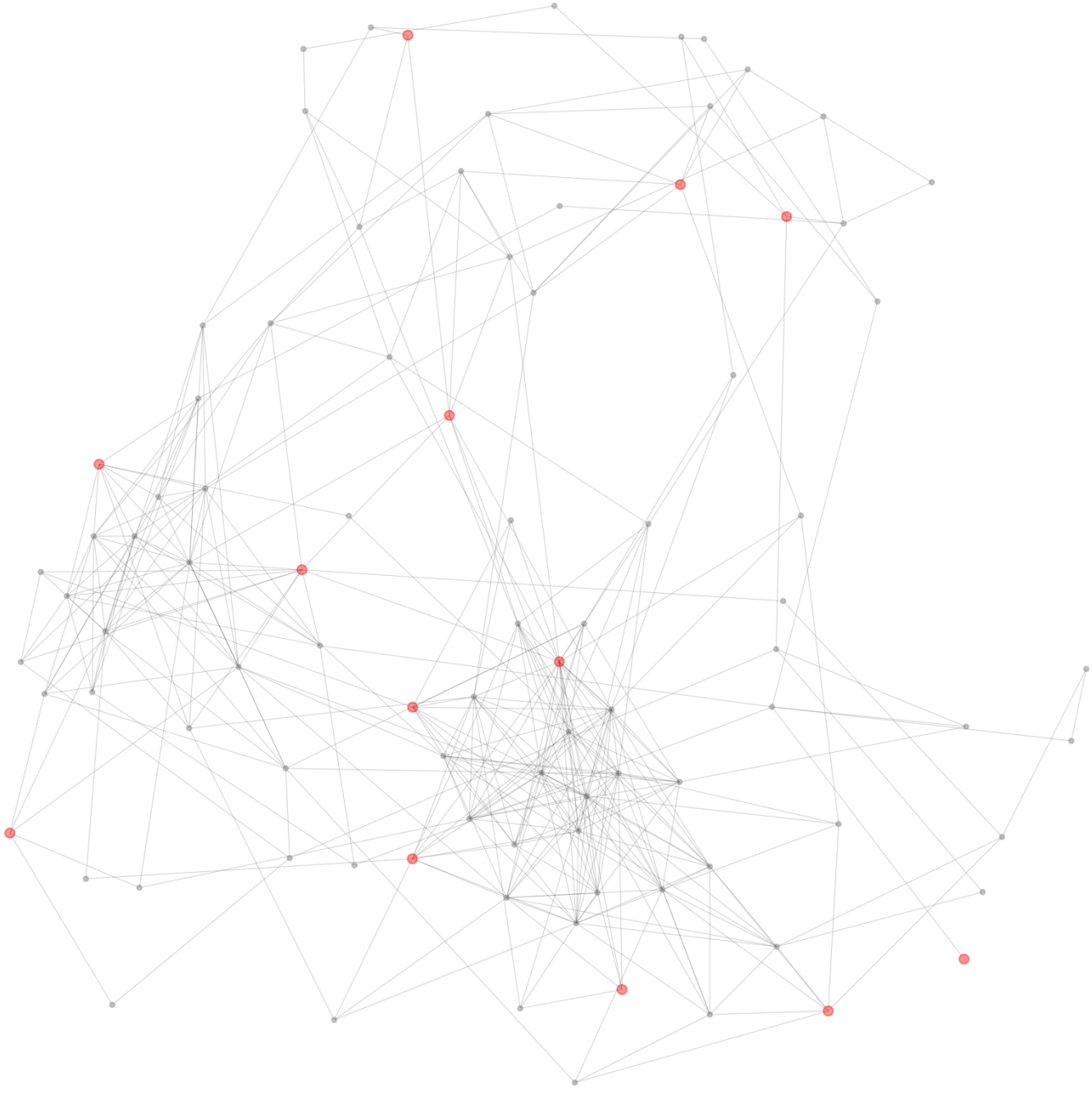
Maxence BROCHARD
Duc-Vinh TRAN
Samy BOUCHNAIF

Pour le 19/05/2017

Rapport de projet SE334

Statistical Analysis of Network Data with applications in Marketing

Viralité des hashtags dans un réseau social



Propriétés du document :

Classification	Publique -
Version	Version 1.0
Auteurs	Maxence BROCHARD maxence.brochard@ensae-paristech.fr Duc-Vinh TRAN duc.vinh.tran@ensae-paristech.fr Samy BOUCHNAIF samy.bouchnaif@ensae-paristech.fr
Responsable Encadrants	Professeur Eric D KOLACZYK kolaczyk@bu.edu
Pages	17

Note de rédaction :

Ce rapport a été rédigé dans le cadre du projet qui sanctionne le cours « *Statistical Analysis of Network Data with applications in Marketing* » de l'ENSAE ParisTech. Il constitue la synthèse du travail réalisé et sert également de support de compréhension pour le livrable fourni à l'adresse suivante : https://github.com/dvp-tran/Statistical_Graph

N'hésitez pas à nous solliciter pour tout complément d'information. Nous restons à votre disposition pour de plus amples informations, ainsi que pour d'éventuelles remarques et suggestions qui pourraient notamment nous permettre de rectifier et compléter notre travail.

Les contributeurs de ce projet sont Maxence BROCHARD, Duc-Vinh TRAN, et Samy BOUCHNAIF. Nous souhaitons par la présente remercier le Professeur KOLACZYK pour son encadrement, son expertise, ses conseils, et sa grande disponibilité qui nous ont permis de mener à bien ce projet.

Précautions :

- Le code a été réalisé en langage Python, et le livrable prend la forme d'un Notebook Jupyter. N'hésitez pas à nous solliciter si vous souhaitez récupérer le fichier brut de code Python.

- De nombreuses libraires ont été utilisées pour ce projet. Il vous faudra préalablement les installer sur votre machine avant de pouvoir charger ces libraires et faire tourner le code.

- Les calculs ont été réalisés sur un ordinateur à la configuration suivante :
  - Operating System : Ubuntu 17.04
  - Processeur : i7-7700HQ (2.8Ghz*4)
  - RAM : 32 Go

- Par abus de langage, nous dirons à plusieurs reprises : utilisateurs/individus infectés pour hashtag lorsque celui-ci à fait usage (tweet, retweet, mention) de ce hashtag. Il est également possible de parler d'adoption du hashtag.

Sommaire

I. Introduction - contexte et motivations :

La viralité est, dans les réseaux sociaux, une question importante pour les entreprises, les campagnes politiques et les personnes influentes, car elles déploient énormément de ressources et d'efforts pour rendre leurs produits ou messages viraux afin de capter l'attention et de propager/diffuser leur influence/activité à un public plus large.

Ainsi, la compréhension du mécanisme complexe de la viralité peut aider à contrôler ses effets sur le réseau :

- Comment la structure du réseau affecte-t-elle la diffusion ?
- Comment modeler la contagion, etc. ?

Pour répondre à ces questionnements, nous nous sommes penchés sur ce papier : https://arxiv.org/abs/1403.6199

L'article part de l'idée générale que les communautés de réseau permettent de prédire la viralité par son modèle de diffusion précoce. Une approche simple et populaire dans l'étude de la diffusion des hashtags est de considérer les hashtags comme des maladies et d'appliquer des modèles épidémiques. Cependant, des études récentes démontrent que les maladies et les comportements diffèrent différemment.

Nous pouvons voir une énorme potentialité pour les applications dans le marketing des réseaux sociaux : les réseaux sociaux pourraient donner de meilleurs conseils à leurs utilisateurs quant aux publications susceptibles de donner le meilleur retour sur investissement.

Nous avons voulu étudier de façon plus spécifique la propagation des hashtags sur le réseau social *Twitter*, en nous inspirant du dataset "Sampled public tweets from Twitter streaming API between March and April 2012" présent ici : http://carl.cs.indiana.edu/data/#virality2013

Dans un premier temps, nous allons tenter de modéliser la propagation d'un hashtag dans un réseau et sa viralité à travers des structures classiques de graphes aléatoires, et étudier les différents types de propagation.

Ensuite, nous jetterons un oeil sur le jeu de données et expliquerons que de par son volume et la taille des graphes, une étape préliminaire de sampling sera nécessaire avant d'analyser les graphes. Nous verrons comment la phase de sampling influencera nos études et comment les prendre en compte pour la modélisation.

Puis, nous nous pencherons sur une modélisation rapide de la propagation dans un sous-réseau (évolution de la propagation, son intensité quotidienne en fonction du nombre de malades, etc). Dans le cadre d'un réseau infecté, nous regarderons également les features et le comportement du réseau ; l'idée étant d'extraire des features utiles dans le but de faire de la prédiction (machine learning).

II. Analyse du jeu de données et modélisation de la propagation d'un hashtag dans un réseau :

Cette partie traitera des différentes questions que nous nous sommes posées au tout début de ce projet. Elle concernera notamment les différentes difficultés rencontrées lors des chargements de données ou encore sur les temps de calculs et comment nous avons tentés de résoudre ces problèmes.

1. Analyses simples du graphe complet

Il est intéressant de débuter notre étude par une analyse simple du graphe complet. A partir des données des relations Tweeter entre les utilisateurs, nous avons réussi à établir un graphe non dirigé des utilisateurs. Ce graphe est immense et compte exactement :

- 595 460 Nœuds
- 14 273 311 Edges
- degré moyen : 24 environ.

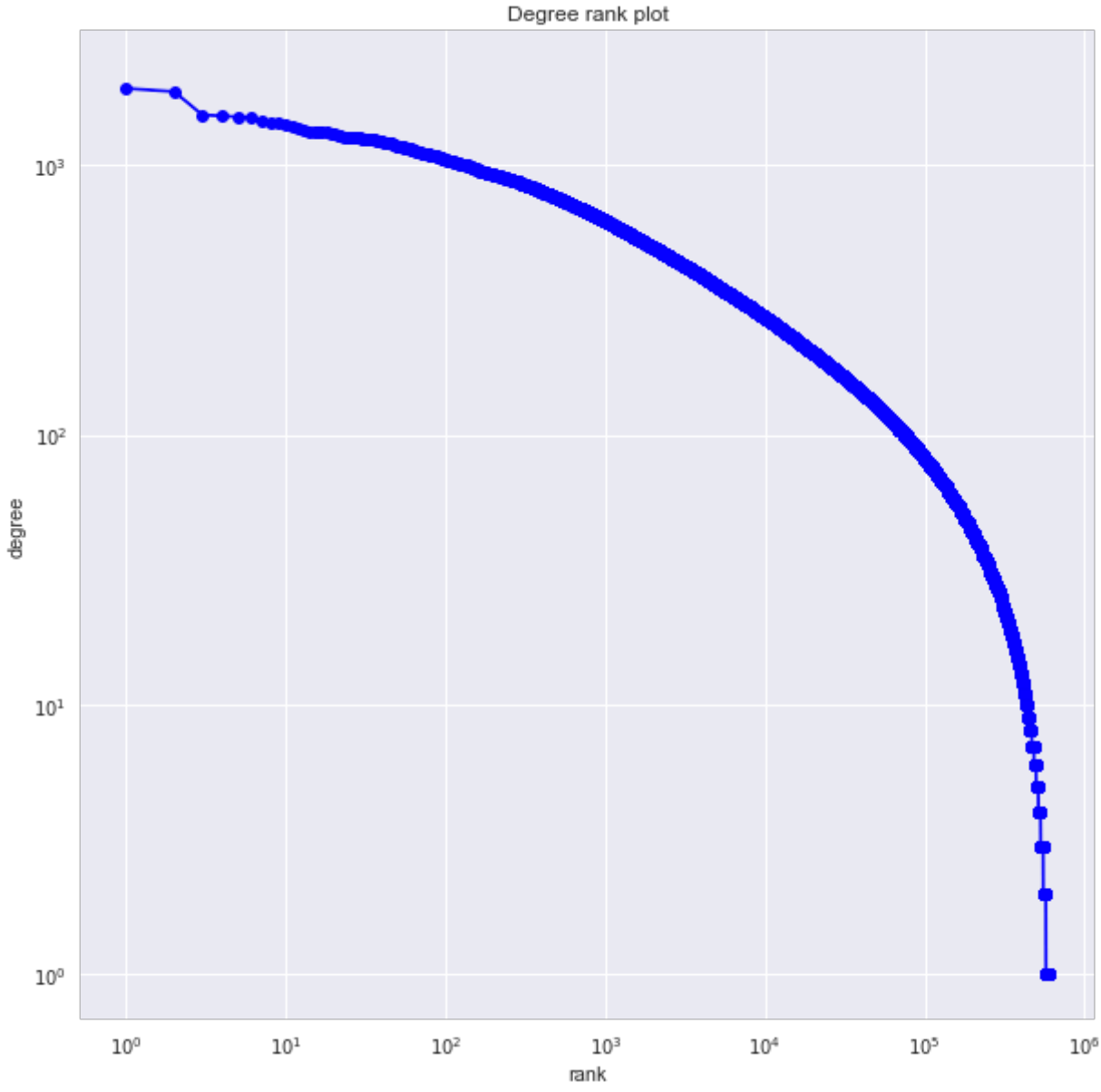


Fig. 1 : Distribution des degrés au sein du graphe complet

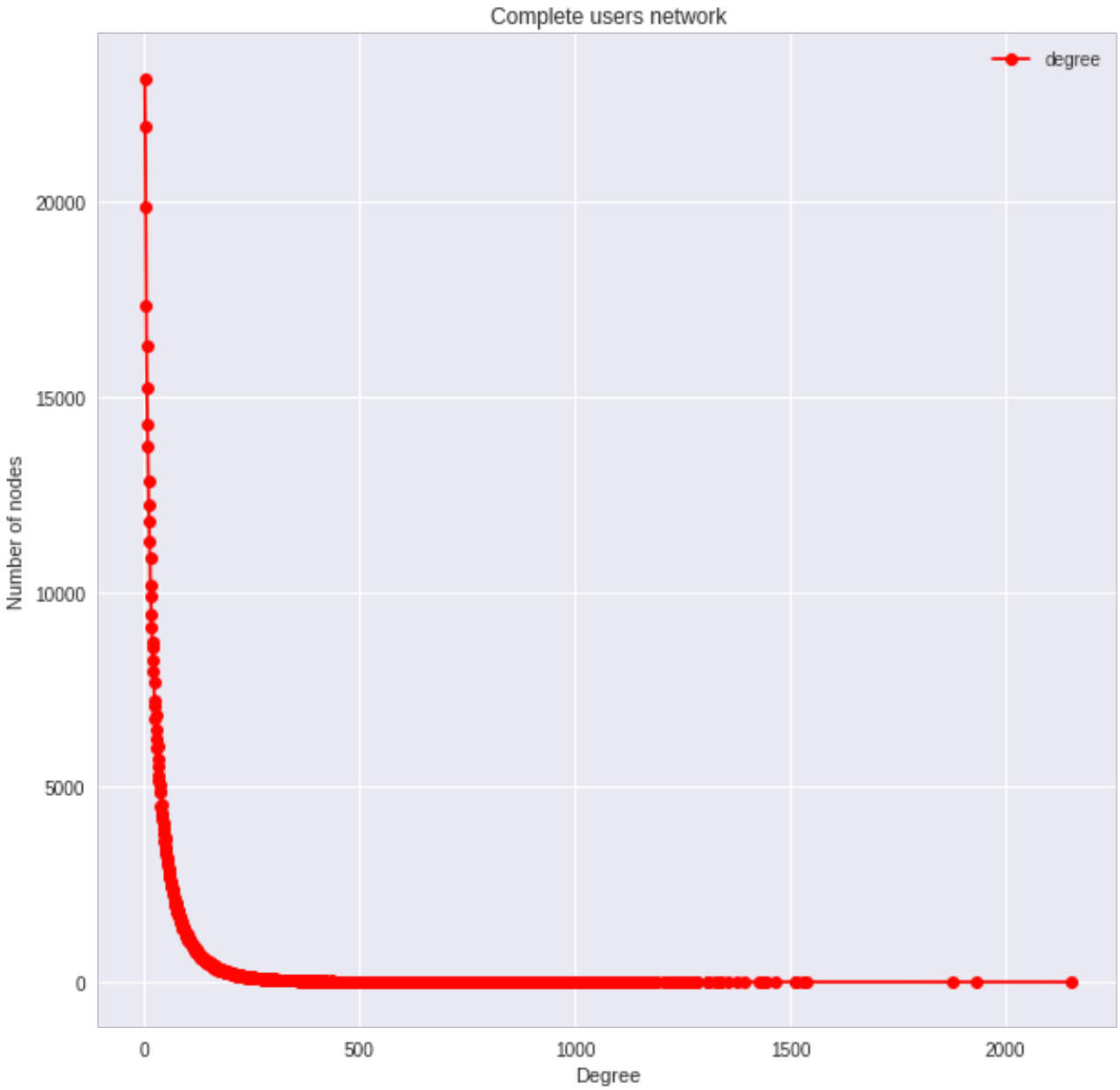


Fig. 2 : Distribution des degrés au sein du graphe complet

On constate que la majorité des analyses descriptives telles que le clustering sur un graphe de cette taille ne sont pas envisageables pour un ordinateur domestique. Il en est de même pour les représentations graphiques du graphe complet. C'est pourquoi il est intéressant de se poser la question sur comment restreindre notre étude sur un sous ensemble du graphe complet.

2. Extraction d'un sous-graphe

Nous allons tester ici différentes méthodes :

- Approche naïve
- Approche extraction des voisins directs
- Approche de plus courts chemins entre tous les points

Notre critère de validation de la méthodologie serait la conservation des propriétés globales du réseau, c'est à dire une proportionnalité sur la distribution des degrés dans le sous graphe.

L'approche naïve consiste tout simplement à extraire le sous-graphe à partir de seulement les nœuds infectés par un hashtag donné.

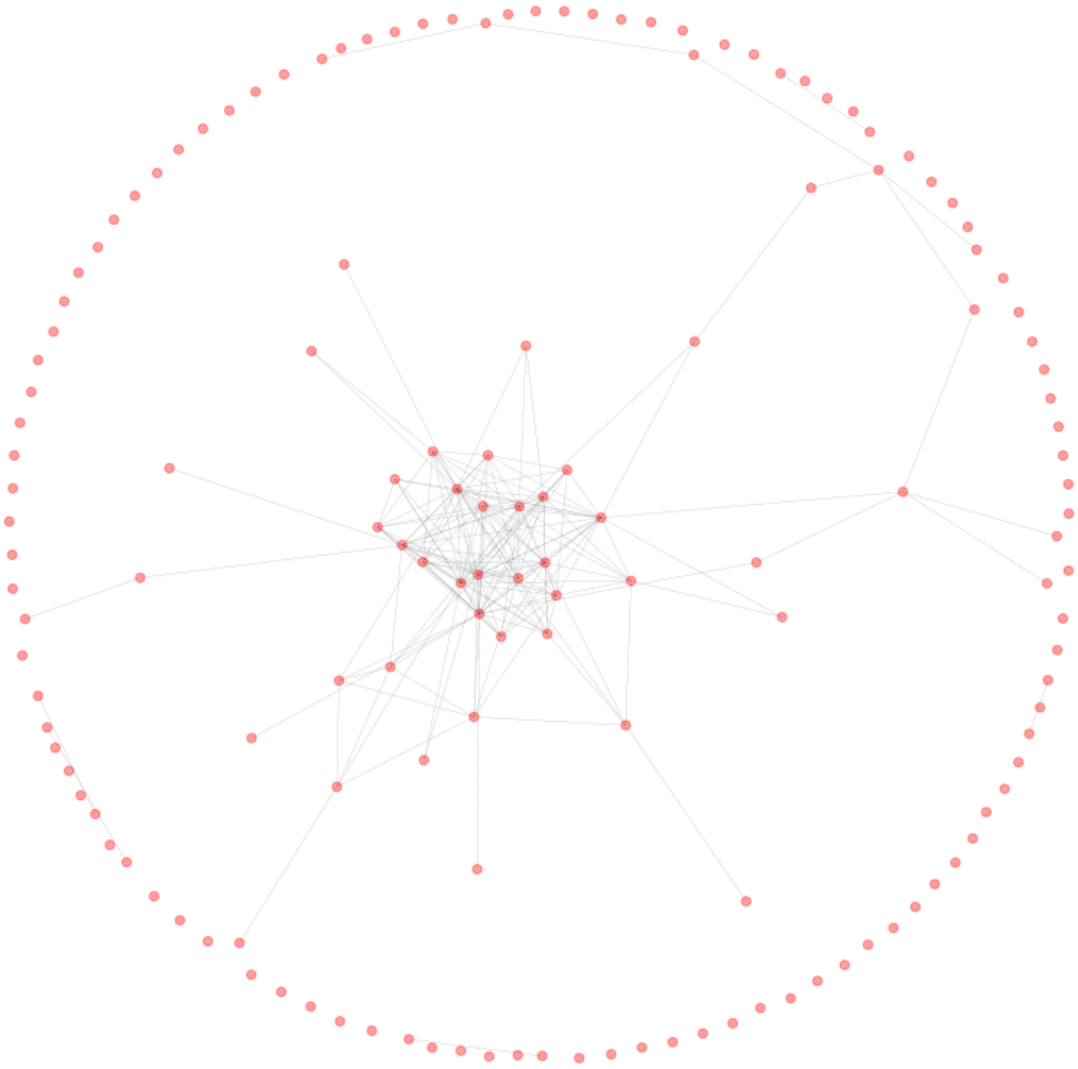


Fig. 3 : Représentation de la méthode naïve de sous-graphe, hashtag : Madonna

On constate que pour cette approche naïve, de nombreux points se retrouvent en périphérie car non connectés au reste du réseau signifiant que soit ces utilisateurs utilisent spontanément hashtag sans influence de la part du réseau (donc sont des points d'origines de la propagation), soit il manque des informations vis à vis du réseau d'infection. Ce dernier point est une hypothèse probable car le flux API streaming de Tweeter ne capte seulement que 10% du flux, il est donc possible que d'autres points d'infections n'apparaissent pas car non récupérés par l'API.

Cela impliquera que, par la suite, nous chercherons à prédire plutôt le nombre d'infectés du réseau plutôt que l'infection ou non-infection d'un seul nœud.

L'approche d'extraction des voisins directs consiste à :

- Identifier les utilisateurs infectés pour un hashtag donné
- A partir du graphe complet, extraire le sous graphe composé de tous les voisins directs pour chaque individu identifié juste ci-dessus.

Ainsi nous obtenons un sous-graphe avec des propriétés particulières telles que des composantes sous formes de « grappes » :

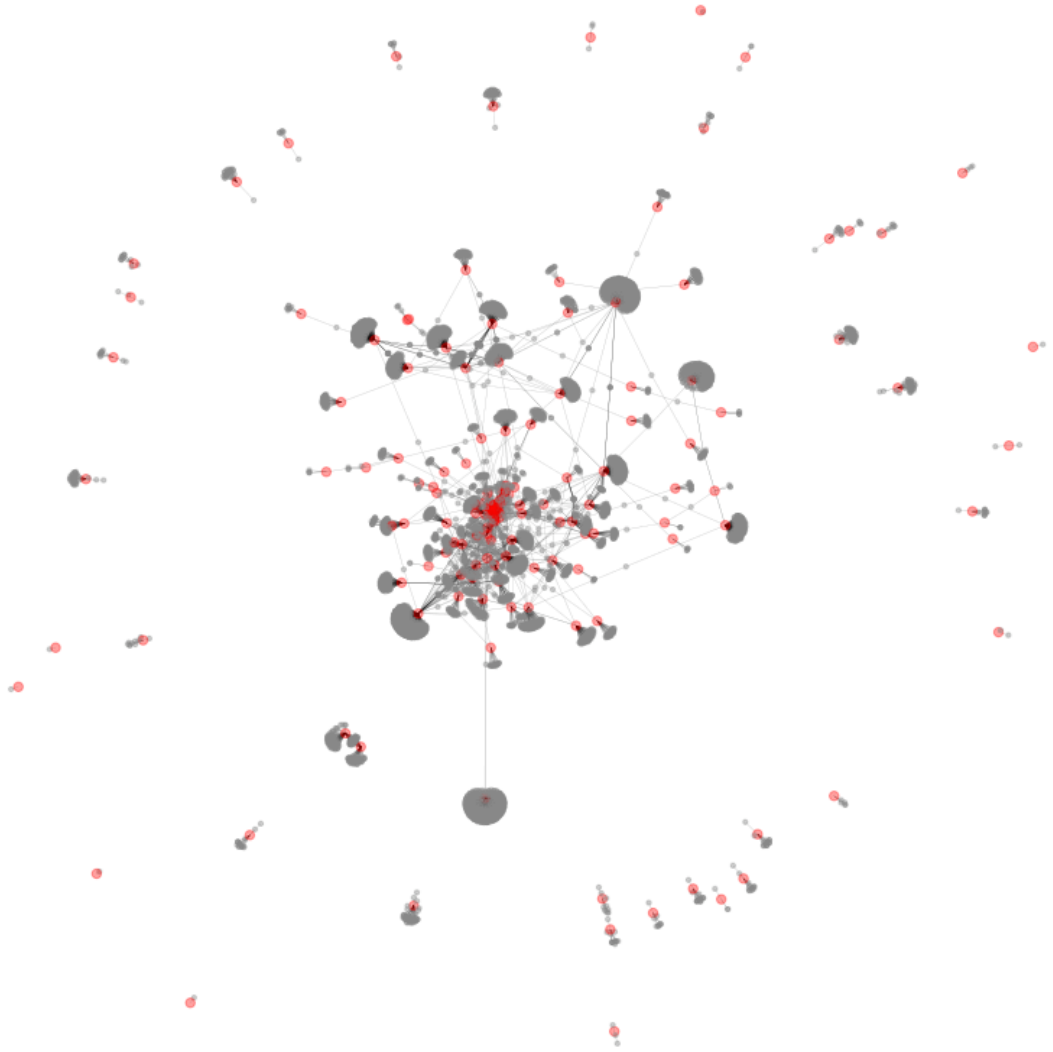


Fig. 4 : Représentation du sous réseau par voisins directs, hashtag : Madonna

Directement par observations, nous pouvons constater que cette extraction de sous graphe ne peut convenir :

- La distribution des degrés est bien plus erratique
- La formation des grappes indépendantes, détachées du reste du sous-réseaux : nombreux nœuds infectés par un hashtag (en rouge) se retrouve dissociés des autres.

Quant à **l'approche d'extraction du sous-graphe par plus courts chemins** entre tous les points, la méthodologie est la suivante :

- Identifier les individus infectés, supposons qu'il y a N-nombre d'individus concernés
- Réaliser tous les combinaisons de paires d'individus possibles : $\binom{N}{2}$ paires possibles.
- Pour chaque paire, rechercher le plus court chemin séparant les deux points de cette paire.

Pour accélérer le processus qui est long nous avons parallélisés les calculs des plus courts chemins sur les différents cœurs du processeur. Néanmoins le temps de calcul reste très coûteux.

Nous obtenons certes un graph bien plus dense et fortement interconnecté, mais bien plus représentatif :

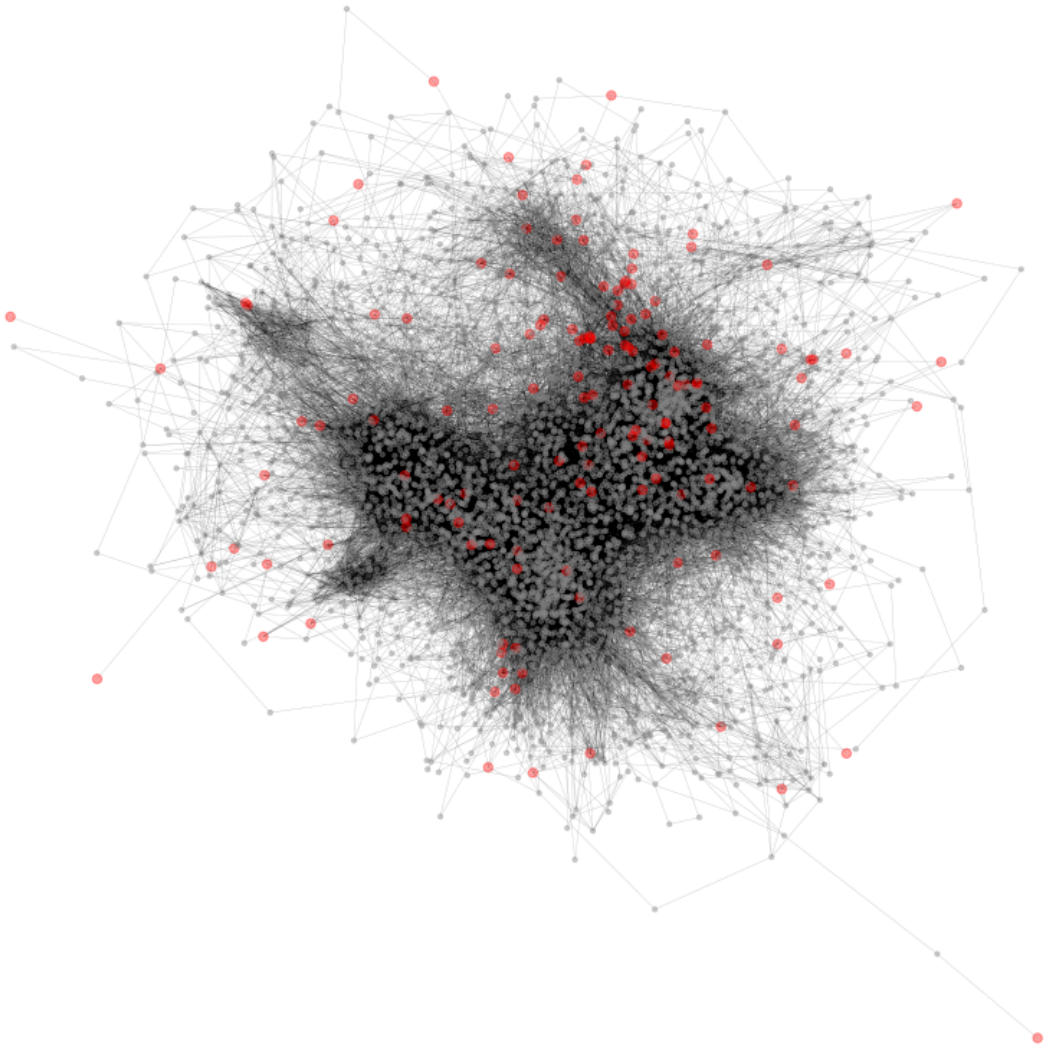


Fig. 5 : Représentation du sous réseau par plus court chemin, hashtag : Madonna

Concernant la distribution des degrés d'un tel sous-graphe, le résultat est bien meilleur. On conserve une distribution proche du graphe complet à un facteur de proportionnalité près. :

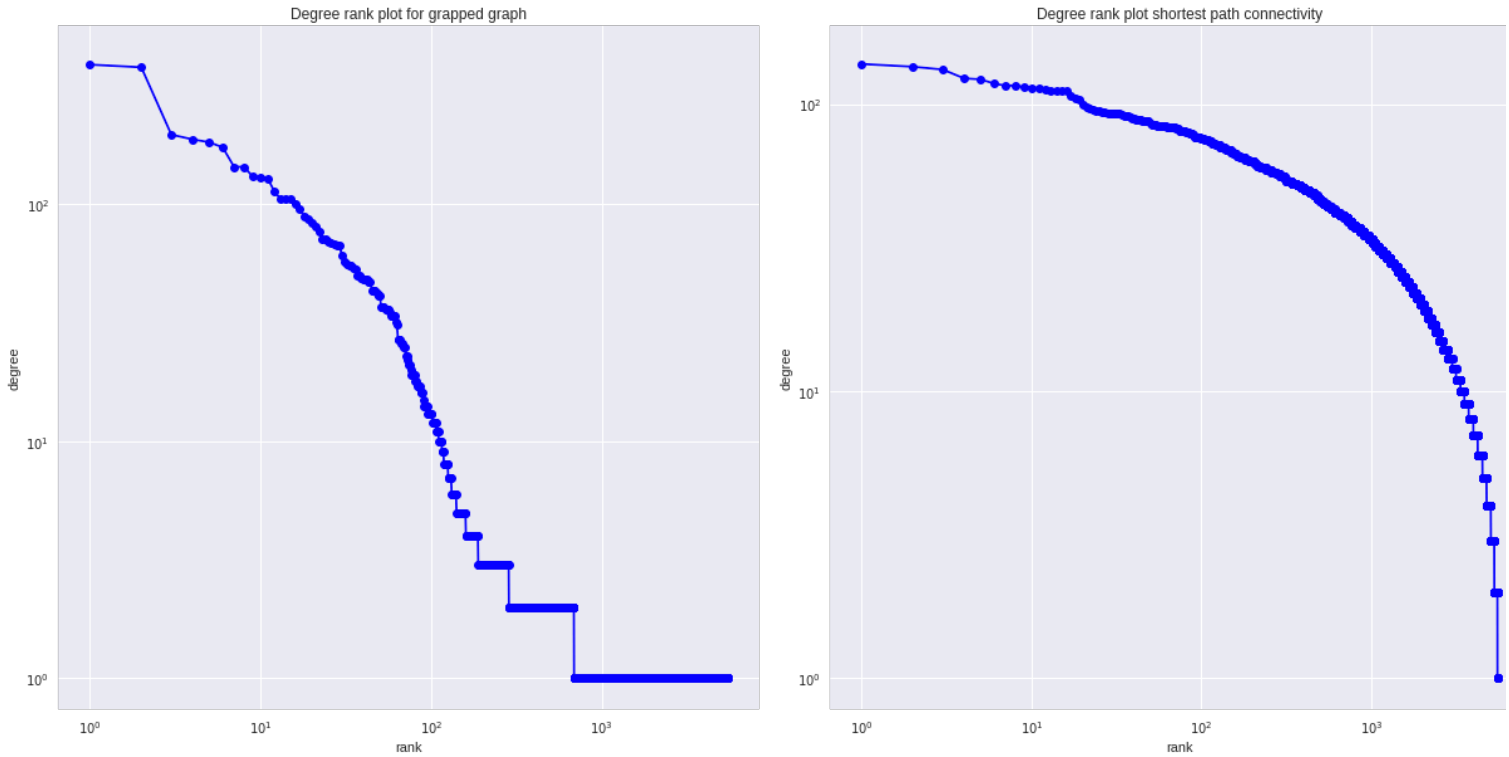


Fig. 6-7 : Comparaison des distributions des degrés, gauche : voisins direct, droite : shortest path

Par la suite, nous allons nous intéresser au « pouvoir interprétatif » de cette représentation de sous-graphes en essayant d'identifier des communautés de population. Pour cela, nous considérons trois différents hashtags : #news en GRIS, #foxnews en BLEU, et #madonna en JAUNE. A partir de ces hashtags, on extrait le sous-graphe associé et on les représente ensemble :

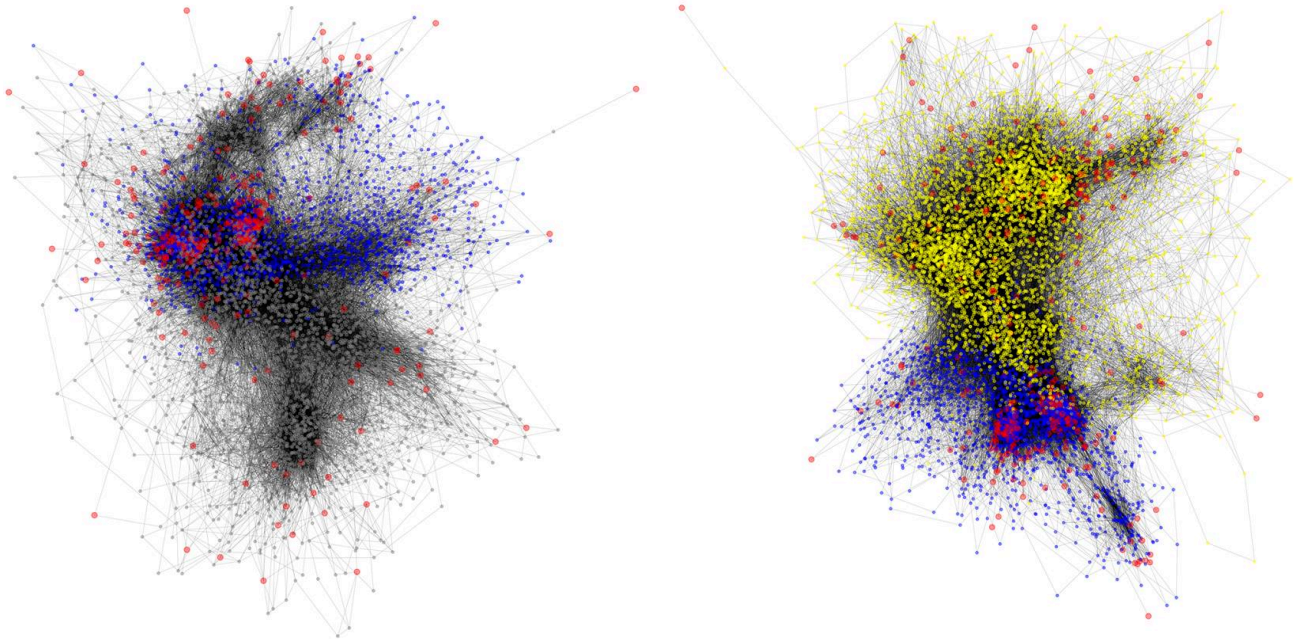


Fig. 8-9 : Relation entre les différentes communautés de population infectée

On constate qu'il y a en effet un réel intérêt à cette représentation des sous-graphes : en effet, des hashtags portant sur des sujets similaires tels que *foxnews* et *news* sont liés de façons inclusives, c'est à dire que la communauté *foxnews* est bien comprise dans celle de news. Contrairement à la communauté *madonna* qui se distingue bien de celle de *foxnews* car ils traitent des sujets indépendants.

3. Modèle de propagation

Les données concernant la propagation des hashtags et leur adoptions par les utilisateurs sont générées à partir du fichier : *timeline_tag.anony.dat*, une ligne de se fichier correspondant à un hashtag donné :

Hahstag | timestamp1,user1 |timestamp2,user2 | etc.

On a récupère ainsi 1 345 913 différents hashtags pouvant infectés de 2 à 363 519 individus.

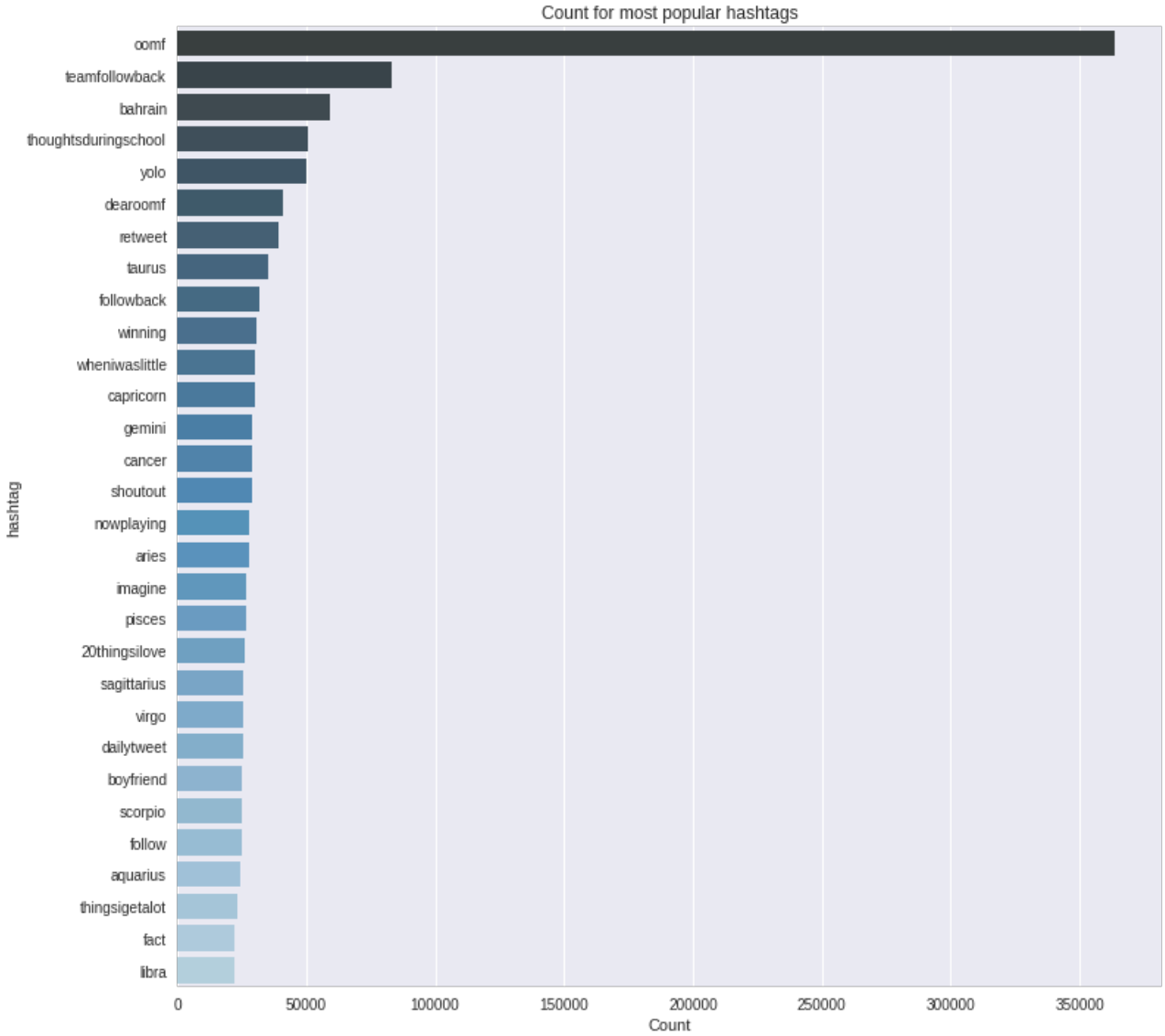


Fig. 10 : Top 30 des hashtags les plus virulents

Grâce au timestamp, il est possible d'avoir une résolution temporelle très fine, nous avons décidés d'étudier l'évolution du nombres d'infectés par jour. Pouvant ainsi déterminer des pics d'intensités liées à une date particulière.

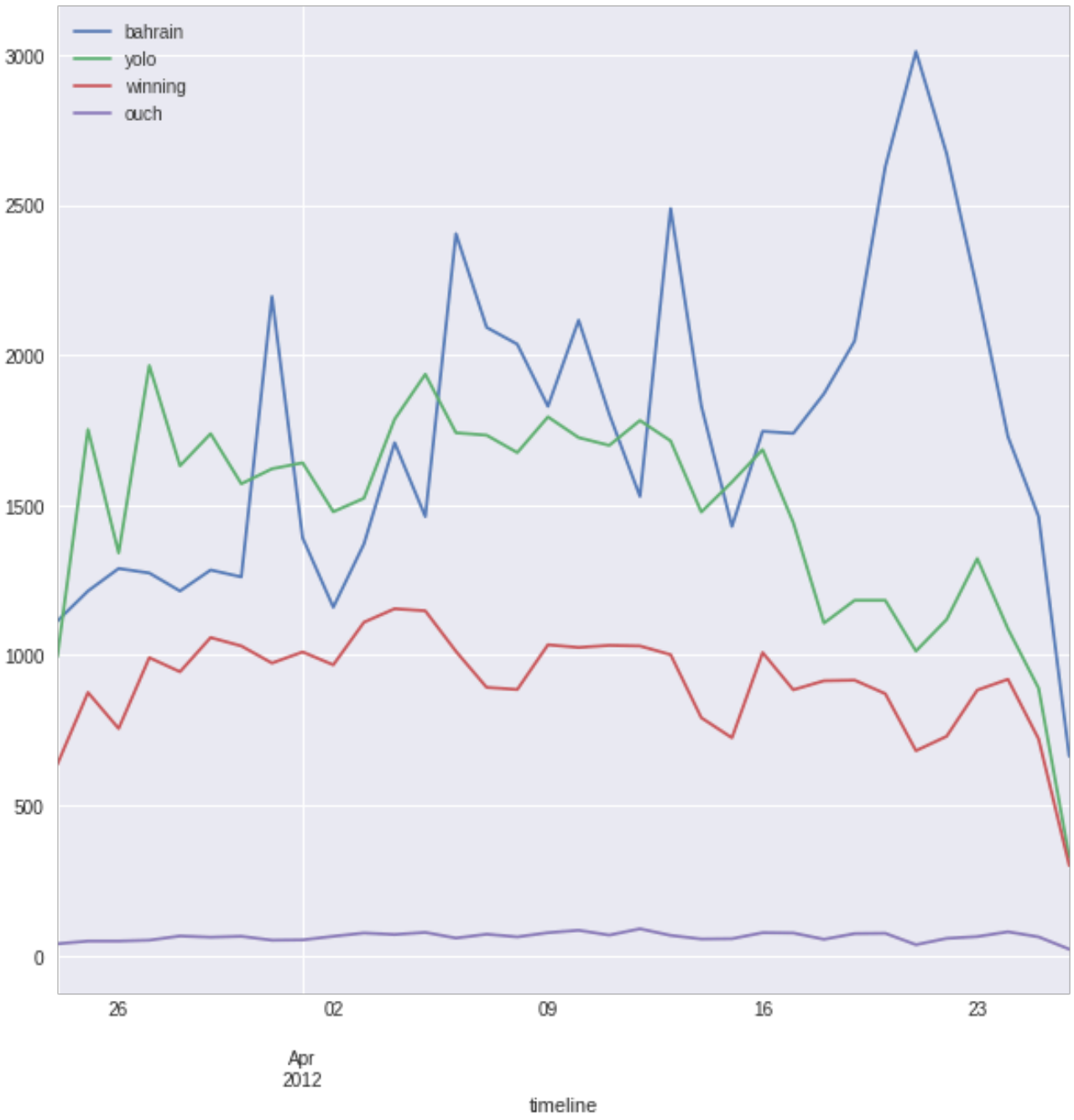


Fig. 11 : Nombre de nouveaux infectés par jour pour différents hashtags

On constate, par exemple, un fort pic le 20 avril 2012 pour le #bahrain, en effet il s'agit du moment où a eu lieu des affrontements à Bahrain à deux jours du grand prix de Formule 1, en effet la région était sous tension en raison de nombreuses révoltes au cours de cette période de printemps 2012 (essentiellement des étudiants qui voulaient sonner la révolte dans leur pays face à leur gouvernement ; pour la petite histoire).

On constate pour les hashtags peu virulents et très courants tels que #ouch, la propagation se réalise à vitesse constante. Comme nous pouvons l'illustrer avec le graphique des infections cumulées ci-dessous. Cela nous permet d'avoir l'intuition que virulence d'un hashtag, à comprendre sa puissance à se propager et à être adopté par des nouveaux utilisateurs, peut être déterminé par les premiers jours d'observations.

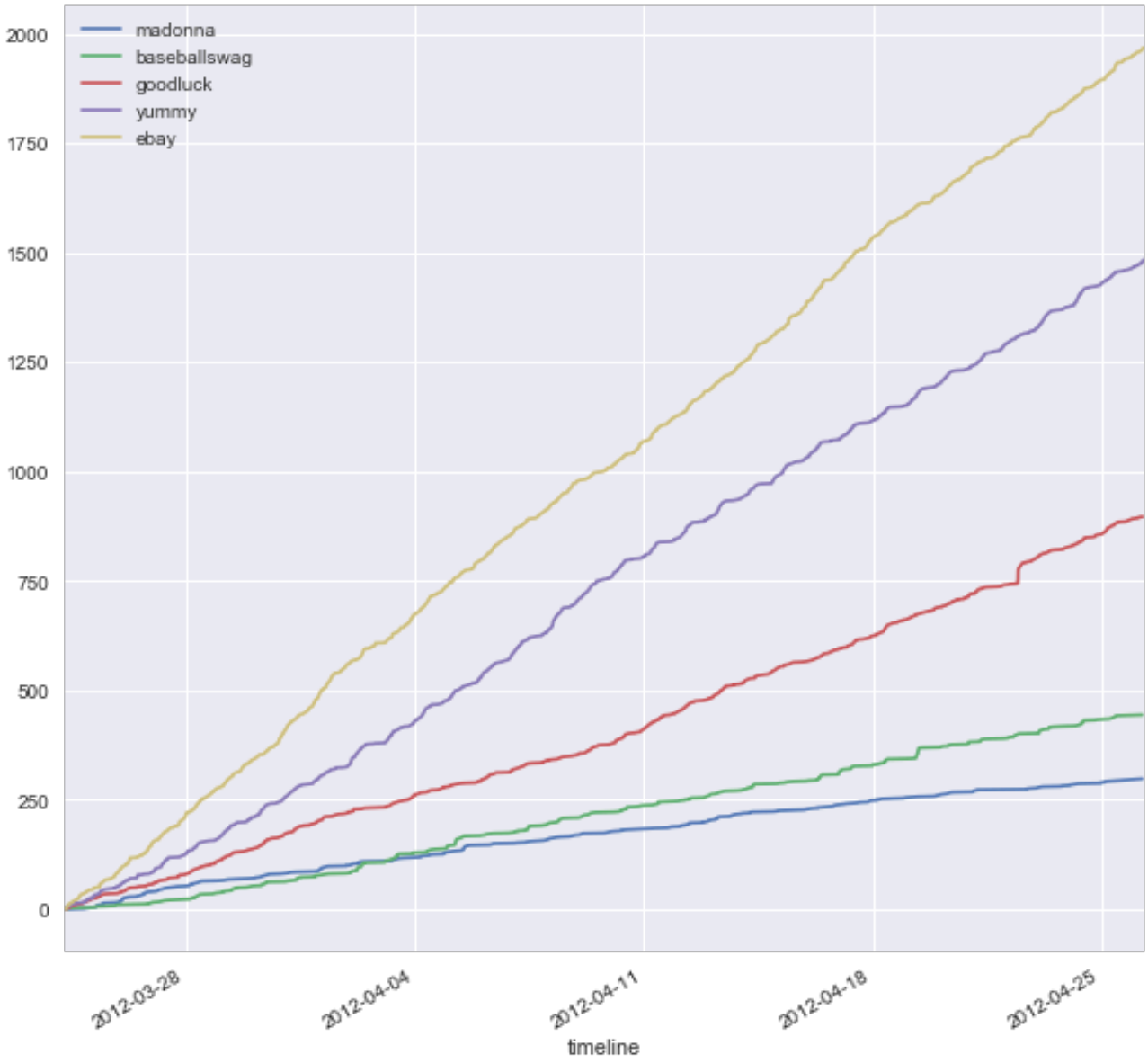


Fig. 12 : Nombre cumulé d'infectés par jour pour différents hashtags

Pour modéliser la propagation nous avons tout d'abord tenté de réaliser un modèle d'infection de proche en proche avec une probabilité fixe, c'est-à-dire que pour un nœud infecté, on observe ces voisins directs et parmi ces voisins on réalise un tirage avec probabilité fixe, ces derniers seront les nouveaux nœuds infectés.

On en déduit qu'il s'agit d'un processus dit explosif, car plus le nombre d'infectés augmente plus le nombre de nouveaux infectés augmente :

- Le graphique supérieur représente le taux d'infection par jour (nouveaux infectés par jours)
- Le graphique inférieur représente l'évolution de l'infection sur le réseau

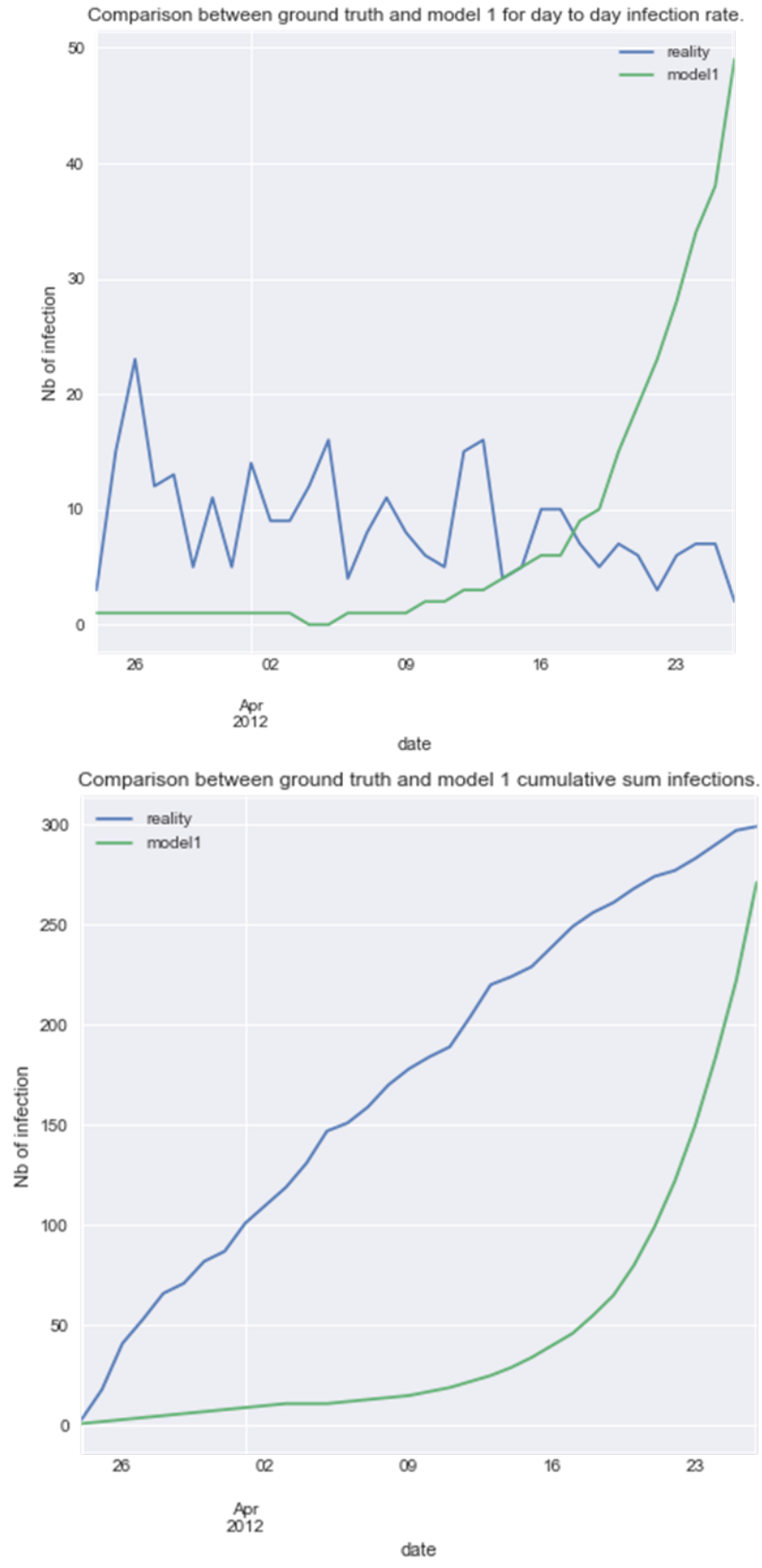


Fig. 13-14 : Profil d'infection pour le hashtag Madonna sur 34 jours

III. Construction et extraction des différents features :

Dans le but de réaliser la prédiction de la viralité d'un hashtag, il nous faut construire les différentes features qui nous permettront de caractériser cette viralité. Nous avons pour cela choisi plusieurs caractéristiques dont certaines issues de Weng, Menczer, Ahn (2014). Nous avons donc retenu plusieurs variables qui décrivent :

- La structure du réseau
- Les communautés au sein de ce réseau
- La croissance du réseau

Nous allons décrire la construction des différentes variables utilisées dans notre modèle dans la partie suivante.

1. La structure du réseau

a. Nombre d' « early-adopters »

Cette caractéristique correspond au nombre d'adopters ayant utilisé un hashtag *h* parmi les *n* premiers tweets :

$$|A_n(h)|$$

h : hashtag
n : nombre de premiers tweets
A : adopters

Si cette valeur est faible, cela indique qu'un petit nombre d'utilisateurs a généré la majorité des tweets. D'après Weng, Menczer et Ahn (2014), un petit nombre de premiers adopters indique un faible potentiel de viralité du hashtag en question.

b. Taille de la première surface

Ce sont tous les voisins des premiers adopters qui n'ont pas été infectés au cours des *n* premiers tweets. Ils correspondent aux potentiels adopters suivants.

$$\left|S\left(A_n(h)\right)\right|$$

c. Taille de la seconde surface

La taille de la seconde surface correspond à tous les voisins non infectés de la première surface :

$$\left|S\left(A_n(h)\right)\right|$$

d. Distance moyenne

Pour une séquence d'adopters $\langle a_1^h, a_{2,}^h \ldots a_n^h \rangle$ des *n* premiers tweets d'un hashtag *h*, on mesure le chemin le plus court $d(a_i^h, a_{i+1}^h)$ entre chaque couple d'adopters a_i^h et a_{i+1}^h avec *i* $1 < i <$ *n-1* :

$$\overline{d_n(h)} = \frac{1}{n-1} \sum_{i=1}^{n-1} d(a_i^h, a_{i+1}^h)$$

e. Diamètre

Le diamètre est la distance maximale entre deux n'importe quel couple d'adopters d'un tweet *h* au sein des *n* premiers tweets.

$$D_n(h) = \max_{1 \le i \ne j \le n-1} d(a_i^h, a_j^h)$$

2. Les communautés au sein du réseau

Nous avons réalisé la détection de communautés à l'aide d'un algorithme Infomap (Rosvall Bergstrom, 2008). Cet algorithme permet de détecter les communautés disjointes au sein d'un graph.

a. Nombre de communautés infectées

Cette feature nous permet de déterminer le nombre de communautés ayant au moins du adopter un hashtag *h* parmi les *n* premiers tweets :

$$|C_n(h)|$$

3. La croissance du réseau

a. Durée moyenne entre chaque tweets

Prenons la série temporelle des *n* premiers tweets d'un hashtag *h*, $\langle t_1^h, t_{2,}^h \ldots t_n^h \rangle$. La durée moyenne entre chaque tweet est exprimée de la manière suivante :

$$\overline{\Delta t_n(h)} = \frac{\sum_{i=1}^{n-1} t_{i+1}^h - t_i^h}{n-1} = \frac{t_n^h - t_1^h}{n-1}$$

b. Le coefficient de variation de la durée entre chaque tweets

$$C_v(\Delta t_n(h)) = \frac{1}{\Delta t_n(h)} \sqrt{\frac{\sum_{i=1}^{n-1} (t_{i+1}^h - t_i^h - \Delta t_n(h))^2}{n-2}}$$

IV. Prédiction de viralité des hashtags :

Nous avons donc implémenté l'extraction de features pour un hashtag donné, puis nous avons généralisé l'extraction des features pour 201 hashtags (l'extraction de features pour 201 hashtags nous a tout de même pris 1h30 ; un vrai trou noir). L'objectif de prédiction que nous nous sommes fixés est de prédire le nombre de communautés infectées par la propagation d'un hashtag. Grâce à l'algorithme de *Random Forest*, nous espérons pouvoir prédire ce nombre de communautés infectées par un hashtag totalement inédit. La structure très complexe de la donnée à notre disposition nécessitait un travail préalable de features engineering. Vous trouverez dans un Notebook cette phase de *machine learning*. Nous nous retrouvons avec de très bonnes performances dans la mesure où notre algorithme de *Random Forest* d'environ 10,65%. C'est très bien, c'est certes perfectible, mais c'est prometteur pour la suite.

Nous envisageons d'approfondir ce travail par la suite pour réaliser cette prédiction sur un nombre d'hashtags bien plus ample, et également réaliser des prédictions sur les autres features. Nous n'avons pas pu le faire par manque de temps, mais c'est un travail intéressant qui permettrait de mesurer l'influence d'un feature sur le destin d'un hashtag.

V. Conclusion :

En définitive, la prédiction de viralité permet de mieux saisir les phénomènes de *« buzz »* sur les réseaux sociaux. Les hashtags Twitter constituent d'excellents cobayes pour un tel travail. En plus de mieux saisir les influences des tendances, et de mieux saisir la psychologie des fervents utilisateurs de Twitter, nous avons pris beaucoup de plaisir à travailler sur cette thématique, en particulier sur un dataset aussi riche (bien que complexe). Des ressources plus importantes en termes de puissance de calcul sont nécessaires, mais le jeu en vaut la chandelle, et la possibilité d'étendre ce genre de prédictions à des domaines comme la santé peut constituer un challenge fort intéressant et bénéfique pour tous.

VI. Bibliographie :

- *Weng, Menczer, Ahn (2014) - Predicting Successful Memes using Network and Community Structure*

- *Rosvall, Bergstrom (2008) - Maps of random walks on complex networks reveal community structure*