Un Modèle de Diffusion de l'Information dans les Réseaux Sociaux

Cédric Lagnier*, Eric Gaussier*

*Université Grenoble 1 / Laboratoire LIG
Campus - Bat. B
385 avenue de la Bibliothèque
38400 Saint Martin d'Hères
{Cedric.Lagnier,Eric.Gaussier}@imag.fr

Résumé. Les réseaux sociaux sont un outil que les gens utilisent de plus en plus pour communiquer et partager de l'information. Un certain nombre d'études ont été effectuées, sur les réseaux sociaux, la propagation de l'innovation et les maladies afin de comprendre et de modéliser la diffusion dans des graphes d'utilisateurs. Dans un premier temps, nous présentons ici un modèle de diffusion de l'information dans les graphes de contenu alliant aussi bien l'influence des voisins que celle de la proximité avec le contenu, avant d'illustrer notre modèle par des exemples de diffusion sur des réseaux générés manuellement et des réseaux réels. Puis, dans une seconde partie, nous introduisons une dynamique de groupe afin de considérer ensemble les utilisateurs similaires au sein du réseau social.

1 Introduction

Les réseaux sociaux permettent à n'importe qui de partager de l'information avec un grand nombre de personnes, cette information pouvant aller du simple texte (documents, annotations, commentaires, ...) aux images, vidéos, ou tout autre contenu. Chaque utilisateur sur un réseau possède un profil qui lui est propre et qui guide ses choix sur le réseau. De plus, il est relié à un certain nombre d'autres par choix personnel. Que ces liens soient appelés "lien d'amitié", "lien de suivi" ou autre, leur signification reste la même : ils définissent une proximité entre les utilisateurs. Un utilisateur partageant de l'information sur ces réseaux le fera donc principalement avec les autres utilisateurs avec lesquels il est relié.

On peut distinguer plusieurs types de réseaux sociaux, dont voici une liste non exhaustive. Les réseaux de blogs sont regroupés autour d'une thématique ciblée. Il existe des liens d'amitié entre les bloggueurs qui diffusent de l'information à la fois par création de billets et par dépôt de commentaires. Un certain nombre de réseaux sont basés sur l'annotation collaborative. On peut citer par exemple *Bibsonomy* ou *Del.icio.us* qui permettent aux utilisateurs de partager des articles et des liens puis de les annoter. Des réseaux tels que *Flickr* ou *Youtube* sont fondés sur le partage de contenu des utilisateurs. Que ce soient les photos pour *Flickr* ou les videos pour *Youtube*, les utilisateurs peuvent ensuite annoter et commenter les contenus ainsi diffusés. Les grands journaux se sont aussi lancés dans ces nouveaux moyens de communication que sont les réseaux sociaux. On voit ainsi *TimesPeople* ou *Le monde* permettre à leurs utilisateurs

de recommander des articles de presse à leurs voisins. Enfin, nous citerons *Twitter* et *Facebook* qui sont un peu différents des précédents dans la mesure où ils sont fondés sur le partage d'information de tout type, *Twitter* permettant aux utilisateurs de diffuser des messages courts (140 caractères maximum) et *Facebook* donnant la possibilité aux utilisateurs d'échanger tout type d'information.

Au sein de ces réseaux, les utilisateurs vont pouvoir adopter plusieurs types de comportements. Tout d'abord, il est possible de poster une information qui n'était pas déjà présente sur le réseau. Dans ce cas de figure, nous définissons l'utilisateur comme étant un producteur. Le contenu peut être autonome ou associé à un autre contenu (ce sera le cas par exemple lors d'annotations). Le second rôle que les utilisateurs peuvent prendre est celui de spectateur/consommateur. Ils sont définis comme tel lorsqu'ils lisent simplement un contenu diffusé sur le réseau. Après avoir consommé un contenu, l'utilisateur peut vouloir faire suivre celui-ci à ses contacts ; il prend alors le rôle de diffuseur qui se distingue de celui de producteur dans le sens ou l'information est déjà présente sur le réseau. Nous distinguons deux types de diffusion : la diffusion explicite qui consiste à re-diffuser un contenu tel quel et la diffusion implicite qui consiste en une diffusion d'un lien vers un contenu interne au réseau. Deux rôles supplémentaires un peu différents peuvent être observés. Le modérateur veille à l'intégrité de l'information présente dans le réseau et le nuisible cherche à nuire au bon fonctionnement du réseau. Cela peut aller du simple "troll" au filtrage intempestif ou même à la falsification d'information.

Comme nous venons de le voir, les réseaux sociaux et leurs utilisateurs ont un certain nombre de caractéristiques que nous pouvons chercher à exploiter pour résoudre différents problèmes. L'identification de communautés d'utilisateurs ou de thèmes sur les contenus peuvent permettre de trier plus facilement les informations et les utilisateurs. Dans ce même sens, on peut vouloir chercher à filtrer l'information d'une part en utilisant la topologie du réseau et d'autre part par le contenu qui circule. Un autre problème est la détection de liens dans les réseaux : avec quels autres utilisateurs un utilisateur donné aura envie de se connecter et pour quelles raisons ? Le problème qui va nous intéresser dans cet article est celui de la diffusion de l'information.

Lorsqu'un contenu est diffusé dans un réseau, il atteint un certain nombre d'utilisateurs en un certain temps. En partant du principe qu'il existe un chemin entre un utilisateur source et un utilisateur cible dans le réseau, est-ce que l'information va atteindre la cible, et si oui au bout de combien de temps ? Au bout de combien de temps le contenu aura été diffusé auprès d'un pourcentage donné du réseau ? A partir de quels nœuds est-il préférable de diffuser une information car elle sera le plus rapidement diffusée ? Pour répondre à ces questions, nous définissons un modèle pour représenter la diffusion de l'information dans un réseau social. Ce modèle a pour but de caractériser la volonté de diffusion d'un contenu par chacun des utilisateurs du réseau. Sa caractéristique principale est de tenir aussi bien compte du graphe des utilisateurs, et donc des voisins, que de la similarité entre le contenu diffusé et les centres d'intérêt de l'utilisateur.

Dans la partie suivante, nous présentons les différents domaines et travaux qui sont reliés à notre problématique. Dans la partie 3, nous présentons une première version de notre modèle puis l'illustrons par des exemples dans la partie 4. La partie 5 présente une variante de notre modèle en définissant une dynamique de groupe. Nous concluons et présentons les perspectives de ce travail dans la partie 6.

2 Travaux reliés

L'étude des réseaux sociaux commence par l'étude de leurs caractéristiques (Mercklé (2004)). Le nombre d'utilisateurs, la densité moyenne du réseau, la centralité d'un utilisateur, etc permettent une première approche pour comprendre les interactions au sein du réseau (Mazzoni (2006)). Chaque utilisateur a des comportements que l'on peut essayer de catégoriser afin de comprendre et prédire les actions de chacun (Golder et Donath (2004)). En effet, on s'aperçoit que les utilisateurs prennent des décisions en suivant un ensemble de rôles qui diffèrent selon le contexte, leur place au sein de la communauté ou encore leur utilisation du réseau. De plus, la topologie du réseau joue grandement sur la diffusion, et on verra souvent des regroupements caractéristiques d'invidus être favorables ou défavorables à la propagation d'information (Lieberman et al. (2005)).

Plusieurs études ont été réalisées sur la diffusion dans les grands réseaux sociaux (Cha et al. (2009)). On y découvre notamment que la diffusion ne réussit pas pour tous les contenus. La réussite ou non de la diffusion d'un contenu est très dépendante des diffuseurs initiaux. C'est d'ailleurs l'une des principales questions : comment maximiser l'influence au sein d'un réseau social, ou comment choisir les diffuseurs initiaux de manière à ce que la diffusion soit optimale. Il a été montré sur un certain nombre de modèles que le calcul exact de cet ensemble de meilleurs diffuseurs initiaux est NP-Complet (Kempe et al. (2003)), et il importe dans ce cas de bien approcher la solution optimale (Kimura et al. (2007)).

La diffusion dans les réseaux est principalement étudiée au sein de trois communautés : la diffusion de maladie dans un groupe d'individus, la propagation d'innovations dans une population et la diffusion d'information dans un réseau social. Tous utilisent des modèles similaires en les adaptant aux problèmes posés. Les principales caractéristiques qui ressortent dans ces modèles sont la contagion, qui peut être à la fois interne (provenant d'individus du réseau) ou externe, et le taux d'adoption, qui est considéré comme une variable d'étude privilégiée (selon la finalité, on va chercher à maximiser ce taux dans le cas de la diffusion d'un produit, ou le minimiser dans le cas de la propagation d'une maladie). Les principaux modèles sont de deux formes différentes. Tout d'abord les modèles de seuil introduit par Granovetter (1978) dans lesquels un utilisateur s'active si un nombre prédéfini de ses voisins (le paramètre de seuil) sont déjà actifs. Le second type de modèle est le modèle de cascade indépendante où un utilisateur qui devient actif a une probabilité non nulle d'activer ses voisins. Dans ces modèles, on retrouve deux états : non actif et actif. On peut trouver une utilisation de ce type de modèles dans Saito et al. (2010). On retrouve ici une correspondance avec un des premiers modèles en épidemiologie, SI (Susceptible-Infected), dans lequel une personne est tout d'abord saine mais avec la possibilité d'être infectée, puis devient infectée quand ses proches le sont. Dans ce modèle, on définit des probabilités de transition entre les états. Les modèles ont ensuite evolué pour voir apparaitre d'autres états comme immunisé ou exposé (Trottier et Philippe (2001)).

On trouve une evolution de ces modèles standards dans Abrahamson et Rosenkopf (1997) qui présente un modéle de propagation des innovations fondé sur un modèle de seuil, en étudiant l'impact des autres utilisateurs sur l'adoption d'une innovation. Liben-Nowell et Kleinberg (2008) montrent un modèle de diffusion de chaînes de mail dérivé des cascade indépendantes en ajoutant plusieurs paramètres : un paramètre de défausse qui autorise l'utilisateur à ne pas s'intéresser à l'information, il peut s'agir par exemple d'un anti-spam et un paramètre de latence qui va permet à un utilisateur qui accepte un contenu de ne pas le faire à l'étape présente mais avec une certaine latence.

Pour chaque modèle, les auteurs ont une représentation de ce qu'est la diffusion. Certains définissent une diffusion comme une cascade : un sous-graphe du graphe des utilisateurs composé seulement des utilisateurs étants devenus actifs lors d'une diffusion. Dans Leskovec et al. (2007a), la diffusion de recommendations de produits au sein d'une population sur un site de commerce en ligne est étudiée. Les auteurs montrent notamment que l'on ne voit que pour très peu de produits une réelle diffusion. Une autre étude, présentée dans Leskovec et al. (2007b), cible cette fois-ci la diffusion d'information dans les blogs. La finalité de ce travail est un premier modèle de génération de cascades qui produit des cascades dont les caractéristiques sont similaires aux cascades observées sur les réseaux réels. Enfin, un modèle probabiliste de génération de cascades dans un graphe d'utilisateurs est défini dans Gomez-Rodriguez et al. (2010).

D'autres modèles, étudiés pour la propagation des innovations, se rapprochent plus de notre travail en définissant des modèles probabilistes fondès sur des équations aux différences : Young (2009) et Lopez-Pintado (2008). Le but est de caractériser la dynamique du taux d'adoption des innovations afin de pouvoir étudier l'évolution de la diffusion au cours du temps. On recherche ici les états stationnaires lorsqu'il y a convergence, mais ces modèles permettent aussi d'étudier le fonctionnement moyen d'un système (étape par étape), et pas seulement son fonctionnement à la limite (au bout d'un temps très long).

3 Modèle général

3.1 Présentation du modèle

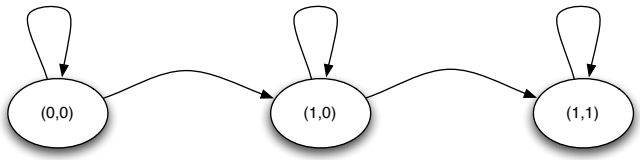


FIG. 1 – *3 états*

Soit $G = (V, E)$ un graphe orienté où V est un ensemble d'utilisateurs et E est l'ensemble des relations entre ces utilisateurs. Nous ne traiterons ici qu'un seul type de relation que nous appellerons relation de proximité. Nous étudions la diffusion d'un seul document, mais le modèle peut être adapté pour plusieurs documents. Un utilisateur a deux actions possible lorsqu'au moins un de ses voisins lui a diffusé le document : l'accepter (le lire, ne pas le jeter) et le diffuser auprès de ses propres voisins. Pour pouvoir diffuser le document, un utilisateur doit obligatoirement l'avoir accepté auparavant. Notre modèle est donc composé de trois états qui sont représentés sur la figure 1 avec les transitions possibles entre eux et qui correspondent aux actions précédentes. L'état (0,0) correspond à un utilisateur qui n'a pas encore accepté le document (et donc pas encore diffusé), l'état (1,0) correspond à un utilisateur qui a accepté le document mais qui ne l'a pas encore diffusé. Enfin, l'état (1,1) correspond à un utilisateur qui a accepté et diffusé le document. A tout moment, un utilisateur se trouve dans chacun des trois états avec une certaine probabilité. Soit $\rho^{0,0}(i,t)$ la probabilité que l'utilisateur i se trouve dans

l'état (0,0) au temps t. Il en est de même pour les états (1,0) et (1,1). On a alors la contrainte suivante : $\rho^{0,0}(i,t) + \rho^{1,0}(i,t) + \rho^{1,1}(i,t) = 1$.

Il y a deux probabilités à définir qui sont la probabilité d'acceptation $F_i(t)$ (pour passer de l'état (0,0) à l'état (1,0)) et la probabilité de diffusion $G_i(t)$ (pour passer de l'état (1,0) à l'état (1,1)). La probabilité de rester dans l'état (0,0) se définit alors par $1 - F_i(t)$. Il en va de même pour celle de rester dans l'état (1,0). Chacune de ces probabilités dépend des paramètres de l'utilisateur i :

$$F_i(t) = F(\nu_{id}, s_{id}, a_i(t)) \text{ (probabilité d'acceptation)}$$

où :

- ν_{id} est le taux de propagation externe du document d pour un nœud/utilisateur i. Il représente toutes les informations qui sont diffusées hors du réseau et qui influencent quand même l'utilisateur. Si l'on se place dans un réseau social de type *Twitter*, la télévision fera partie des moyens de communication externes au réseau qui contribueront quand même à la diffusion de l'information au sein de celui-ci ;
- $s_{id} = sim(\phi_i, d)$ est la similarité entre le profil (centres d'intérêt) de l'utilisateur i et le document d. Ce paramètre pénalise la probabilité d'acceptation si le document ne fait pas partie des centres d'intérêt de l'utilisateur et l'augmente dans le cas inverse ;
- $a_i(t)$ correspond au nombre de voisins (entrants) qui ont déjà diffusé le document d au temps t.

L'intuition qui sous-tend notre modèle est que plus un utilisateur a de contacts qui lui diffusent une information, plus la probabilité qu'il l'accepte est élevée. On voit bien ici les deux clés de la diffusion dans ce modèle : si le contenu est très proche du profil d'un utilisateur, celui-ci va être prompt à le rediffuser. Si par contre le contenu ne correspond que peu, voire pas du tout, aux centres d'intérêt de l'utilisateur, il faudra qu'un grand nombre de ses contacts lui aient diffusé ce même contenu pour qu'il l'accepte. De même, nous définissons :

$$G_i(t) = G(F_i(t), w_i) \text{ (Probabilité de diffusion)}$$

où w_i est un paramètre de volonté (*willingness*) de i. En effet, certains utilisateurs ont une plus grande tendance à diffuser de l'information que d'autres. Par l'intermédiaire de la probabilité d'acceptation, la probabilité de diffusion dépend elle aussi du nombre de voisins diffuseurs et de la proximité de l'information avec le profil de l'utilisateur. Pour diffuser un contenu un utilisateur doit l'avoir accepté auparavant, et il diffusera ce contenu avec une probabilité d'autant plus grande si sa probabilité d'acceptation était élevée.

Notre choix sur la forme de la fonction de transition s'est porté sur les fonctions de seuil. Beaucoup de modèles existants (Lopez-Pintado (2008) et Kempe et al. (2003) par exemple) les utilisent car elles permettent d'activer l'utilisateur si les paramètres atteignent un certain niveau dit "de seuil". Notre modèle étant probabiliste, nous avons choisi une fonction de seuil sous forme de sigmoïde :

$$F_i(t) = \begin{cases} \frac{1}{1+exp(-\lambda_1(s_i-0.5)-\lambda_2 a_i(t)-\lambda_3\nu_i)} & \text{si } a_i(t) \geq 1 \\ 0 & \text{sinon} \end{cases}$$

$$G_i(t) = \frac{1}{1+exp(-\beta_1(F_i(t)-0.5)-\beta_2 w_i)}$$

Certains paramètres (la similarité pour F et F pour G) pénalisent la probabilité lorsqu'ils sont en dessous d'une certaine valeur, alors que les autres ne font que contribuer à l'augmentation de la probabilité.

On peut maintenant exprimer l'évolution des utilisateurs au cours du temps en fonction des probabilités de transition :

$$\rho^{(1,1)}(i,t+1) = \rho^{(1,1)}(i,t) + \rho^{(1,0)}(i,t) \times G_i(t)$$
$$\rho^{(1,0)}(i,t+1) = \rho^{(1,0)}(i,t) \times (1 - G_i(t)) + \rho^{(0,0)}(i,t) \times F_i(t)$$
$$\rho^{(0,0)}(i,t+1) = \rho^{(0,0)}(i,t) \times (1 - F_i(t))$$

Pour chacun des 3 états, la probabilité que l'utilisateur s'y trouve à l'étape $t+1$ est égale à la probabilité qu'il n'en soit pas parti à laquelle s'ajoute la probabilité qu'il y soit entré.

On s'aperçoit avec ces équations que la convergence va se faire lorsque tous les utilisateurs reliés à l'initiateur de la diffusion seront dans l'état (1,1), c'est-à-dire qu'ils auront tous accepté puis diffusé le document. C'est une caractéristique du modèle due au fait qu'il n'y a aucune transition possible pour revenir dans un état antérieur, ni aucune diminution des probabilités d'acceptation et de diffusion au cours du temps.

Le nombre de paramètres dans le modèle que nous venons de présenter est, *a priori*, de l'ordre de $5 + 2|V|$. Il dépend cependant de la façon dont les taux de propagation et les paramètres de volonté sont traités, et peut être largement diminué si l'on se repose sur des hypothèses simplificatrices attribuant une valeur fixe, indépendante des utilisateurs, à ces paramètres. De plus lorsque l'on dispose de données réelles, il est possible d'estimer les valeurs des différents paramètres suivant la méthode du maximum de vraisemblance, à partir des séries de diffusion introduites dans Saito et al. (2009).

3.2 Implantation et Compléxité

L'algorithme 1 décrit le processus de diffusion sur un réseau social. L'initialisation définit les états des utilisateurs au départ (avant de commencer la diffusion) : diffuseur initial ou utilisateur standard. La phase de diffusion correspond ensuite à la mise à jour de l'état de tous les utilisateurs à chaque étape. Un des paramètres important à prendre en compte lors de la réalisation d'un algorithme de diffusion dans un réseau social est le passage à l'échelle. En effet, un grand nombre de réseaux comportent des centaines de milliers d'utilisateurs, voire des millions. Dans cette optique, nous maintenons une liste des voisins entrants de chaque utilisateur ($N^{in}(i)$, l'ensemble des voisins entrants de *i* pour y accéder dans un temps constant. La complexité de notre algorithme s'exprime sous la forme :

$$\text{Complexité} = O(\text{nb\_étapes} * |V| * \tilde{k}_{in})$$

et dépend donc du nombre d'étapes avant convergence, du nombre total d'utilisateurs $|V|$, et du nombre moyen de liens entrants par utilisateur $\tilde{k}_{in}$.

4 Illustrations

Le but de cette section est d'illustrer l'évolution de la densité d'utilisateurs dans chacun des trois états de façon globale sur le réseau, et ce afin d'observer la diffusion d'une information partant d'un unique utilisateur. Nous définissons la densité d'utilisateurs dans un état comme étant la moyenne des probabilités de présence dans l'état sur tous les utilisateurs :

Algorithm 1 Algorithme de diffusion de l'information dans un réseau social

{initialisation}
for all user i **do**
  if i est diffuseur initial **then**
    $\rho^{(0,0)}(i,0) = \rho^{(1,0)}(i,0) = 0$
    $\rho^{(1,1)}(i,0) = 1$
  else
    $\rho^{(0,0)}(i,0) = 1$
    $\rho^{(1,0)}(i,0) = \rho^{(1,1)}(i,0) = 0$
  end if
end for
{diffusion}
for $t = 1$ jusqu'à convergence **do**
  for all user i **do**
    $a_i(t) = \sum_{n \in N^{in}(i)} \rho^{(1,1)}(n, t-1)$
    $F_i(t) = F(s_i, a_i(t))$
    $G_i(t) = G(F_i(t))$
    $\rho^{(0,0)}(i,t) = -\rho^{(0,0)}(i,t-1) \times F_i(t-1)$
    $\rho^{(1,0)}(i,t) = \rho^{(0,0)}(i,t-1) \times F_i(t-1) - \rho^{(1,0)}(i,t-1) \times G_i(t-1)$
    $\rho^{(1,1)}(i,t) = \rho^{(1,0)}(i,t-1) \times G_i(t-1)$
  end for
end for

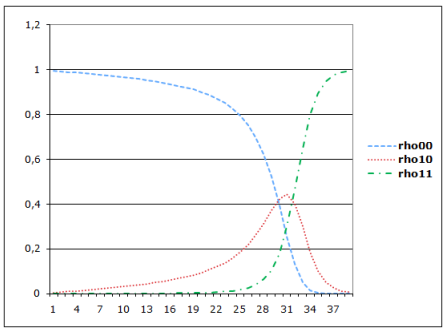
$$\rho^{(0,0)}(t) = \frac{1}{|V|} \sum_{i \in V} \rho^{(0,0)}(i,t)$$
$$\rho^{(1,0)}(t) = \frac{1}{|V|} \sum_{i \in V} \rho^{(1,0)}(i,t)$$
$$\rho^{(1,1)}(t) = \frac{1}{|V|} \sum_{i \in V} \rho^{(1,1)}(i,t)$$

4.1 Sur des réseaux virtuels

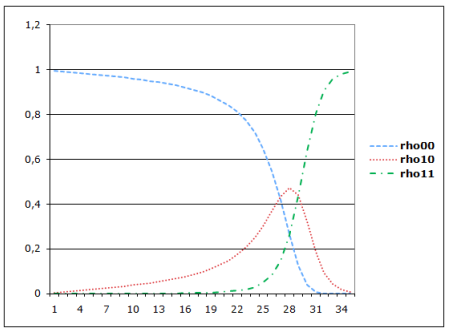
Nous avons généré un certain nombre de graphes suivant une loi de puissance en utilisant le programme *pywebgraph*[1] de façon à simuler l'exécution du modèle sur des réseaux générés artificiellement. Dans ces réseaux virtuels, il n'existe pas de profil des utilisateurs et nous avons utilisé une loi uniforme sur [0,1] pour générer les similarités entre les utilisateurs et le contenu diffusé.

Tout d'abord, nous avons regardé l'influence des paramètres du modèle et de la densité de liens dans le graphe des utilisateurs. La figure 2 montre cinq graphiques représentants l'évolution des densités d'utilisateurs au cours du temps. Nous avons fixé le paramètre d'influence extérieure à 0 ($\lambda_3 = 0$). De plus, nous sommes partis du principe que tous les utilisateurs avaient la même propension à diffuser et ne tenons du coup pas compte du paramètre de vo-

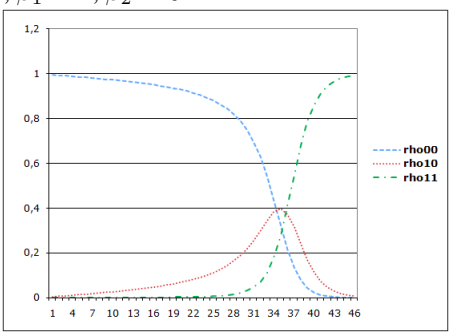
[1] http ://pywebgraph.sourceforge.net/



$\lambda_1 = 0.5, \lambda_2 = 0.5, \beta_1 = 1, \beta_2 = 0$

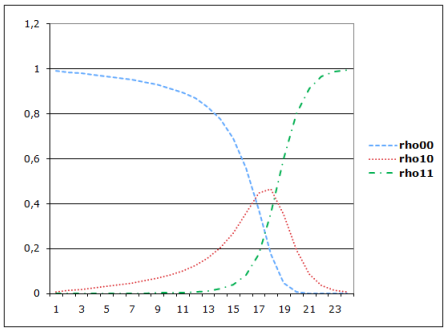


$\lambda_1 = 0, \lambda_2 = 1, \beta_1 = 1, \beta_2 = 0$

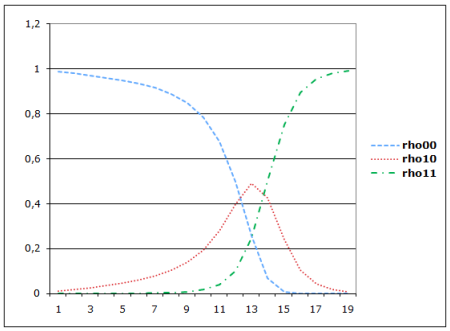


$\lambda_1 = 1, \lambda_2 = 0, \beta_1 = 1, \beta_2 = 0$

(a) Réseau de 5000 nœuds et 5000 liens orientés ($\lambda_3 = 0$)



10000 liens



15000 liens

(b) Réseau de 5000 nœuds ($\lambda_1 = 0.5, \lambda_2 = 0.5, \lambda_3 = 0, \beta_1 = 1, \beta_2 = 0$)

FIG. 2 – *Influence des paramètres et du nombre de liens*

lonté de diffuser ($\beta_2 = 0$). L'utilisateur qui est diffuseur initial est choisi aléatoirement. On voit que quelles que soient les valeurs des paramètres, la forme globale reste la même. Tout d'abord les utilisateurs commencent à accepter l'information puis à la diffuser. La vitesse dépend du nombre d'utilisateurs ayant déjà diffusé l'information. En effet, plus le nombre d'utilisateurs diffuseurs augmente, plus la vitesse de diffusion augmente. Ceci est dû d'une part au fait que l'information a plus de points d'entrées quand le nombre de diffuseurs augmente, et d'autre

part au fait que plus un utilisateur a de voisins diffuseurs, plus il a de chance d'accepter puis de diffuser à son tour l'information. Comme nous l'avons déjà souligné, au bout d'un certain temps, tous les utilisateurs reliés au diffuseur initial (c'est-à-dire l'ensemble des utilisateurs ici) auront diffusé (état (1,1)). Il est par contre intéressant de noter que selon les valeurs des paramètres et la densité du réseau, le nombre d'étapes pour que le système converge varie.

5000 liens			10000 liens	15000 liens
$\lambda_1 = \lambda_2 = 0.5$	$\lambda_1 = 0, \lambda_2 = 1$	$\lambda_1 = 1, \lambda_2 = 0$	$\lambda_1 = \lambda_2 = 0.5$	$\lambda_1 = \lambda_2 = 0.5$
37	34	45	23	19

TAB. 1 – *Influence des paramètres sur un réseau de 5000 nœuds avec des liens orientés : nombre d'étapes avant convergence ($\lambda_3 = 0, \beta_1 = 1, \beta_2 = 0$).*

Le tableau 1 résume le nombre d'étapes nécessaires pour que le système converge pour les cas observés sur la figure 2. Il montre deux aspects importants : tout d'abord, si l'on ne tient compte que du paramètre de similarité et pas du nombre de voisins diffuseurs, la diffusion est plus longue que lorsque l'on ne tient compte que du nombre de voisins diffuseurs. Ceci est dû au fait que le paramètre de similarité avec le contenu peut pénaliser la probabilité de transition si elle est inférieure à 0.5, alors que le nombre de voisins ne fait que l'améliorer. Ensuite, on remarque que plus le réseau est dense, c'est-à-dire plus le nombre de liens par utilisateur est grand, plus la diffusion est rapide. Ceci est dû au fait que plus le réseau est dense, plus le nombre moyen de voisins est élevé et il est donc normal de voir l'information diffusée plus rapidement. A noter également que plus le réseau est dense, plus la centralité de chaque utilisateur augmente, ce qui favorise également la diffusion au sein du réseau.

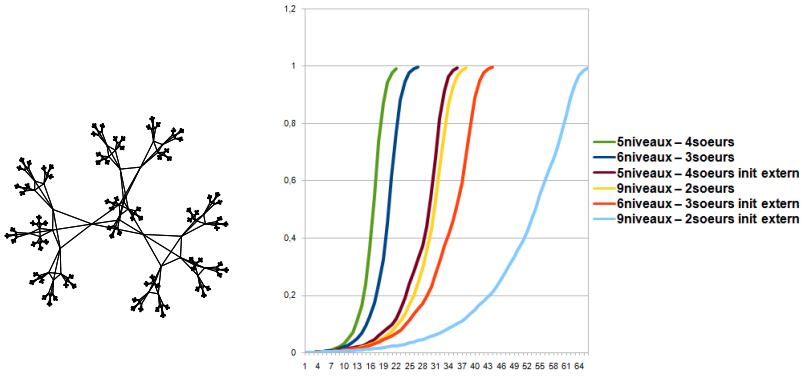
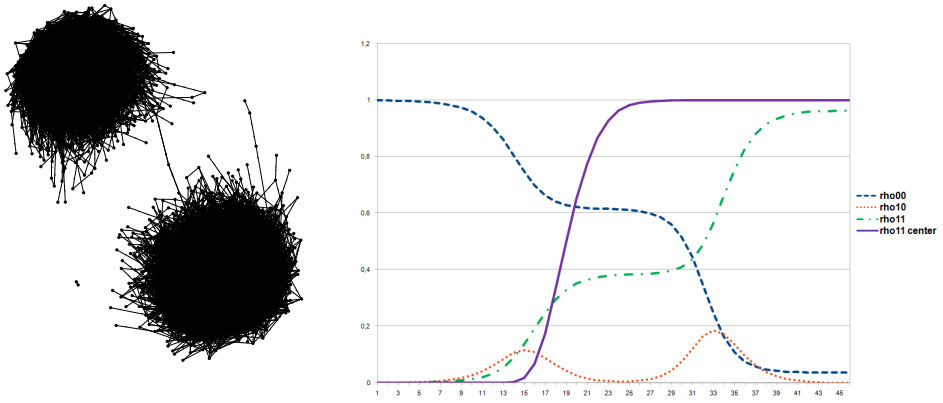


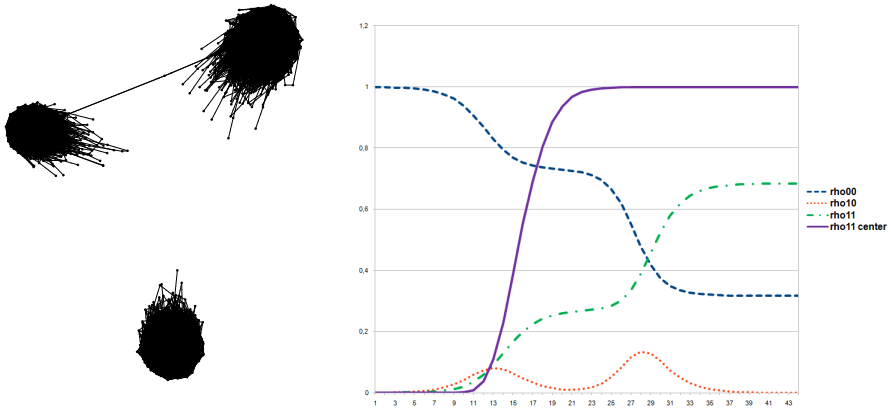
FIG. 3 – *Graphe en étoile (1000 utilisateurs et 4000 liens) ; initialisation au nœud central ou externe : 6 niveaux et 3 sœurs, 9 niveaux et 2 sœurs, 5 niveaux et 4 sœurs ; la légende présente les courbes de gauche à droite. ($\lambda_1 = 0.5$, $\lambda_2 = 0.5$, $\beta_1 = 1$)*

La figure 3 représente des graphes en étoile sur lesquels nous avons travaillé. La représentation a été faite avec le programme *gephi* (Bastian et al. (2009)). Ces graphes sont définis par

deux paramètres : le niveau de profondeur et le nombre de fils de chacun des nœuds. On part du nœud central, on lui crée ses fils, ensuite on relie ces fils entre eux et on ré-itère le processus pour chacun des fils jusqu'à atteindre la profondeur souhaitée. Nous avons créé trois graphes avec des paramètres différents ayant tous environ 1000 nœuds et 4000 liens, puis nous avons diffusé une information depuis le centre ou depuis la périphérie des graphes et étudié l'évolution. On remarque que plus la profondeur est faible et plus le nombre de fils est grand, plus l'information se diffuse vite. Ceci est toujours dû au fait que les utilisateurs sont plus centraux dans ces graphes. On remarque aussi que le départ de l'information de la périphérie du graphe double le temps que l'information met à se diffuser dans tout le graphe. Malgré le fait que les frères sont liés entre eux, l'information a besoin de remonter vers le centre pour se diffuser dans les autres branches.



(a) 2 communautés reliés (graphe des noeuds et évolution de la diffusion) : 2500 (1000+1500) nœuds et 2000 liens ; présence d'un nœud central reliant les deux communautés



(b) 3 communautés dont seulement 2 reliées (graphe des noeuds et évolution de la diffusion)

FIG. 4 – *Graphes de communautés* ($\lambda_1 = 0.5$, $\lambda_2 = 0.5$, $\beta_1 = 1$)

La figure 4 représente des réseaux formés de plusieurs communautés reliées par un unique nœud et par un unique lien. Dans le graphe 4(a), seulement deux communautés sont reliées entre elles. Ce qu'il y a de flagrant sur ce graphique, c'est l'évolution en deux étapes. Tout d'abord l'information se propage dans la première communauté, puis atteint le nœud central. C'est alors qu'elle peut être diffusée dans la seconde communauté mais avec de nouveau un seul point entrant, ce qui explique les deux étapes bien distinctes. En somme, diffuser une information au sein de deux communautés faiblement reliées revient a diffuser deux fois l'information, comme si les deux communautés étaient des graphes distincts. Si l'on diffuse l'information depuis le nœud central, les deux étapes disparaissent puisque la diffusion va se faire en parallèle au sein des deux communautés. La figure 4(b) montre le même phénomène sur un réseau formé de trois communautés dont seulement deux sont reliées entre elles. La diffusion se déroule comme dans le cas précédent, mais on remarque que la densité de diffuseurs converge à une valeur de 0.7 après seulement deux étapes de diffusion distinctes car la troisième communauté ne reçoit jamais l'information.

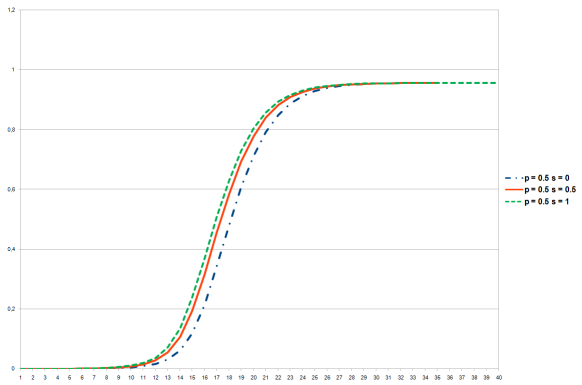


FIG. 5 – *Influence de la similarité* ($\lambda_1 = 0.5$, $\lambda_2 = 0.5$, $\beta_1 = 1$)

Enfin, nous avons modifié la similarité du nœud central dans le réseau à deux communautés puis nous avons diffusé une information depuis un nouveau nœud seulement relié au nœud central de façon à voir l'influence de la similarité sur la diffusion. Le paramètre λ_1 est fixé à 2 pour augmenter l'influence de la similarité par rapport à celle des voisins. La figure 5 illustre les résultats obtenus : une similarité plus faible entraîne une diffusion plus lente sans pour autant jouer un rôle prépondérant sur l'évolution du système. Nous comptons dans la suite estimer les paramètres sur un réseau réel de façon à mieux déterminer l'influence de la similarité dans notre modèle.

4.2 Sur des réseaux réels

Dans cette seconde partie nous avons utilisé les graphes d'utilisateurs de deux jeux de données réels :

– le graphe généré à partir des échanges de mails du jeu de données Enron[2] composé d'environ 37000 nœuds et 370000 liens ;
– le graphe du jeu de données utilisé pour la conférence ICWSM[3] qui est un ensemble de billets de blogs. Il comporte environ 85000 nœuds et 130000 liens.

Nous n'avons pour cette étude utilisé que le graphe des utilisateurs de chacun de ces réseaux de manière à illustrer notre modèle sur des rèseaux réels. Nous avons utilisé des similarités générées aléatoirement. Nous utiliserons les diffusions d'information présentes dans ces jeux de données pour estimer les paramètres du modèle dans une prochaine étude.

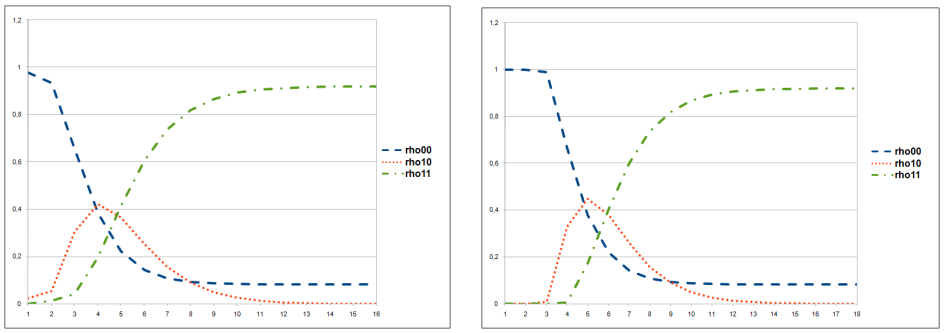


FIG. 6 – *Diffusion sur le graphe d'utilisateurs du jeu de données Enron* ($\lambda_1 = 0.5$, $\lambda_2 = 0.5$, $\beta_1 = 1$)

La figure 6 représente la diffusion à partir de deux utilisateurs différents sur le jeu de données Enron. La première chose que l'on peut remarquer est le fait que la diffusion ne s'effectue que sur une partie des utilisateurs, ceci étant dû au fait que tous les utilisateurs ne peuvent pas être atteints à partir du diffuseur initial. En effet, la convergence se fait lorsque 91% des utilisateurs ont diffusé l'information. On voit de plus que la vitesse de la diffusion varie en fonction du nœud qui en est à l'origine. Dans le premier graphe, la diffusion a été lancée par l'utilisateur le plus connecté du réseau, alors que dans le second le premier diffuseur est un nœud standard.

La figure 7 montre la diffusion dans le réseau de blogs utilisé pour la conférence ICWSM. Par rapport aux deux diffusions sur le graphe d'Enron, la diffusion touche ici un nombre beaucoup plus restreint d'utilisateurs, ceci étant dû au fait que le réseau est bien moins connecté. Cet exemple permet de mettre en valeur l'importance de bien choisir la source de la diffusion d'une information. D'une part celle-ci détermine la vitesse de diffusion mais aussi l'ensemble des utilisateurs que l'on peut atteindre. Dans un réseau peu connecté comme celui-ci, il faut avoir un grand nombre de diffuseurs initiaux pour avoir la possibilité d'atteindre la totalité des utilisateurs. Une des pistes encore ouverte est celle de la sélection des meilleurs diffuseurs initiaux avec le modèle que nous venons de présenter.

[2] http ://www.cs.cmu.edu/ enron/
[3] http ://www.icwsm.org/2010/data.shtml

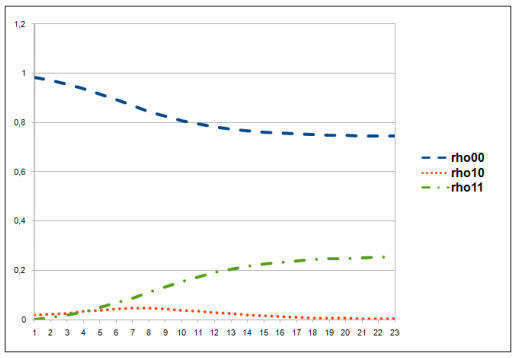


FIG. 7 – *Diffusion sur le graphe d'utilisateurs du jeu de données ICWSM09 ($\lambda_1 = 0.5$, $\lambda_2 = 0.5$, $\beta_1 = 1$)*

5 Dynamique de groupe

Dans cette partie nous introduisons un nouveau concept : les utilisateurs qui agissent de la même façon doivent pouvoir être regroupés afin que leurs actions soient considérées comme identiques. On peut imaginer regrouper tous les utilisateurs d'une même communauté, les utilisateurs ayant les mêmes caractéristiques comme le nombre de liens entrants ou sortants, les utilisateurs ayant des profils similaires, et ainsi de suite. Nous définissons ici une dynamique de groupe pour notre modèle présenté précédemment pour tenir compte de ce phénomène. Nous ne fixons aucune règle sur la manière de regrouper les utlisateurs, mais décrivons seulement l'evolution du système en fonction de ces groupes. On va distinguer deux cas extrêmes : lorsque tous les utilisateurs sont dans le même groupe, on aura alors les décisions de tous les utilisateurs du réseau liées entre elles quelles que soient les caractéristiques de chacun, et lorsque chaque utilisateur se retrouve seul dans un groupe.

Soit $\mathcal{G}$ un groupe d'utilisateurs, la probabilité d'acceptation du groupe est la moyenne pour chacun des membres du groupe de l'espérance de sa probabilité de transition. Les équations de la dynamique de groupe devienent alors :

$$\begin{aligned}
P^{(1,0)}_{(0,0)}(\mathcal{G},t) &= \tfrac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} E[F_i(t)]_{P(a_i(t))} \\
&= \tfrac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\sum_{a=1}^{k_i^{in}} F_i(t)\binom{k_i^{in}}{a}\theta_{\mathcal{G}}(t)^a(1-\theta_{\mathcal{G}}(t))^{k_i^{in}-a} \\
P^{(1,1)}_{(1,0)}(\mathcal{G},t) &= \tfrac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} E[G_i(t)]_{P(a_i(t))} \\
&= \tfrac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\sum_{a=1}^{k_i^{in}} G_i(t)\binom{k_i^{in}}{a}\theta_{\mathcal{G}}(t)^a(1-\theta_{\mathcal{G}}(t))^{k_i^{in}-a}
\end{aligned}$$

où $P^{(1,0)}_{(0,0)}(\mathcal{G},t)$ est la probabilité pour un membre du groupe $\mathcal{G}$ de passer de l'état (0,0) à l'état (1,0) et $P^{(1,1)}_{(1,0)}(\mathcal{G},t)$ est la probabilité pour un membre du groupe $\mathcal{G}$ de passer de l'état (1,0) à

l'état (1,1). $\theta_{\mathcal{G}}(t)$ est la probabilité pour qu'un lien entrant dans le groupe $\mathcal{G}$ provienne d'un diffuseur et k_i^{in} est le nombre de liens entrant au nœud i. $\theta_{\mathcal{G}}(t)$ est défini de la façon suivante :

$$\theta_{\mathcal{G}}(t) = \frac{\sum_{i \in \mathcal{G}} \sum_{n \in N^{in}(i)} \rho_{G(n)}^{(1,1)}}{\sum_{i \in \mathcal{G}} k_i^{in}}$$

où $N^{in}(i)$ est l'ensemble des voisins entrants de i, $G(n)$ est le groupe de l'utilisateur n et k_i^{in} est le nombre de voisins entrants de l'utilisateur i.

Comme pour la dynamique par utilisateur présentée précédemment, nous appelons $\rho^{(0,0)}(\mathcal{G}, t)$ la densité d'utilisateurs du groupe $\mathcal{G}$ présents dans l'état (0,0) à un instant t. Il en est de même pour $\rho^{(1,0)}(\mathcal{G}, t)$ et $\rho^{(1,1)}(\mathcal{G}, t)$. Nous pouvons définir l'évolution du système par l'évolution de ces densités :

$$\begin{aligned}
\rho^{(1,1)}(\mathcal{G}, t+1) &= \rho^{(1,1)}(\mathcal{G}, t) + \rho^{(1,0)}(\mathcal{G}, t) \times P_{(1,0)}^{(1,1)}(t) \\
\rho^{(1,0)}(\mathcal{G}, t+1) &= \rho^{(1,0)}(\mathcal{G}, t) \times (1 - P_{(1,0)}^{(1,1)}(t)) + \rho^{(0,0)}(\mathcal{G}, t) \times P_{(0,0)}^{(1,0)}(t) \\
\rho^{(0,0)}(\mathcal{G}, t+1) &= \rho^{(0,0)}(\mathcal{G}, t) \times (1 - P_{(0,0)}^{(1,0)}(t))
\end{aligned}$$

Comme pour la dynamique par utilisateur, la convergence sera atteinte lorsque tous les utilisateurs seront dans l'état (1,1). Les taux de transition et l'évolution globale de la diffusion des contenus sont alors définis par les équations suivantes :

$$\begin{aligned}
P_{(1,0)}^{(1,1)}(t) &= \tfrac{1}{N_{\mathcal{G}}} \sum_{\mathcal{G}} P_{(1,0)}^{(1,1)}(\mathcal{G}, t) \\
&= \tfrac{1}{N_{\mathcal{G}}} \sum_{\mathcal{G}} \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \sum_{a=1}^{k_i^{in}} G_i(t) \binom{k_i^{in}}{a} \theta_{\mathcal{G}}(t)^a (1 - \theta_{\mathcal{G}}(t))^{k_i^{in} - a} \\
\rho^{(1,1)}(t) &= \tfrac{1}{N_{\mathcal{G}}} \sum_{\mathcal{G}} \rho^{(1,1)}(\mathcal{G}, t)
\end{aligned}$$

Il en est de même pour les autres taux de transition et densités d'utilisateurs. Ces équations permettent d'observer l'évolution de la diffusion de manière globale.

La dynamique de groupe définie ici permet de représenter la diffusion de l'information au sein d'un réseau d'utilisateurs, et de considérer plusieurs utilisateurs comme ayant des réactions proches, des prises de décisions similaires. Lorsque l'on manque de données pour pouvoir caractériser correctement les utilisateurs, il est pratique de pouvoir les regrouper, ceci afin d'éviter de prendre une décision se fondant sur trop peu de données.

6 Conclusion et perspectives

Nous avons présenté un modèle de diffusion de l'information dans les graphes de contenu qui permet de voir évoluer l'état (acceptation et diffusion) des utilisateurs d'un réseau au cours du temps. Dans un premier temps, nous avons défini la dynamique de ce modèle au niveau de chaque utilisateur. Nous avons vu que selon les valeurs des paramètres, la diffusion est plus ou moins dirigée par certaines caractéristiques comme la similarité entre un utilisateur et le

contenu d'une information ou l'influence des voisins dans l'acceptation puis la diffusion d'un contenu. Nous avons aussi vu que la topologie du réseau ainsi que le diffuseur initial jouent beaucoup dans le processus. En effet, un utilisateur très excentré va beaucoup moins bien diffuser une information qu'un utilisateur central. Dans un second temps, nous avons défini une dynamique de groupe pour ce modèle afin de pouvoir considérer ensemble les utilisateurs ayant des volontés et actions similaires. La prochaine étape va consister à tester ce modèle sur des diffusions réelles de façon à estimer la valeur des paramètres automatiquement. Une fois ces paramètres appris, il sera possible de prédire les comportements des utilisateurs (ou groupes d'utilisateurs) ainsi que l'évolution du réseau lors de la diffusion d'une information, et répondre aux questions soulevées.

Références

Abrahamson, E. et L. Rosenkopf (1997). Social network effects on the extent of innovation diffusion : A computer simulation. pp. 289–309.

Bastian, M., S. Heymann, et M. Jacomy (2009). Gephi : An open source software for exploring and manipulating networks.

Cha, M., A. Mislove, et K. P. Gummadi (2009). A measurement-driven analysis of information propagation in the flickr social network. In *WWW '09 : Proceedings of the 18th international conference on World wide web*, New York, NY, USA, pp. 721–730. ACM.

Golder, S. A. et J. Donath (2004). Social roles in electronic communities. In *in Association of Internet Researchers (AoIR) conference Internet Research 5.0*.

Gomez-Rodriguez, M., J. Leskovec, et A. Krause (2010). Inferring networks of diffusion and influence. *CoRR abs/1006.0234*.

Granovetter, M. (1978). Threshold models of collective behavior. *The American Journal of Sociology 83*(6), 1420–1443.

Kempe, D., J. Kleinberg, et E. Tardos (2003). Maximizing the spread of influence through a social network. In *In KDD*, pp. 137–146. ACM Press.

Kimura, M., K. Saito, et R. Nakano (2007). Extracting influential nodes for information diffusion on a social network. In *AAAI'07 : Proceedings of the 22nd national conference on Artificial intelligence*, pp. 1371–1376. AAAI Press.

Leskovec, J., L. A. Adamic, et B. A. Huberman (2007a). The dynamics of viral marketing.

Leskovec, J., M. McGlohon, C. Faloutsos, N. Glance, et M. Hurst (2007b). Cascading behavior in large blog graphs.

Liben-Nowell, D. et J. Kleinberg (2008). Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences 105*(12).

Lieberman, E., C. Hauert, et M. A. Nowak (2005). Evolutionary dynamics on graphs. *Nature 433*(7023), 312–316.

Lopez-Pintado, D. (2008). Diffusion in complex social networks. *Games and Economic Behavior 62*(2), 573–590.

Mazzoni, E. (2006). Du simple tracement des interactions à l'évaluation des rôles et des fonctions des membres d'une communeauté en réseau : une proposition dérivée de l'analyse des

réseaux sociaux.

Mercklé, P. (2004). Les origines de l'analyse des réseaux sociaux.

Saito, K., M. Kimura, K. Ohara, et H. Motoda (2009). Learning continuous-time information diffusion model for social behavioral data analysis. In *ACML '09 : Proceedings of the 1st Asian Conference on Machine Learning*, Berlin, Heidelberg, pp. 322–337. Springer-Verlag.

Saito, K., R. Nakano, et M. Kimura (2010). Prediction of information diffusion probabilities for independent cascade model. In I. Lovrek, R. J. Howlett, et L. C. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems*, Volume 5179 of *Lecture Notes in Computer Science*, Chapter 9, pp. 67–75. Berlin, Heidelberg : Springer Berlin Heidelberg.

Trottier, H. et P. Philippe (2001). Deterministic modeling of infectious diseases : Theory and methods. *The Internet Journal of Infectious Diseases 1*.

Young, H. P. (2009). Innovation diffusion in heterogeneous populations : Contagion, social influence, and social learning. Open access publications from university of oxford, University of Oxford.

Summary

Social networks are becoming one of the preferred tools to communicate and share information, and several studies have been conducting on social networks to understand how the information is propagated in such networks. Such studies bear strong similarities with the way diseases are propagated in populations. We first present here an information diffusion model which has the particularity to take into account (a) neighbors influence and (b) the similarity bewteen users profile and the content propagated to determine how a document is disseminated over a social network. We then illustrate this model on several, mainly artificial, networks.