

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.....	2
	Objet.....	3
	Commentaires	3
	Résultat	3
6 -	VISUALISATION.....	7
	Phase exploratoire	Erreur ! Signet non défini.
	Visualisation du réseau social	7
7 -	CONCLUSION.....	8
8 -	RÉFÉRENCES.....	8

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

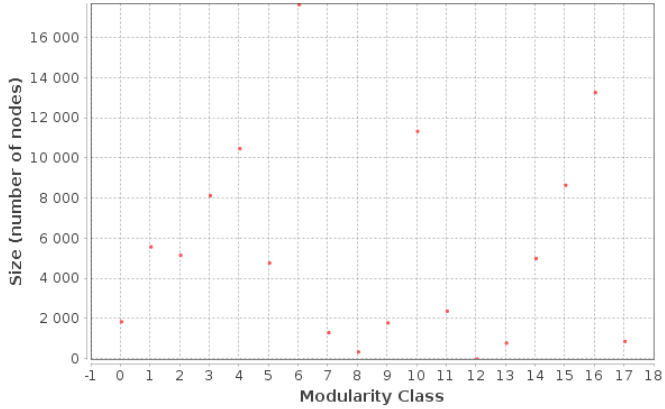
1. Plateforme Spark –version 2.2.0 (version java : « 1.8.0_171 »)
2. Notebook Jupyter --version 4.3.0
3. Kernels : Scala –version 2.11.8

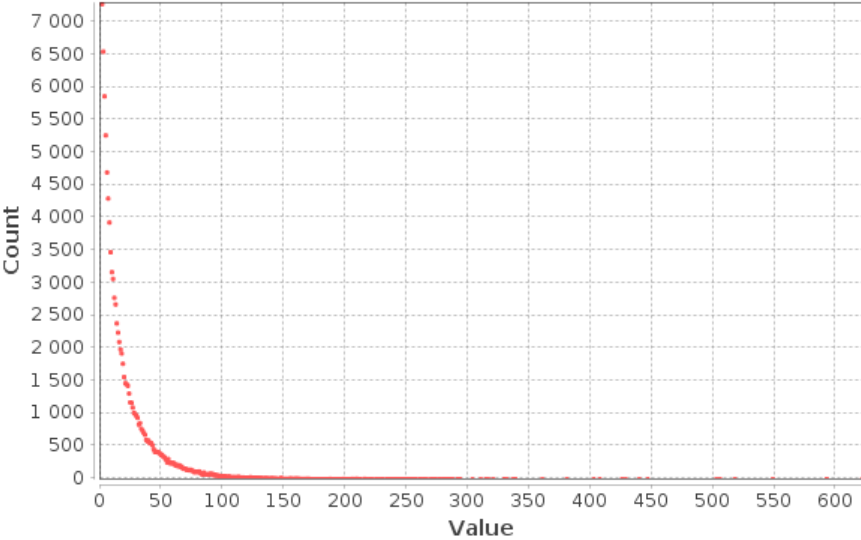
4 - DÉMARCHE

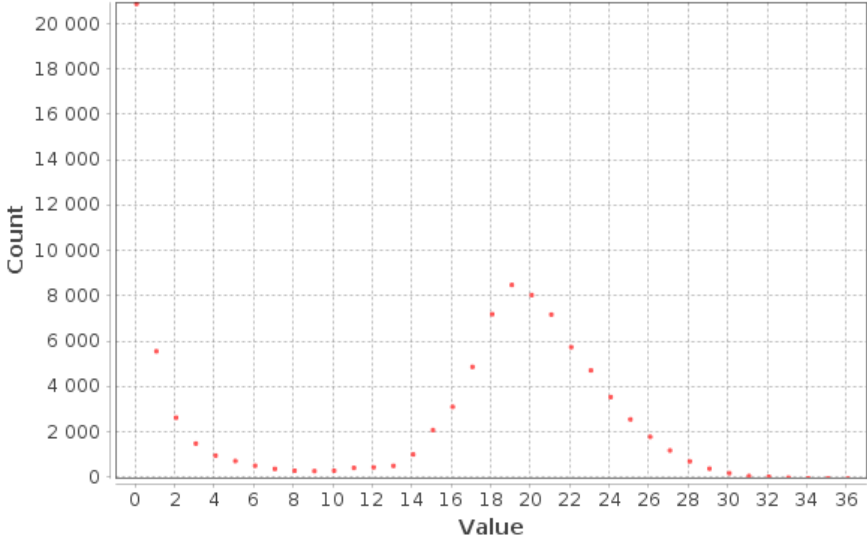
1. Création du sous graphe à partir des données
2. Analyse topologique de la structure globale du réseau
3. Analyse spectrale de la structure globale du réseau
4. Visualisation du sous graphe

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 représentants des relations différentes entre les utilisateurs. Le scrapping des données a sans doute rapatrié également des données sur le partage de contenu.</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>	<ul style="list-style-type: none"> Le graphe complet : Type: Graph Number of nodes: 5 689 498 Number of edges: 14 067 887 Average degree: 4.9452 Densité : 0 Nombre de cliques : 12 889 424 Test Erdős-Gallan => True Test method="hh" => True 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 Nombre de nœuds source distincts : Nombre de nœuds cible distincts :
Connexité	Le graphe est entièrement connexe. Il n'existe pas de nœud isolé. C'est une composante connexe géante.	<p>Nombre de composante connexe : 1 seule composante faiblement connectée.</p> <p>Même résultat trouvé par programme en spark.</p>

Objet	Commentaires	Résultat
Recherche de communauté	<p>Il existe bien des communautés car on trouve beaucoup d'ensembles de sommets très liés entre eux et peu liés vers le reste du graphe.</p> <p>La modularité calculée est cohérente avec la présence de communauté en comparaison à celle du graphe aléatoire utilisé comme « témoin ».</p>	<p>Nombre de communauté : 18</p> <p>Modularité: 0.691 (Paramétrage effectué pour une chercher ce niveau de précision. En augmentant la précision on peut détecter plus de 26 000 communautés)</p> <p>Size Distribution</p>  <p>Coefficient de clustering :</p> <ul style="list-style-type: none"> - résultat avec Spark : 26,75 % - moyenne des coefficient de clustering : 0,134
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>Validation de la séquence de degrés</p> <p>Selon les algorithmes de Erdős-Gallai et de Havel-Hakimi la séquebce est valide.</p> <p>Classements des utilisateurs par ordre décroissant de degrés :</p>

Objet	Commentaires	Résultat
	<p style="text-align: center;">Degree Distribution</p>  <p>La courbe est caractéristique des réseaux de petit monde.</p>	<p>(3687,621) ; (3552,592) ; (3570,592) ; (3553,548) ; (82,517) ; (3542,505) ; (212,502) ; (3674,446) ; (2680,439) ; (3651,428)</p>
Centralité	<p>La centralité d'intermédiation (betweenness-centrality) est étudiée ici. (nombre de plus courts chemins passant par un sommet)</p>	

Objet	Commentaires	Résultat
Vue générale des liens	<p>Eccentricity Distribution</p>  <p>The plot shows a distribution of eccentricity values. The x-axis represents the 'Value' (eccentricity) from 0 to 36, and the y-axis represents the 'Count' of nodes from 0 to 20,000. The distribution is unimodal, peaking at a value of 19 with a count of approximately 8,500. There is a significant outlier at value 0 with a count of over 20,000.</p>	Distance du parcours moyen : 7.3
Matrice d'adjacence		
Liste d'adjacence		
Comparaison avec un graphe aléatoire	<p>Utiliser un modèle uniforme, en se donnant n sommets et en décidant que chaque arête a une probabilité $0 < p < 1$ d'exister (modèle de Gilbert, noté $G(n,p)$).</p> <p>On observe alors en général une composante connexe géante, un diamètre proche de $\log(n)$, un clustering très faible, une</p>	

Objet	Commentaires	Résultat
	distribution de degrés homogène, pas de structure communautaire (cf liste précédente sur les réseaux réels).	
PageRank	J'ai supposé que les arêtes sont orientées (1 est « followers » de 2). J'ai donc construit un graphe de « followers » pour appliquer un l'algorithme PageRank. On trouve 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.	

6 - VISUALISATION

Analyse statistiques

1. Histogrammes
2. Insérez votre texte ici

Visualisation du réseau social

1. 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 3giga octet). Seconde tentative réussit avec le graphe analysé plus haut.
2. Choix de l'algorithme de spatialisation
L'algorithme de Yifan Hu est basé sur les forces et l'analyse multi-niveaux. C'est un algorithme d'une complexité

d'ordre $O(N \cdot \log(N))$, ce qui tout à fait approprié étant donné la taille du graphe à traiter (100199 nœuds). De plus cet algorithme terminant en un temps polynomial il s'arrête tout seul. Les paramètres utilisés sont les suivants :

- ▶ Ratio du pas (pour une qualité optimale au détriment de la vitesse) : 0.95
- ▶ Distance optimale (pour une bonne séparation sans trop étaler étant donné le nombre imposant de nœuds à afficher) : 100
- ▶ Thêta de l'algorithme de Barnes-Hut (donne une bonne précision) : 1.2

3. Résultat (obtenu après une très longue attente)

7 - CONCLUSION

Les propriétés du réseau Friendster sont assez différentes du graphe aléatoire généré :

- faible densité : oui
- fort clustering : oui
- faible distance moyenne : <à faire>
- distribution de degré très hétérogène : oui
- composante connexe géante : 1 oui
- présence de communautés : oui

La structure de petit monde se vérifie. Il existe une structure distribuée globalement et localement centralisé sur quelques individus plus populaires.

8 - RÉFÉRENCES

1. Documentation Spark – Gestion dynamique de la mémoire :
2. Librairies Networkx ; GraphX
3. Repository GitHub pour les codes sources : <https://github.com/MLC06800/RCP216>