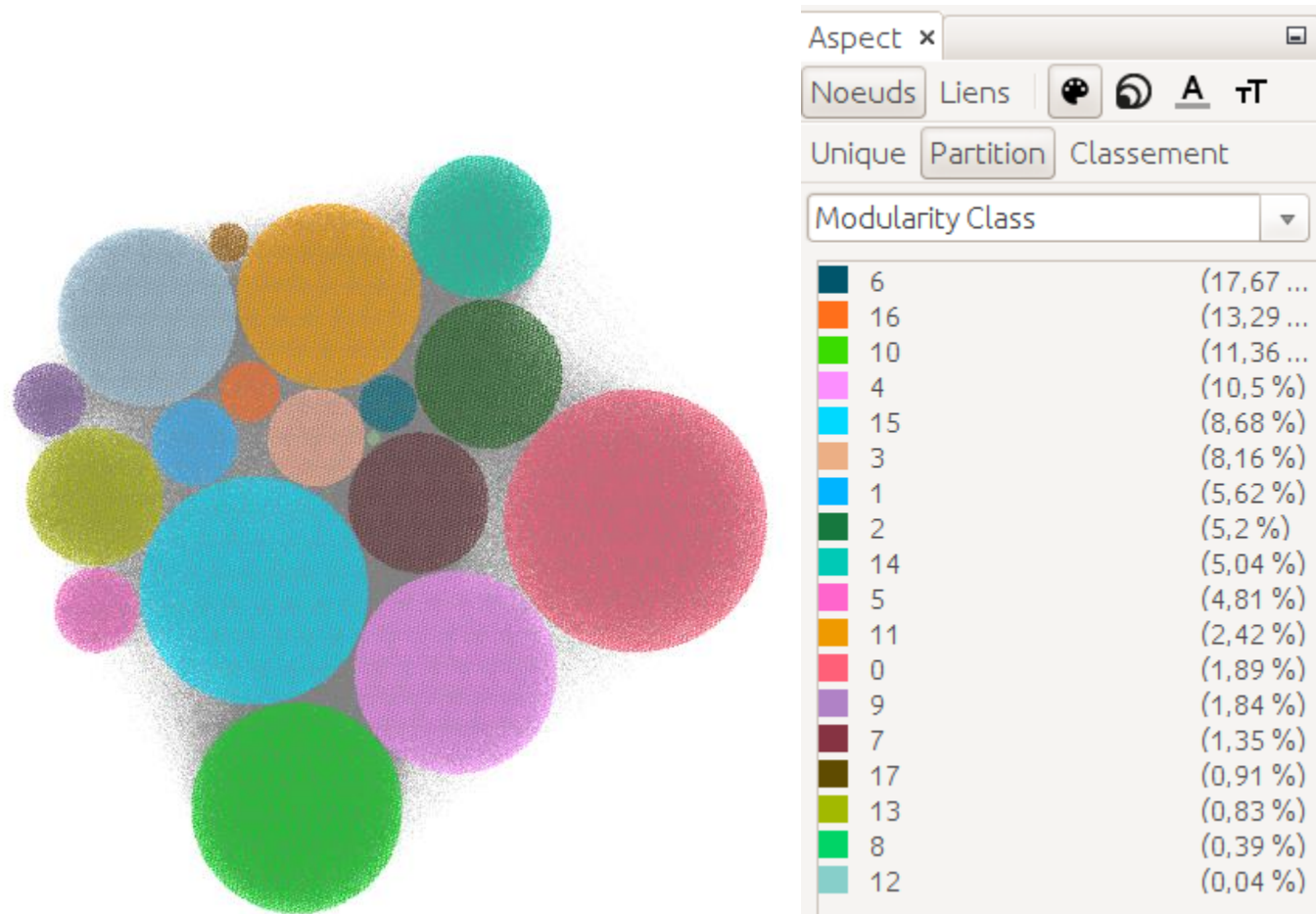


Figure 4 : Représentation des 18 communautés proportionnellement à leur taille

L'excentricité a été finalement retenue pour hiérarchiser les communautés et surtout voir si l'on pouvait vérifier le phénomène d'attraction préférentiel identifié plus haut sur le nombre d'ami.

#### Visualisation du réseau social

##### 1. Choix de l'algorithme de spatialisation

Selon Tutte une bonne représentation doit limiter le croisement des arcs. Etant donné le nombre important de ponts, une structuration hiérarchique s'imposait.

Le Layout Circle Pack a permis de repérer les communautés à l'aide de la coloration par classe de modularité mais cette représentation ne permet pas de se rendre compte de la taille de chacune d'entre elles. Le critère de hiérarchisation retenu ici est l'excentricité. Il donne une meilleure représentation des communautés par rapport à au phénomène de propagation.

##### 2. Résultat

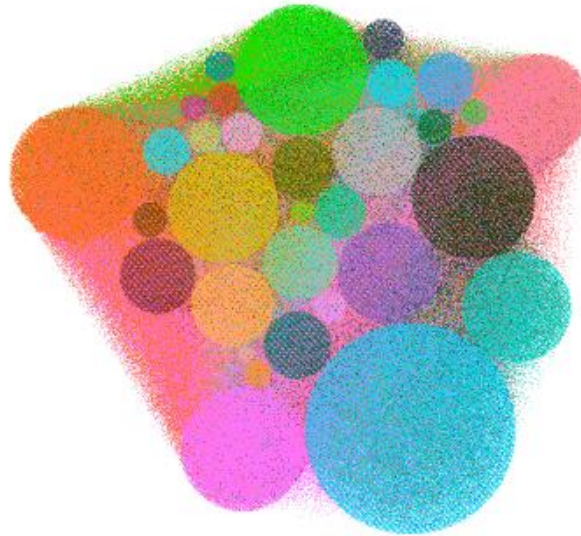


Figure 5 : Spatialisation des communautés Friendster et représentation de leurs affinités

## 7 - CONCLUSION

Les propriétés du réseau Friendster sont bien représentative de la structure d'un graphe de terrain de type petit monde:

- faible densité : oui
- fort clustering : oui
- faible distance moyenne : 7
- distribution de degré très hétérogène : réseau sans échelle (suit une loi de puissance)
- composante connexe géante : oui, une seule
- présence de communautés : oui

Les différents algorithmes consistant à supprimer des arêtes pour déconnecter le graphe et détecter les communautés, même en effectuant un parcours en largeur ne termine pas en un temps polynomiale. Les difficultés liés à l'espace mémoire nécessaire pour traiter le volume de données sur un ordinateur portable ainsi que les caractéristiques du graphe n'ont pas permis d'exécuter des algorithmes de parcours trop complexe.

## 8 - RÉFÉRENCES

1. Documentation Spark – Gestion dynamique de la mémoire : <https://spark.apache.org/docs/latest/>
2. Librairies Networkx <https://networkx.github.io/documentation/latest/>
3. GraphX : <https://spark.apache.org/docs/latest/graphx-programming-guide.html>
4. Repository GitHub pour les codes sources : <https://github.com/MLC06800/SourcesRCP216>