Modéliser l'utilisateur pour la diffusion de l'information dans les réseaux sociaux

Cédric Lagnier, Éric Gaussier, François Kawala

▶ **To cite this version:**

Cédric Lagnier, Éric Gaussier, François Kawala. Modéliser l'utilisateur pour la diffusion de l'information dans les réseaux sociaux. Revue des Sciences et Technologies de l'Information - Série ISI : Ingénierie des Systèmes d'Information, 2012, 17 (3), pp.1-22. hal-00741416

HAL Id: hal-00741416
https://hal.science/hal-00741416

Submitted on 12 Oct 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Modéliser l'utilisateur pour la diffusion de l'information dans les réseaux sociaux

Cédric Lagnier [1], **Eric Gaussier** [1], **François Kawala** [1,2]

1. *Université Joseph Fourier / Grenoble 1 / CNRS*
   Laboratoire LIG
   Bat. CE4, allée de la Palestine
   38610 GIERES
   {cedric.lagnier,eric.gaussier,francois.kawala}@imag.fr

2. *Société BestofMedia, 4 rue des méridiens*
   Immeuble Le Calypso, PARC SUD GALAXIE
   Echirolles, 38130
   fkawala@bestofmedia.com

ABSTRACT. *Predicting information diffusion in social networks is a hard task which can lead to interesting applications: recommending relevant information for users, choosing the best entry points in the network for the best diffusion of a given piece of information, etc. We present new models which take into account three main characteristics: the number of neighbors who have disclosed the information, the relevance of the information for each user and the willingness of users to diffuse information. After this presentation, we propose to estimate the parameters of our models and illustrate their behavior through a comparison with standard information diffusion models on a real dataset. We also propose a study of the influence maximization problem associated with these new models.*

RÉSUMÉ. *Prédire la diffusion d'information dans les réseaux sociaux est une tâche difficile qui peut cependant permettre de répondre à des problèmes intéressants : recommandation d'information, choix des meilleurs points d'entrée pour une diffusion, etc. Nous présentons de nouveaux modèles de diffusion qui tiennent compte de trois caractéristiques : le nombre de voisins ayant déjà diffusé l'information, l'intéret que l'utilisateur peut porter à l'information et la tendance d'un utilisateur à diffuser. Après cette présentation, nous proposons une méthode pour estimer les paramètres de nos modèles et illustrons leur comportement sur un jeu de données réel à travers une comparaison avec des modèles standards de diffusion de l'information. Nous proposons aussi une étude de la maximisation de l'influence associée à ces nouveaux modèles.*

KEYWORDS: *Social Networks, Information Diffusion, Machine Learning*

MOTS-CLÉS : *Réseaux Sociaux, Diffusion d'Information, Apprentissage Automatique*

DOI:10.3166/ISI.22.1.1-22 © 2012 Lavoisier

1. Introduction

Les modèles de propagation ont pour but de reproduire les phénomènes que l'on peut observer dans les réseaux sociaux, mais aussi dans le marketing viral ou dans la propagation de maladies. La communication entre les utilisateurs acteurs de ces réseaux donne lieu à un certain nombre de problématiques comme la découverte de pôles d'influence, le choix des diffuseurs initiaux pour une diffusion maximale, ou encore l'identification des liens à supprimer pour limiter la diffusion.

La plupart des modèles récemment proposés pour la diffusion d'information sont des extensions des modèles à cascades indépendantes (IC - *Independent Cascade*) (Goldenberg *et al.*, 2001) et des modèles à seuil linéaire (LT - *Linear Threshold*) (Granovetter, 1978). De telles extensions sont pas exemple proposées dans (Prakash *et al.*, 2010 ; Saito *et al.*, 2011). Si de tels modèles peuvent bel et bien être utilisés pour modéliser ou prédire la diffusion d'information dans un réseau social, ils présentent néanmoins un certain nombre de défauts :

– Tout d'abord, ils ne tiennent pas compte du contenu de l'information diffusée, alors même que cette information semble cruciale dans plusieurs cas. En particulier, au sein d'un même réseau social, deux informations différentes se propageront de façon différente selon les champs d'intérêt des utilisateurs impliqués dans la diffusion ;

– Ensuite, il ne tiennent pas compte des caractéristiques des utilisateurs du réseau social : quels sont les centres d'intérêt de tel ou tel utilisateur, quels sont les rôles (actif, passif) joués par tel ou tel utilisateur dans le réseau social ?

– Enfin, ils reposent sur des hypothèses fortes sur les processus de diffusion, hypothèses que ne sont pas forcément vérifiées en pratique.

Nous présentons ici une famille de modèles "centrée utilisateur" qui revient sur un certain nombre de ces points. En particulier, ces modèles intègrent (a) la pression sociale, mesurée à partir du nombre de voisins actifs, (b) l'intérêt d'un utilisateur pour l'information diffusée, mesuré à partir de la similarité entre le contenu de l'information diffusée et les centres d'intérêt de l'utilisateur, et (c) le rôle de chaque utilisateur, caractérisé par sa propension à rediffuser une information.

La suite de cet article est organisée de la façon suivante : la prochaine section décrit différents travaux existant dans le domaine de la diffusion d'information, de façon à mieux situer notre approche. La section 3 décrit les modèles centrés utilisateur que nous introduisons dans cet article. Dans la section 4, nous présentons des expériences que nous avons faites pour valider la qualité de ces nouveaux modèles en comparant leurs résultats avec des modèles standards. La section 5 traite du problème de la maximisation de l'influence dans ces modèles. Enfin, la section 6 conclut notre étude en rappelant les principales contributions réalisées et en ébauchant un certain nombre de perspectives.

2. Travaux reliés

Nous classons les modèles de diffusion de l'information en deux grandes catégories : modèles de contagion et modèles d'influence sociale. Une troisième catégorie, correspondant aux modèles d'apprentissage social, est parfois considérée (Young, 2009). Cette troisième catégorie repose sur le fait que l'adoption d'un produit par un utilisateur, par exemple, dépend de l'utilité observée du produit pour d'autres utilisateurs (Munshi, 2004). S'il existe un certain nombre de travaux qui tentent d'inclure un paramètre d'utilité dans des modèles de contagion ou d'influence, ils reposent en général sur des hypothèses fortes (comme le fait que l'utilité d'un produit adopté par un utilisateur est quantifiable et connue) qui sont difficilement réalisables en pratique, et ne s'appliquent pas directement aux réseaux de contenu. Outre ces deux grandes catégories, contagion et influence sociale, deux approches sont généralement utilisées pour modéliser la diffusion au sein de réseaux. La première approche consiste à établir des équations différentielles (ou plus généralement aux différences) régissant l'évolution du réseau au cours du temps. La deuxième approche consiste à modéliser la diffusion étape par étape à travers un mécanisme explicite. Comme nous le verrons, ces deux approches mènent parfois aux mêmes modèles. Enfin mentionnons que certains travaux s'intéressent à la dynamique globale de diffusion, la variable étudiée étant alors le taux d'utilisateurs actifs et son évolution au cours du temps, et d'autres à une dynamique plus locale, où le statut de chaque utilisateur (inactif ou actif) est étudié à chaque étape de temps. Bien évidemment, la dynamique globale de la diffusion peut être déduite de la dynamique locale, mais pas le contraire.

Dans les modèles de contagion, les utilisateurs s'activent dès qu'ils sont en contact avec une personne active. Les modèles de contagion hérités de l'épidémiologie s'intéressent en général à la dynamique globale de la contagion et sont fondés sur des équations différentielles qui régissent le passage d'un état sain à un état infecté (modèle SI - Susceptible/Infected) voire à d'autres états au cours du temps (comme le modèle SIR - Susceptible/Infected/Recovered). (Trottier, Philippe, 2001), (Newman, 2003) et (Brauer, Castillo-Chavez, 2001) fournissent une bonne description de ces modèles ; (López-Pintado, 2008) ou (Young, 2009) proposent plusieurs variantes de ces modèles dans divers cadres. Il est en fait possible d'appliquer ces mêmes équations différentielles localement en considérant des taux d'infection non plus constants sur le réseau mais dépendant des utilisateurs en contact. Le modèle SI devient alors très proche du modèle IC (voir (Kimura *et al.*, 2007) pour cette relation). Le modèle IC (Independent Cascade) (Goldenberg *et al.*, 2001) est un modèle de cascades indépendantes qui a suscité un grand nombre de développements. Il est fondé sur le principe simple suivant : dès qu'un nœud u est actif, il a une unique chance d'activer chacun de ses voisins directs v, et ce avec une probabilité $P_{u,v}$. Que cette activation réussisse ou échoue, u ne sera plus à même de contaminer v par la suite. Ce modèle est un modèle chronologique qui procède par étapes d'activation. Les paramètres $P_{u,v}$ peuvent être appris par maximum de vraisemblance (Saito *et al.*, 2008), après avoir observé un certain nombre de diffusions. Tout comme le modèle SI déjà cité, on peut montrer que le modèle IC correspond à un processus de percolation de liens (processus de pro-

pagation principalement utilisé en physique) sur le graphe du réseau social considéré ((Newman, 2003), (Kempe *et al.*, 2003), (Kimura *et al.*, 2007)).

Le modèle IC a récemment été étendu de façon à avoir un modèle continu en temps (et pas seulement fondé sur des étapes d'activation) et à corriger le fait que l'activation à partir d'un nœud ne peut avoir lieu que lorsque ce nœud vient d'être activé. Le modèle ASIC (Asynchronous IC), (Saito *et al.*, 2009), introduit un délai, régi par une distribution exponentielle, entre le moment où un utilisateur devient actif et celui où il active ses voisins, la probabilité d'activation décroissant avec le temps. L'algorithme EM (Expectation-Maximization) de maximisation de la vraisemblance peut être utilisé pour estimer les paramètres du modèle. Plus récemment, (Gomez-Rodriguez *et al.*, 2011) considèrent différentes distributions de probabilité pour le delai dans la contamination : exponentielle, loi de puissance et distribution de Rayleigh. La famille de modèles qu'ils ont définit est appelé NetRate. La version basée sur la distribution exponentielle est en fait un cas particulier du modèle ASIC (obtenu quand les paramètres $k_{u,v}$ sont fixés comme des constantes). Un des avantages des modèles considérés dans ce dernier travail est que l'estimation de leurs paramètres est un problème d'optimisation du maximum de vraissemblance avec des contraintes de paramètres positifs pour lequel la fonction de vraissemblance est convexe impliquant une unique solution qui peut être trouvée avec des méthodes d'optimisation standards. Elles ne souffrent donc pas du problème de maximum local que l'on peut retrouver pour l'optimisation de la méthode ASIC. Le but original de NetRate est la prédiction de liens (elle peut être vue comme une extension de NetInf ((Gomez-Rodriguez *et al.*, 2010))). Malgré leur but original, cette famille de modèles peut être utilisée pour prédire des diffusions étant donné que les probabilités calculées pour estimer les liens sont directement basées sur les probabilités de diffusion dans le réseau.

Dans les modèles d'influence sociale, également appelés modèles à seuil, un individu est activé si le nombre ou la proportion de ses voisins déjà activés est supérieur à un seuil qui lui est propre. C'est donc la pression sociale qui est déterminante ici pour l'activation. Les premiers travaux sur ces modèles sont décrits dans (Schelling, 1971) et (Granovetter, 1978) - le nom de modèle LT (Linear Threshold) est souvent associé au modèle de Granovetter. Ils ont depuis été repris et étendus dans (Granovetter, Soong, 1988), (Macy, 1991), (T. W. Valente, 1995), (T. Valente, 1996), (Abrahamson, Rosenkopf, 1997), (Richardson, Domingos, 2002), (Dodds, Watts, 2004), (López-Pintado, Watts, 2008) et (Borodin *et al.*, 2010) par exemple. Dans la version la plus courante du modèle LT, un nœud v est activé si la somme des poids des liens entrant est supérieure à un seuil θ_v propre à v, choisi de façon aléatoire dans de nombreuses instanciations de ce modèle (voir (Kempe *et al.*, 2003) par exemple). La dynamique globale de diffusion engendrée par ces modèles diffère en partie de celle observée pour le modèle IC. En particulier, il est possible de montrer sous certaines conditions (voir par exemple (Young, 2009)) que lorsque la diffusion augmente, elle le fait de façon super-exponentielle. Toutefois, comme pour le modèle IC, on peut montrer (Kempe *et al.*, 2003) que le modèle LT est équivalent à un processus de percolation de liens sur le graphe du réseau, ce qui place ce modèle et ses extensions dans la même classe générale que celle des modèles de contagion, $i.e.$ la classe des modèles de percolation.

Enfin, des versions généralisées des modèles IC et LT sont proposées dans (Kempe *et al.*, 2003). La généralisation du modèle IC permet de tenir compte de la "pression sociale" : quand un nœud u est activé et tente d'activer un de ses voisins v, il le fait avec une probabilité $P_{u,v}(S)$ qui tient compte de l'ensemble S des voisins de v qui ont déjà tenté d'activer v et échoué. Pour que cette probabilité rende bien compte de la pression sociale, il faut que la probabilité $P_{u,v}(S)$ soit croissante avec la taille de S. On suppose de plus qu'elle est indépendante de l'ordre avec lequel les voisins de v qui ont échoué dans leur tentative d'activation sont considérés. Le modèle IC est un cas particulier de cette cascade généralisée obtenu en considérant $P_{u,v}(S)$ constant et égal à $P_{u,v}$. Pour le modèle LT, c'est la fonction de combinaison des poids des liens entrant en v qui est généralisée : le nœud v est activé à l'étape n si $f_v(S) \geq \theta_v$, où S est l'ensemble des voisins de v actifs à l'étape $n-1$ et f est une fonction croissante avec la taille de S. Le modèle LT correspond au cas où f est la somme des poids des liens entre les éléments de S et v. Ces deux généralisations permettent d'une part d'obtenir un mécanisme de cascade tenant compte de la pression sociale, et d'autre part d'établir un pont entre les modèles de contagion et les modèles d'influence. En effet, (Kempe *et al.*, 2003) montrent que chaque modèle à cascade généralisée peut être reformulé comme un modèle à seuil généralisé équivalent, et vice versa.

Tous les modèles présentés ignorent un certain nombre de facteurs cruciaux pour la diffusion d'information dans les réseaux sociaux : (a) l'intérêt d'un utilisateur pour le contenu diffusé, et (b) le rôle que les utilisateurs prennent dans les réseaux sociaux (actif ou passif). Les modèles que nous introduisons dans cet article prennent en compte ces facteurs supplémentaires.

Enfin, nous voulons terminer ce survol des travaux reliés en mentionnant le problème de maximisation/minimisation de l'influence, problème qui consiste à déterminer, pour un réseau, un type d'information et un nombre k donnés, les k diffuseurs initiaux qui maximisent/minimisent la diffusion de l'information. Ce problème, originellement étudié dans (Domingos, Richardson, 2001) puis dans toute une série de travaux depuis (comme (Kempe *et al.*, 2003), (Kimura *et al.*, 2007) ou (Leskovec *et al.*, 2007)) est connu, pour les modèles étudiés, pour être NP-difficile. Il est donc nécessaire de trouver des heuristiques qui fournissent de bonnes approximations de la solution optimale.

3. Modèles centrés utilisateur

3.1. Notations

Nous travaillons sur des graphes sociaux dirigés $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ composés d'un ensemble de nœuds/utilisateurs $\mathcal{N} = \{n_1, ..., n_N\}$ et d'un ensemble de liens $\mathcal{E}$. Un utilisateur n_i est relié à un utilisateur n_j si $(n_i, n_j) \in \mathcal{E}$. Nous utilisons les définitions/notations suivantes :

– nous parlerons de voisins entrants ou de voisins sortants en fonction du sens des liens. Soit $\mathcal{B}(n_i)$ l'ensemble des voisins entrants de l'utilisateur n_i (les utilisateurs qui

ont un lien vers n_i) :

$$\mathcal{B}(n_i) = \{n_j/(n_j, n_i) \in \mathcal{E}\}$$

De la même manière, soit $\mathcal{F}(n_i)$ l'ensemble des voisins sortants de l'utilisateur n_i (les utilisateurs vers qui n_i a un lien) :

$$\mathcal{F}(n_i) = \{n_j/(n_i, n_j) \in \mathcal{E}\}$$

– tous les utilisateurs ont un profil décrivant leurs centres d'intérêt. Soit $\mathcal{U}$ l'ensemble des profils des utilisateurs et $\forall i, 1 \leq i \leq N, u_i$ représente le profil de l'utilisateur n_i. Le profil est en général un vecteur de termes déduit des documents écrits ou (re-)diffusés par l'utilisateur ;

– $\mathcal{Q} = (q_1, ..., q_M)$ est l'ensemble des différents contenus qui se propagent à travers le réseau. Sur un réseau comme Twitter, $\mathcal{Q}$ correspond à l'ensemble de tous les tweets. Un élément de $\mathcal{Q}$ sera indifférement appelé contenu, requête ou information. Un contenu est codé de la même façon que les profils utilisateurs.

– Dans les modèles sur lesquels nous travaillons, nous nous intéressons aux utilisateurs qui diffusent un contenu, et, quand cela est fait, on dira qu'ils sont actifs (ou contaminés). Dans les processus que nous modélisons, il n'y a pas de retour-arrière (un utilisateur qui est actif ne peut redevenir inactif ; en d'autres termes, il ne peut nier avoir diffusé l'inforamtion).

3.2. Modèles

Nous proposons ici l'utilisation de trois facteurs caratérisant la diffusion d'information dans les réseaux sociaux. Il s'agit de :
(a) la "pression sociale" que chaque utilisateur subit, qui peut être mesurée par le nombre de voisins entrants de l'utilisateur qui sont déjà contaminés,
(b) l'intérêt d'un utilisateur pour un contenu donné, que l'on peut mesurer par la similarité entre son profil et le contenu, et
(c) le rôle, actif ou passif, qu'un utilisateur joue dans le réseau, qui peut être mesuré en fonction de l'activité d'un utilisateur dans les diffusions passées.
L'influence de la pression sociale a été étudiée par le passé, notamment dans le cadre des modèles à seuil linéaire ; il n'existe pas à notre connaissance de modèles intégrant les deux autres facteurs. Nous allons maintenant introduire de manière plus formelle ces idées et présenter une famille de modèles prenant en compte ces facteurs.

Dans nos modèles, nous définissons la probabilité $P_c(n_i, q_k, t)$ qu'un utilisateur n_i soit contaminé au temps t par un contenu q_k comme la fonction de seuil suivante :

$$P_c(n_i, q_k, t) = \begin{cases} (1 + e^{-\lambda_1(S(n_i,q_k;\theta_s) - \lambda_2 E[|\mathcal{C}^k(n_i,t)|] - \lambda_3 W(n_i;\theta_w)})^{-1} \\ \qquad\qquad \text{si } E[|\mathcal{C}^k(n_i,t)|] > 0 \\ 0 \text{ sinon} \end{cases}$$

où :

– les trois paramètres λ_1, λ_2 et λ_3 de la fonction expontentielle contrôlent l'influence de chacun des facteurs de la contamination (facteurs (a), (b) et (c) introduits précédement). Chaque paramètre de la fonction agit comme un critère de seuil pour l'activation de l'utilisateur n_i ;

– $S(n_i, q_k; \theta_s) = sim(u_i, q_k) - \theta_s$, $sim(u_i, q_k)$ représente la similarité entre le contenu diffusé q_k et le profil de l'utilisateur u_i. θ_s est un seuil qui permet de diminuer la probabilité d'activation si l'intérêt de l'utilisateur pour le contenu est trop faible (c'est-à-dire si la similarité est inférieure à θ_s). Dans notre étude, u_i est un vecteur dans $\mathbb{R}^v$ correspondant à une moyenne de tous les contenus diffusés par le passé par l'utilisateur n_i. De plus, nous utilisons une mesure de similarité cosinus pour la fonction sim, tout en gardant à l'esprit que d'autres choix sont possibles ;

– $\mathcal{C}^k(n_i, t)$ est l'ensemble des voisins entrants de l'utlisateur n_i qui sont actifs au temps t. $E[|\mathcal{C}^k(n_i, t)|]$ correspond donc à l'espérance du nombre de voisins entrants de l'utilisateur n_i qui sont déjà contaminés par le contenu q_k au temps t. Dans le cas où l'état d'un utilisateur est binaire (actif ou inactif), cette espérance correspond au nombre de voisins entrants actifs ;

– $W(n_i; \theta_w) = act(n_i) - \theta_w$, où act est une mesure de l'activité de l'utilisateur n_i dans les diffusions passées ; si cette activité est inférieure au seuil θ_w (c'est le cas pour les utilisateurs passifs dans le réseau), la probabilité d'activation en est diminuée. Nous définissons $act(n_i)$ comme le ratio du nombre de contenus reçus et (re-)diffusés par l'utilisateur n_i sur le nombre total de contenus reçus par n_i dans son activité passée. D'autres choix sont bien sûrs possibles, fondés par exemple sur une connaissance *a priori* de l'activité des utilisateurs dans d'autres réseaux sociaux.

L'estimation des paramètres λ_1, λ_2, λ_3, θ_s et θ_w est décrite dans la setion 3.3. La probabilité de non contamination d'un utilisateur, $P_{nc}(n_i, q_k, t)$, est simplement $1 - P_c(n_i, q_k, t)$. Nous utilisons le terme "centré utilisateur" (User-Centric) pour parler de ces modèles car, mis à part la pression sociale, toutes les informations utilisées sont fondées sur les caractéristiques des utilisateurs. Nous allons maintenant décrire trois modèles à cascade que nous avons définit qui intègrent ces éléments :

– un modèle à cascades simple prenant en compte ces caractéristiques

– un modèle à cascade plus complexe, prenant en compte le temps

– un dernier modèle pour palier une défaillance du second modèle sur le long terme

3.2.1. UC

Dans le modèle UC (pour *User-Centric model*), chaque utilisateur n_j, contaminé au temps t, a une et une seule chance de contaminer chacun de ses voisins sortants n_i au temps $t + 1$ sur la base de la probabilité $P_c(n_i, q_k, t)$. S'il y arrive, une valeur de contamination de 1 est associée à n_i à partir du temps $t + 1$, 0 sinon. Ce processus est similaire à celui du modèle à cascade indépendente IC (Independant Cascade), mais en diffère cependant par le fait que si n_j échoue dans la contamination de n_i, il contribuera par la suite, dans le cas où un autre voisin entrant de n_i tente d'activer

ce dernier, par l'intermédiaire de $E[|\mathcal{C}^k(n_i,t)|]$. Dans cette configuration, le nombre de voisins actifs est connu et nous avons $E[|\mathcal{C}^k(n_i,t)|] = |\mathcal{C}^k(n_i,t)|$. Le processus de contamination se termine lorsque plus rien ne change dans le réseau, c'est-à-dire lorsqu'il n'y a plus de contamination.

3.2.2. RUC

Dans le modèle RUC (pour *Reinforced User-Centric model*), à la différence du modèle UC, un utilisateur n'est pas contaminé ou non contaminé, mais a une probabilité d'être contaminé qui évolue au cours du temps en fonction de l'environnement de l'utilisateur. Soit $P_c(n_i,q_k,\leq t)$ la probabilité que l'utilisateur n_i soit contaminé par un contenu q_k avant le temps t (la probabilité de non contamination est donc $P_{nc}(n_i,q_k,\leq t) = 1 - P_c(n_i,q_k,\leq t)$). L'équation suivante définit l'évolution au cours du temps de cette quantité :

$$P_c(n_i,q_k,\leq t+1) = P_c(n_i,q_k,\leq t) + [1 - P_c(n_i,q_k,\leq t)]\, P_c(n_i,q_k,t) \qquad (1)$$

En d'autres mots, un utilisateur contaminé avant le temps $t+1$ a soit été contaminé avant le temps t, soit a été contaminé au temps t. Par définition, $P_c(n_i,q_k,\leq 0) = 1$ pour les diffuseurs initiaux, et vaut 0 pour les autres. Par récurrence on obtient :

$$P_c(n_i,q_k,\leq t) = \sum_{t'=0}^{t-1} P_c(n_i,q_k,t') \prod_{\tau=0}^{t'-1} (1 - P_c(n_i,q_k,\tau)) \qquad (2)$$

Contrairement au modèle UC, le modèle RUC n'étant pas binaire, on n'a plus directement accès au nombre de voisins entrants d'un utilisateur donné à un moment donné. Il est nécessaire ici de calculer explicitement l'espérance du nombre de voisins actifs $E[|\mathcal{C}^k(n_i,t)|]$. La valeur de cette espérance est définie par : $E[|\mathcal{C}^k(n_i,t)|] = \sum_{m=0}^{|\mathcal{B}(n_i)|} m\ P(|\mathcal{C}^k(n_i,t)| = m)$, où $P(|\mathcal{C}^k(n_i,t)| = m)$ est la probabilité que le nombre de voisins entrants de l'utilisateur n_i qui sont actifs au temps t soit égal à m. On peut montrer :

$$E[|\mathcal{C}^k(n_i,t)|] = \sum_{n_j \in \mathcal{B}(n_i)} P_c(n_j,q_k,\leq t) \qquad (3)$$

Le principal problème de ce modèle est que les probabilités $P_c(n_j,q_k,\leq t)$ ne peuvent décroître et augmenteront dès lors que $P_c(n_j,q_k,t)$ est non nulle. Ce phénomène est dû au fait que chaque utilisateur garde une influence forte sur ses voisins alors même qu'il a pu diffuser l'information dans un passé lointain. Nous corrigeons ce problème dans le modèle suivant.

3.2.3. DRUC

Dans le modèle DRUC (pour *Decaying Reinforced User-Centric model*), nous introduisons un nouveau paramètre pour diminuer l'influence qu'ont les voisins entrants ayant diffusé un contenu il y a longtemps. Ce paramètre permet donc de rendre compte

du fait que plus une information est récente, plus il y a de chances qu'un utilisateur veuille la relayer.

Pour cela, nous introduisons l'influence $I(n_j, q_k, t)$ d'un utilisateur n_j sur ses voisins sortants au temps t pour un contenu q_k :

$$I(n_j, q_k, t+1) = \alpha \times I(n_j, q_k, t) + [1 - P_c(n_j, q_k, \leq t)]\, P_c(n_j, q_k, t) \quad (4)$$

où $0 \leq \alpha \leq 1$, et $I(n_j, q_k, 0) = 1$ pour les diffuseurs initiaux et 0 pour les autres. Pour $\alpha < 1$, l'influence qu'un utilisateur perçoit de ses voisins entrants diminue avec le temps, et ce jusqu'à ce qu'un nouveau voisin soit contaminé. On peut alors redéfinir l'espérance du nombre de voisins entrants contaminés par :

$$E[|\mathcal{C}^k(n_i, t)|] = \sum_{n_j \in \mathcal{B}(n_i)} I(n_j, q_k, t) \quad (5)$$

Les valeurs $P_c(n_j, q_k, \leq t)$ correspondent toujours aux équations 1 and 2. Dans le cas particulier où $\alpha = 1$, on voit que $I(n_j, q_k, t) = P_c(n_j, q_k, \leq t)$ et le modèle se comporte comme le modèle RUC.

3.3. Estimation des paramètres

Pour fixer la valeur du paramètre de seuil θ_s, nous calculons dans un premier temps, sur un ensemble d'entraînement constitué des diffusions passées, la similarité cosinus entre chaque contenu et chaque utilisateur, qu'il soit contaminé ou pas par le contenu. Nous pouvons ensuite, pour chaque valeur β dans $[0; 1]$, determiner le nombre (moyenné sur tous les contenus de l'ensemble d'entrainement) d'utilisateurs actifs et inactifs qui ont une similarité plus grande que β. Nous cherchons enfin (à travers une recherche par ligne de pas de 0.05) la valeur au-delà de laquelle il y a plus d'utlisateurs actifs que d'utilisateurs inactifs. Le seuil θ_s est fixé à cette valeur, qui correspond donc à la valeur de similarité au-dessus de laquelle un utilisateur a plus de chances d'être actif qu'inactif. Un raisonnement similaire sur la propension à diffuser de chaque utilisateur conduit à fixer θ_w à 0.5.

Pour les paramètres λ_1, λ_2 et λ_3, nous utilisons un critère de maximisation de la vraisemblance sous contraintes. Soit $\mathcal{L}(\lambda_1, \lambda_2, \lambda_3)$ la vraissemblance calculée sur l'ensemble d'entraînement. Le problème à résoudre est le suivant :

$$\begin{cases} \text{argmax}_{\lambda_1, \lambda_2, \lambda_3} \mathcal{L}(\lambda_1, \lambda_2, \lambda_3) \\ \text{avec : } \lambda_1 \geq 0,\ \lambda_2 \geq 0,\ \lambda_3 \geq 0 \end{cases}$$

où les contraintes de positivité sont dictées par le choix de la fonction de contamination sur laquelle sont fondés nos modèles. Ces contraintes étant des contraintes "d'intervalles", nous pouvons utiliser la méthode du gradient projeté pour résoudre ce problème et estimer les valeurs des paramètres. Cette méthode consiste à effectuer une projection des valeurs des paramètres dans les intervalles admissibles après chaque étape de descente/montée de gradient. On obtient dans notre cas, après calcul des dérivées

partielles, les formules de mise à jour suivantes (entre les étapes p et $p+1$, en notant $\mathcal{LL}$ la log-vraissemblance) :

$$\forall i \in \{1,2,3\} : \left\{ \begin{array}{l} \lambda_i^{(p+1)} = \lambda_i^{(p)} + \gamma \frac{\partial \mathcal{LL}(\lambda_1^{(p)}, \lambda_2^{(p)}, \lambda_3^{(p)})}{\partial \lambda_i} \\ \text{Si } \lambda_i^{(p+1)} < 0, \text{ alors } \lambda_i^{(p+1)} = 0 \end{array} \right.$$

où γ contrôle le pas de descente le long du gradient de $\mathcal{LL}$.

La vraissemblance pour le modèle UC est donnée par :

$$\mathcal{L}(\lambda_1, \lambda_2, \lambda_3) = \prod_{k=1}^{|\mathcal{Q}|} \prod_{t=1}^{T^k} [\prod_{n_i \in D^k(t)} P_c(n_i, q_k, t-1) \prod_{n_i \in D^k(t-1)} \prod_{n_j \in F(n_i) \setminus C^k(t)} (1 - P_c(n_j, q_k, t-1))]$$

où $D^k(t)$ est l'ensemble des utilisateurs qui ont été contaminés au temps t et $C^k(t)$ est l'ensemble des utilisateurs qui ont été contaminés avant le temps t. On a : $C^k(t) = \cup_{t'=0}^{t} D^k(t')$.

La vraissemblance pour les modèles RUC et DRUC prend une forme plus simple, fondée sur la probabilité de chaque utilisateur d'être actif à chaque étape de temps. Elle s'exprime de la façon suivante :

$$\mathcal{L}(\lambda_1, \lambda_2, \lambda_3) = \prod_{k=1}^{|\mathcal{Q}|} \prod_{t=1}^{T^k} [\prod_{n_i \in C^k(t)} P_c(n_i, q_k, \leq t) \prod_{n_i \notin C^k(t)} (1 - P_c(n_i, q_k, \leq t)]$$

Afin de réduire le coût du calcul du gradient de ces vraisemblances à chaque étape, nous utilisons l'équation 1 pour calculer les dérivés partielles, puis stockons, pour chaque utilisateur, les valeurs des probabilités $P_c(n_j, k, \leq t)$ et de leurs dérivées à chaque étape de temps. La dérivée de l'équation 1 est la suivante :

$$\begin{aligned} \frac{\partial P_c(n_i, q_k, \leq t+1)}{\partial \lambda_i} = & \quad \frac{\partial P_c(n_i, q_k, \leq t)}{\partial \lambda_i} (1 - P_c(n_i, q_k, t)) \\ & + \frac{\partial P_c(n_i, q_k, t)}{\partial \lambda_i} (1 - P_c(n_i, q_k, \leq t)) \end{aligned} \quad (6)$$

et les dérivées des équations pour la mise à jour des probabilités à chaque étape si $E[|\mathcal{C}^k(n_i,t)|] > 0$:

$$\frac{\partial P_c(n_i,q_k,t)}{\partial \lambda_1} = \frac{(S(n_i,q_k;\theta_s))(e^{-\lambda_1(S(n_i,q_k;\theta_s)-\lambda_2 E[|\mathcal{C}^k(n_i,t)|]-\lambda_3 W(n_i;\theta_w)})}{(1+e^{-\lambda_1(S(n_i,q_k;\theta_s)-\lambda_2 E[|\mathcal{C}^k(n_i,t)|]-\lambda_3 W(n_i;\theta_w)})^2} \quad (7)$$

$$\frac{\partial P_c(n_i,q_k,t)}{\partial \lambda_2} = \frac{(\frac{\partial E[|\mathcal{C}^k(n_i,t)|]}{\partial \lambda_2})(e^{-\lambda_1(S(n_i,q_k;\theta_s)-\lambda_2 E[|\mathcal{C}^k(n_i,t)|]-\lambda_3 W(n_i;\theta_w)})}{(1+e^{-\lambda_1(S(n_i,q_k;\theta_s)-\lambda_2 E[|\mathcal{C}^k(n_i,t)|]-\lambda_3 W(n_i;\theta_w)})^2} \quad (8)$$

$$\frac{\partial P_c(n_i,q_k,t)}{\partial \lambda_3} = \frac{(W(n_i;\theta_w))(e^{-\lambda_1(S(n_i,q_k;\theta_s)-\lambda_2 E[|\mathcal{C}^k(n_i,t)|]-\lambda_3 W(n_i;\theta_w)})}{(1+e^{-\lambda_1(S(n_i,q_k;\theta_s)-\lambda_2 E[|\mathcal{C}^k(n_i,t)|]-\lambda_3 W(n_i;\theta_w)})^2} \quad (9)$$

sinon $\frac{\partial P_c(n_i,q_k,t)}{\partial \lambda_i} = 0$.

4. Validation expérimentale

Nous voulons ici comparer les modèles développés en section 3.2 avec un certain nombre d'autres méthodes discutées en section 2 : IC, ASIC et NetRate. Le but de ces comparaisons est d'illustrer l'apport des nouveaux facteurs pris en compte dans les modèles centrés utilisateur.

4.1. Données

Table 1. Description des jeux de données

Jeu de données	Nb. utilisateurs	Nb. liens	Nb. termes	Nb. cascades	Durée
ICWSM	5000	17746	173014	30075	31j

Nous avons effectué les tests sur un jeu de données utilisé lors du concours de ICWSM 2009 (Burton *et al.*, 2009). Il s'agit d'un ensemble de billets provenant de blogs. Le contenu des billets regroupe à la fois le texte des billets et les liens, soit vers d'autres billets du jeu de donnés, soit vers des sites externes. Les blogs sont considérés comme étant les utilisateurs du réseau. La diffusion observée sur ce réseau est explicite par les liens entre billets : si un billet p_2 dans un blog b_2 a un lien vers un billet p_1 d'un blog b_1, on considère que b_2 a rediffusé un contenu venant de b_1. Une cascade est donc un ensemble de billets (et donc de blogs) qui sont tous liés les uns aux autres sous forme de graphe.
Nous avons effectué les opérations de pré-traitement suivantes :

- ne garder que les billets sur une durée de un mois ;
- filtrage des billets pour ne garder que ceux en anglais ;
- suppression des mots vides ;
- utilisation du stemmer de Porter pour la racinisation ;
- filtrage des mots qui apparaissent moins de cinq fois dans le jeu de données.

Nous avons ensuite sélectionné les 5000 utilisateurs ayant posté le plus de billets et n'avons gardé que les cascades dans lesquelles il y a une diffusion (dont la taille est supérieure à 1). Deux tiers des cascades (≈ 20000) sont ensuite utilisés pour l'entraînement des modèles, le tiers restant (≈ 10000) constituant l'ensemble de test. Les liens du graphe entre les utilisateurs ont été calculés à partir des cascades de l'ensemble d'entraînement : si un blog b_2 a publié au moins un billet contenant un lien vers un billet du blog b_1, alors on considère qu'il y a un lien de diffusion de b_1 vers b_2. Le profil des blogs a été calculé comme la moyenne des vecteurs de descripteurs des billets qu'ils ont "écrits". Le tableau 1 donne les détails de ce jeu de données.

4.2. Mesures d'évaluation

Afin de comparer les résultats des différentes méthodes de diffusion, nous avons sélectionné deux mesures d'évaluation :

– **l'erreur** entre la prédiction du modèle et la réalité : pour chaque cascade et chaque utilisateur, nous calculons la valeur absolue de la différence entre la probabilté que l'utilisateur soit actif annoncée par le modèle et la valeur réelle (1 s'il est actif, 0 s'il ne l'est pas). Pour avoir une meilleure lisibilité des résultats, nous n'avons pas normalisé cette mesure ;

– **les courbes précision-rappel** : pour chaque cascade, nous classons les utilisateurs par probabilité d'activation prédite par le modèle. Nous calculons ensuite la précision de la liste obtenue à chaque point de rappel (utilisateur réellement actif). Cette précision est ensuite moyennée sur toutes les cascades (cf. (Manning *et al.*, 2008)). Toutes les cascades n'ayant toutefois pas le même nombre de points de rappels, la variance des estimations sur les derniers points de rappel est plus importante que celle sur les premiers points.

4.3. Résultats

Les expériences ont été faites sur le jeu de données présenté en section 4.1 en utilisant les mesures d'évaluation présentées en section 4.2. L'estimation des paramètres des modèles centrés utilisateurs a été effectuée en utilisant les algorithmes présentés en section 3.3. Pour les autres modèles, l'estimation a été effectuée en utilisant des algorithmes EM pour les méthodes IC ((Saito *et al.*, 2008)) et ASIC ((Saito *et al.*, 2009)), et une descente de gradient avec contraintes pour NetRate ((Gomez-Rodriguez *et al.*, 2011)) - nous avons également utilisé un gradient projeté dans ce dernier cas.

Le tableau 2 montre les erreurs obtenues pour chacun des modèles. La colonne *diffuseurs* représente l'erreur sur les utilisateurs qui sont actifs et donc la capacité des modèles à trouver les acteurs d'une diffusion. La colonne *non diffuseurs* représente l'erreur sur tous les utilisateurs du réseau qui ne sont pas acteurs de la diffusion. Elle représente donc la tendance du modèle à surdiffuser. La colonne total est la somme des deux autres. Les valeurs présentées n'étant pas normalisées (comme précisé précédemment), la première ligne fournit l'erreur maximum que l'on peut avoir avec ce

Table 2. Erreur entre la prédiction du modèle et la réalité

	diffuseurs	non diffuseurs	total
Erreur max	9590	50×10^6	50×10^6
IC	6778	637	7415
UC	7177	2005	9183
RUC	**4954**	3248	8203
DRUC	5112	1584	**6696**
ASIC	8192	288	8480
NetRate	8582	**81**	8664

jeu de données. Toutes les valeurs sont sommées sur tous les utilisateurs et toutes les cascades.

La première remarque que l'on peut faire est que les modèles à cascades ont tendance à très peu diffuser l'information, ce qui explique leurs très bons résultats sur les *non diffuseurs*, mais entraîne aussi un fort taux d'erreur sur les *diffuseurs*. Les modèles centrés utilisateur ont un meilleur résultat sur les *diffuseurs* mais une erreur plus importante sur les *non diffuseurs*. Les modèles à cascades fondés sur le temps ont une diffusion encore plus faible que le modèle IC. Ceci est dû au fait qu'ils permettent de retarder une diffusion, éventuellement en dehors de la fenêtre temporelle considérée (31 jours ici, durée sur laquelle s'étend notre jeu de données). Enfin, nous remarquons une nette amélioration du modèle DRUC par rapport au modèle RUC sur l'erreur sur les *non diffuseurs*, due au fait que le paramètre d'oubli permet au modèle de stopper en partie la diffusion vers les utilisateurs les moins enclins à devenir actifs.

Afin de comparer ces méthodes sous un autre angle, nous proposons d'étudier leur précision aux différents points de rappel. La figure 1 montre les courbes de précision/rappel pour les six modèles que nous étudions. Malgré la différence de résultats pour ce qui est de l'erreur, les courbes des modèles RUC et DRUC sont très proches (quasiment confondues sur le graphe). Le modèle RUC diffuse vers beaucoup plus d'utilisateurs *non diffuseurs* sans pour autant le faire avec une plus grande force, ce qui permet aux utilisateurs *diffuseurs* de rester bien classés. Ils ont tout deux des résultats au dessus des autres modèles. Nous pouvons voir que les modèles IC et UC, fondés sur le même processus mais avec des probabilités différentes, ont des résultats similaires, même si le modèle centré utilisateur est légèrement meilleur que le modèle à cascades.

Au vu des résultats de ces expériences, nous pouvons constater que les modèles centrés utilisateur obtiennent de meilleurs résultats que les modèles à cascades. Malgré leur propension à sur-diffuser, leur meilleure modélisation du processus de diffusion permet *in fine* d'obtenir un classement des utilisateurs meilleurs que celui des modèles standards. De plus, pour la méthode DRUC, qui obtient une erreur sur les *non diffuseurs* plus basse que les autres méthodes centrées utilisateur, l'erreur totale est plus faible que celle de tous les autres modèles. Ces résultats valident donc le bien fondé de la famille des modèles centrés utilisateurs, et en particulier du modèle DRUC, qui

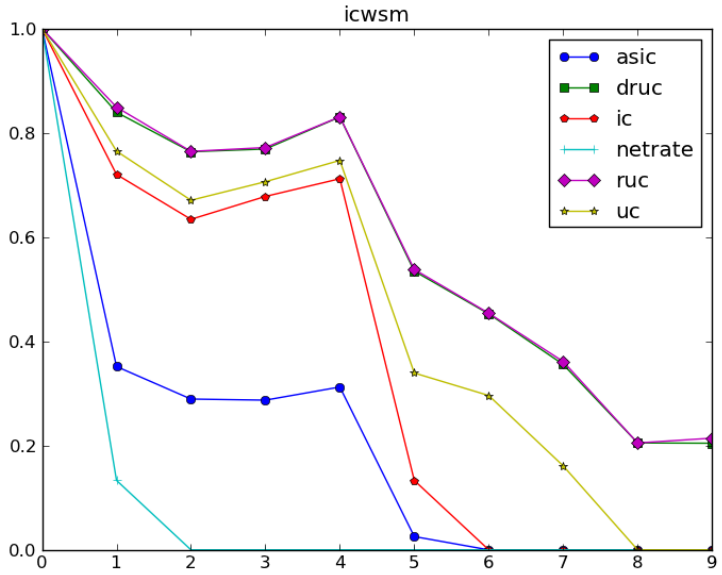

Figure 1. Précision pour chaque point de rappel

fournit les meilleures performances globales pour les deux mesures d'évaluation retenues.

Nous nous tournons maintenant vers le problème de maximisation de l'influence pour les modèles centrés utilisateur.

5. Maximisation de l'influence

Le problème de maximisation de l'influence vise à identifier les k diffuseurs initiaux qui maximisent la diffusion d'une information au sein d'un réseau social. Ce problème a été traité par Kempe *et al.* dans (Kempe *et al.*, 2003) pour différents modèles de diffusion de l'information. Nous montrons ici que, tout comme pour les autres modèles, ce problème est *NP-difficile* pour le modèle RUC. Nous nous concentrons sur ce modèle car les résultats sur IC se transposent directement à UC, et ceux sur RUC se transposent directement à DRUC. Cette étude débute par la définition des problèmes, et se poursuit par la preuve de complexité. Cette dernière montre l'existence d'une réduction polynomiale depuis le problème de couverture (Set-cover problem) vers le problème de décision associé à la maximisation de l'influence (dMI).

5.1. Définition informelle des problèmes

5.1.1. Problème de maximisation de l'influence

Le problème de maximisation de l'influence est un problème d'optimisation portant sur un "graphe-social" $\mathcal{G} = (Utilisateurs, Relations)$. L'objectif associé à ce problème est de déterminer le sous-ensemble de $Utilisateurs$ le plus influent. Ainsi l'activation de tous les membres de ce groupe d'instigateurs entraîne la plus grande quantité d'activation sur tout le "graphe-social". Cette quantité d'activation est ici décrite par $\sum_{n_i \in \mathcal{N}} P_c(n_i, q_k, \leq t)$. Ce problème est défini par deux paramètres supplémentaires : κ, la taille maximum de l'ensemble d'instigateurs et le modèle de diffusion de l'information (comme UC, RUC, DRUC) utilisé pour calculer la quantité d'activation. Le problème de décision correspondant admet un paramètre supplémentaire : une valeur minimale pour la quantité d'activation. Il répond d'une façon binaire : *vrai* s'il est possible de constituer un ensemble instigateur permettant d'atteindre la quantité d'activation minimum, *faux* dans le cas contraire.

5.1.2. Problème de couverture

Le problème de couverture (SC problem) est un problème de décision dont les paramètres sont : un ensemble $\mathcal{U}$ d'éléments nommé univers, $\mathcal{C}$ une collection de sous-ensembles de $\mathcal{U}$, et un entier κ tel que $\kappa \leq cardinal(C)$. Le résultat du problème de couverture est *vrai* si et seulement si il existe F une famille de sous-ensembles dans $\mathcal{C}$ telle que $cardinal(F) \leq \kappa$ et $\bigcup_{f \in F}(f) = \mathcal{U}$ (*i.e.* la famille F couvre l'univers). Le problème de couverture, une transformation depuis le problème X3C, est l'un des 21 problèmes NP-complet de Karp (Karp, 1972).

5.2. Complexité

Nous allons établir l'existence d'une réduction polynomiale de Karp depuis le problème de couverture vers le problème de décision associé à la maximisation de l'influence (dMI) lorsque le modèle de diffusion de l'information est probabiliste. Pour ce faire nous définissons Γ une application depuis les instances du problème SC vers les instances de problème dMI. Ces dernières comprennent un graphe social et un modèle de diffusion probabiliste spécifique. Si nous connaissions exactement la quantité d'utilisateurs dans l'état actif de la diffusion d'information, et ceci pour chaque étape de temps, nous pourrions répondre à une instance du problème de couverture en utilisant dMI, à condition de pouvoir déterminer la quantité d'activation minimale correspondant à la couverture de l'univers $\mathcal{U}$. Évidemment, le modèle de diffusion que nous utilisons doit être une instance particulière du modèle de diffusion pour lequel nous conduisons notre étude de complexité, ici RUC.

L'application Γ associe chaque élément de $\mathcal{U}$ et chaque élément de $\mathcal{C}$ à un nœud utilisateur du graphe social. Cependant ceux correspondant à un sous-ensemble de

l'univers, sont liés à tous les nœuds/utilisateurs représentant les éléments du sous-ensemble. La figure 2 présente un graphe social produit par l'application Γ.

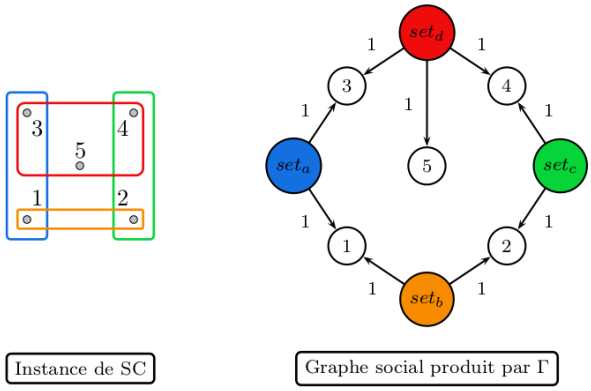

Figure 2. Côte-à-côte : une instance du problème de couverture et le "graphe-social" correspondant, par l'application Γ

L'une des propriétés remarquables de ce graphe social est de ne comporter que des chemins de longueur 1, et que tous les chemins commencent par des nœuds représentant des éléments de $\mathcal{C}$ et terminent par des nœuds représentant des éléments de $\mathcal{U}$. Par ailleurs nous utilisons le paramètre κ, contrôlant la taille de la famille de sous-ensembles pour le problème de couverture, comme valeur pour le paramètre κ contrôlant le nombre d'utilisateurs du groupe instigateur dans le problème dMI.
L'instance du modèle de diffusion RUC que nous cherchons à définir doit décrire, pour un ensemble instigateur donné, une diffusion déterministe sur notre graphe social. Pour ce faire, nous attribuons la valeur 0 à tous les paramètres du modèle, ainsi la probabilité pour un utilisateur de diffuser l'information à un temps donné devient :

$$P_c(n_i, q_k, t) = \begin{cases} (1 + exp(0))^{-1} = \frac{1}{2} & \text{si } n_i \in R_t \\ 0 & \text{Sinon} \end{cases}$$

où R_t désigne l'ensemble des utilisateurs atteignables au temps t, c'est-à-dire l'ensemble des utilisateurs qui ont au moins un voisin entrant ayant une probabilité non nulle de diffuser l'information avant l'étape de temps t.

Le graphe social produit par Γ nous assure ainsi que l'ensemble des utilisateurs est divisé en trois classes disjointes : les instigateurs, les atteignables, et les inatteignables. Cette distinction est effective dès la première étape de temps et n'évolue plus après celle-ci. Ceci implique pour l'instance spécifique de RUC que nous connaissions la probabilité d'activation de chaque nœud, pour chaque étape de temps. Ainsi nous distinguons trois classes de probabilité d'activation :

– Pour les nœuds appartenant à l'ensemble des instigateurs, $P_c(n_i, q_k, \leq 1) = 1$

– Pour les nœuds atteignables,
$P_c(n_i, q_k, \leq 1) = (1 - P_c(n_i, q_k, \leq 0)) * P_c(n_i, q_k, 0) = (1 - 0) * \frac{1}{2}$

– Pour les nœuds inatteignables, $P_c(n_i, q_k, \leq 1) = 0$

Il est possible de prouver par récurrence la propriété suivante : *pour chaque paire de nœuds (n_i, n_j) appartenant à l'une des ces classes, les utilisateurs n_i et n_j verront leurs probabilités d'activation mises à jour de façon identique, restant ainsi égales jusque à la fin de l'observation*. Les probabilités d'activation, à une étape de temps $t > 1$ donnée, sont donc les suivantes :

– Pour les nœuds appartenant à l'ensemble des instigateurs, $P_c(n_i, q_k, \leq t) = 1$

– Pour les nœuds atteignables,

$$\begin{aligned} P_c(n_i, q_k, \leq t) &= P_c(n_i, q_k, \leq t-1) + (1 - P_c(n_i, q_k, \leq t-1)) * P_c(n_i, q_k, t-1) \\ &= 1 - (\tfrac{1}{2})^t \end{aligned}$$

Puisque qu'elle est défini selon une suite arithmetico-géométrique.

– Pour les nœuds inatteignables, $P_c(n_i, q_k, \leq t) = 0$

Connaître ces probabilités d'activation permet de définir une quantité minimale $\mathcal{Q}$ de probabilité d'activation assurant que le problème dMI réponde positivement si et seulement si, chaque nœud représentant un élément de l'univers $\mathcal{U}$ est actif. Cette situation, pour le "graphe social" fourni par l'application $(SC(\mathcal{C},\mathcal{U}))$, correspond à l'existence, dans $\mathcal{C}$, d'une famille couvrant l'univers $\mathcal{U}$. En conséquence nous pouvons répondre au problème de couverture en utilisant dMI. Pour ce faire nous utilisons l'application Γ ci-dessous fournissant une instance du problème dMI pour chaque instance du problème de couverture :

$$\Gamma(\mathcal{U}, \{set_0, \ldots, set_n\}, \kappa) = \langle (V, E), \kappa, \mathcal{Q} = \kappa * 1 + |\mathcal{U}| * 1 - (\frac{1}{2})^t \rangle$$

5.3. Approximation : Algorithme d'escalade de colline

Set A = $\emptyset$
for i = 1 to k **do**
  for all $v \in V \setminus A$ **do**
    compute $\sigma(A \cup \{v\}, q, t)$
    if $\sigma(A \cup \{v\}, q, t)$ is maximal **then**
      $v_{max} = v$
    end if
  end for
  $A \leftarrow A \cup \{v_{max}\}$
end for

Figure 3. L'algorithme d'escalade de colline glouton

Le développement ci-dessus montre que la recherche de l'ensemble optimal d'instigateurs est NP-difficile. Il est donc nécessaire, pour les réseaux relativement importants étudiés en pratique, de proposer un algorithme permettant d'identifier non pas les k meilleurs instigateurs, mais k "bons" instigateurs. C'est ce que nous faisons dans la suite.

L'algorithme d'escalade de colline glouton (greedy hill climbing) appliqué à notre problème est un moyen de trouver une approximation de l'ensemble E de k utilisateurs qui maximise une fonction σ, qui correspond au nombre d'utilisateurs atteints après une diffusion. La figure 3 décrit l'algorithme adapté pour le problème de maximisation de l'influence. Il a été prouvé ((Nemhauser *et al.*, 1978)) que cet algorithme est une approximation de $(1 - 1/e)$ de la solution optimale si σ est une fonction positive, monotone et sous-modulaire.

Nous voulons prouver que la fonction $\sigma(A, q, t)$ est sous-modulaire pour le modèle RUC. Prenons S et T deux ensembles d'utilisateurs tels que $S \subseteq T$. Nous voulons montrer que :

$$\forall v \notin T : \sigma(S \cup \{v\}) - \sigma(S) \geq \sigma(T \cup \{v\}) - \sigma(T) \tag{10}$$

Soit $R(A)$ l'ensemble des utilisateurs atteignables depuis les utilisateurs de l'ensemble A. Le processus de diffusion étant toujours le même, l'équation 10 est équivalente à l'équation suivante :

$$\forall v \notin T : R(\{v\}) \setminus R(T) \subseteq R(\{v\}) \setminus R(S) \tag{11}$$

Il suffit donc de montrer que l'équation 11 est vraie, ce qui se déduit directement du fait que l'ensemble S est inclus dans l'ensemble T (en effet, un utilisateur qui est atteint depuis l'ensemble S l'est aussi par l'ensemble T, par contre l'ensemble T peut atteindre de nouveaux utilisateurs que l'ensemble S n'atteint pas).

5.4. Illustration

Dans le but d'obtenir plus d'informations sur la qualité de la méthode gloutonne, et pas seulement le seuil d'approximation, nous comparons les résultats de cette méthode avec quelques heuristiques simples :

– Plus grand degré sortant : le premier utilisateur choisi est celui qui a le plus grand degré sortant (c'est-à-dire le plus grand nombre de voisins sortants), les autres utilisateurs sont choisis de la même manière jusqu'à en obtenir k.

– Centralité de distance : on choisit l'utilisateur qui est le plus central. La centralité est la distance (nombre de liens séparant les deux utilisateurs) moyenne d'un utilisateur u à tous les autres utilisateurs du réseau. Pour les utilisateurs ne pouvant pas être atteints, la distance est arbitrairement fixée au nombre d'utilisateurs dans le graphe. Après le choix du premier utilisateur, les autres sont choisis de la même manière jusqu'à en obtenir k.

– Aléatoire 100 : on choisit aléatoirement cent fois un ensemble de k utilisateurs dans le réseau et on calcule le résultat moyen sur ces cent ensembles.

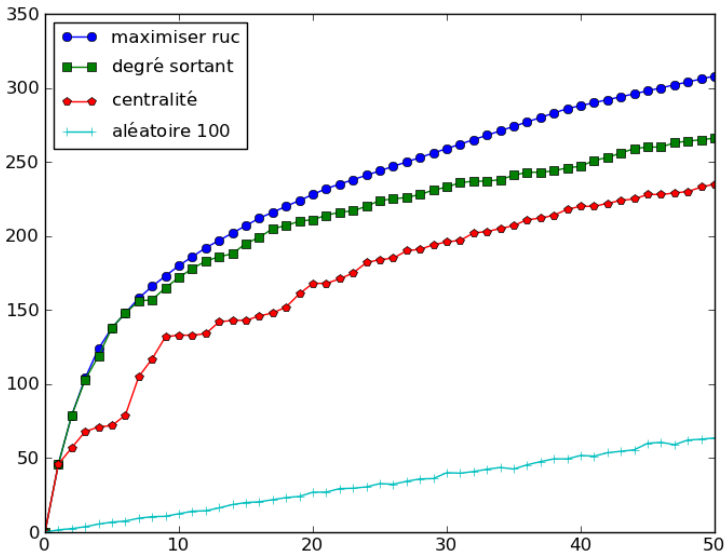

Figure 4. Illustration de la maximisation de l'influence : nombre d'utilisateurs atteints à la fin de la diffusion en fonction du nombre de diffuseurs initiaux

Nous ne pouvons pas comparer ces résultats à l'ensemble optimal de k utilisateurs car ce problème est NP-difficile et une recherche exhaustive ne peut être menée sur nos données. En revanche, nous allons comparer l'algorithme glouton précédent avec les heuristiques ci-dessus.

La figure 4 montre les résultats obtenus par les différentes heuristiques pour le choix des diffuseurs initiaux de la diffusion. Comme nous pouvions nous y attendre, la méthode aléatoire donne des résultats nettement en dessous des autres méthodes. Nous pouvons constater une amélioration entre la méthode gloutonne et les deux autres heuristiques au dessus d'un certain seuil pour la taille de l'ensemble des diffuseurs initaux. Les choix des utilisateurs des trois méthodes sont très similaires au départ (elles choisissent les "hubs" en premier lieu) mais diffèrent par la suite. Ainsi, une heuristique, même simple, fondée sur la méthode de diffusion donne des résultats significativement meilleurs que les heuristiques basées sur la structure du graphe.

6. Conclusion

Nous avons présenté dans cette article trois nouveaux modèles de diffusion de l'information fondés sur des facteurs propres à chaque utilisateur. Une comparaison avec des modèles standards à cascades a montré que ces modèles apportent une réelle amélioration lors de la prédiction de la diffusion d'un contenu. Nous avons aussi montré

que le problème de la maximisation de l'influence en utilisant les modèles de diffusion que nous avons définis est NP-difficile, et avons proposé une adaptation de l'algorithme d'escalade de colline pour pouvoir approcher la solution optimale avec une précision de $(1 - 1/e)$. Les modèles que nous avons proposés dans cet article sont en partie fondés sur le fait qu'un utilisateur est influencé par ses voisins. Ces modèles sont propres au mode de diffusion observé dans les blogs, où un bloggueur décide de re-diffuser l'information donnée par un autre bloggueur. Il serait intéressant d'étudier le modèle dual dans lequel, tout en restant centré sur l'utilisateur, on estime la probabilté qu'un utilisateur puisse activer l'un de ses voisins. Les liens dans un réseau social étant directement liés à la diffusion des contenus, une autre piste de travail intéressante serait d'adapter nos modèles pour estimer des nouveaux liens entre les utilisateurs.

Références

Abrahamson E., Rosenkopf L. (1997). Social network effects on the extent of innovation diffusion : A computer simulation. *Organization Science*, Vol. 8, N° 3, pp. 289-309.

Borodin A., Filmus Y., Oren J. (2010). Threshold models for competitive influence in social networks. In *Wine*, p. 539-550. Springer.

Brauer F., Castillo-Chavez C. (2001). *Mathematical Models in Population Biology and Epidemiology*. Springer.

Burton K., Java A., Soboroff I. (2009). The ICWSM 2009 Spinn3r Dataset. In *The third annual conference on weblogs and social media (icwsm 2009).*

Dodds P., Watts D. (2004). Universal Behavior in a Generalized Model of Contagion. *Physical Review Letters*, Vol. 92, N° 21.

Domingos P., Richardson M. (2001). Mining the network value of customers. In *Proceedings of the seventh acm sigkdd international conference on knowledge discovery and data mining*, p. 57-66. ACM.

Goldenberg J., Libai B., Muller E. (2001). Talk of the Network : A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters*, pp. 211-223.

Gomez-Rodriguez M., Balduzzi D., Schölkopf B. (2011). Uncovering the temporal dynamics of diffusion networks. In L. Getoor, T. Scheffer (Eds.), *Proceedings of the 28th international conference on machine learning (icml-11).*

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

Granovetter M. (1978). Threshold Models of Collective Behavior. *American Journal of Sociology*, Vol. 83, N° 6, pp. 1420-1443.

Granovetter M., Soong R. (1988). Threshold models of diversity : Chinese restaurants, residential segregation, and the spiral of silence. *Sociological Methodology*, Vol. 18, pp. 69-104.

Karp R. (1972). Complexity of Computer Computations, chapter Reducibility among combinatorial problems. *Plenum Press, New York*, pp. 85-103.

Kempe D., Kleinberg J., Tardos E. (2003). Maximizing the spread of influence through a social network. In *Kdd '03 : Proceedings of the ninth acm sigkdd international conference on knowledge discovery and data mining*, p. 137-146. ACM Press.

Kimura M., Saito K., Nakano R. (2007). Extracting influential nodes for information diffusion on a social network. *Proceedings Of The National Conference On Artificial Intelligence*, Vol. 22, N° 2, pp. 1371.

Leskovec J., Krause A., Guestrin C., Faloutsos C., VanBriesen J., Glance N. (2007). Cost-effective outbreak detection in networks. In *Proceedings of the 13th acm sigkdd international conference on knowledge discovery and data mining*, p. 420-429. ACM.

López-Pintado D. (2008). Diffusion in complex social networks. *Games and Economic Behavior*, Vol. 62, N° 2, pp. 573-590.

López-Pintado D., Watts D. J. (2008). Social Influence, Binary Decisions and Collective Dynamics. *Rationality and Society*, Vol. 20, N° 4, pp. 399-443.

Macy M. W. (1991). Chains of Cooperation : Threshold Effects in Collective Action. *American Sociological Review*, Vol. 56, N° 6, pp. 730-747.

Manning C. D., Raghavan P., Schütze H. (2008). *An Introduction to Information Retrieval*. Press, Cambridge U.

Munshi K. (2004). Social learning in a heterogeneous population : technology diffusion in the indian green revolution. *Journal of Development Economics*, Vol. 73, N° 1, pp. 185-213.

Nemhauser G. L., Wolsey L. A., Fisher M. L. (1978). An analysis of approximations for maximizing submodular set functions-I. *Mathematical Programming*, Vol. 14, N° 1, pp. 265-294.

Newman M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review*, Vol. 45, N° 2, pp. 167-256.

Prakash B. A., Tong H., Valler N., Faloutsos M., Faloutsos C. (2010). Virus propagation on time-varying networks : Theory and immunization algorithms. In *Principles of data mining and knowledge discovery*, p. 99-114.

Richardson M., Domingos P. (2002). Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth acm sigkdd international conference on knowledge discovery and data mining*, p. 61-70. ACM.

Saito K., Kimura M., Ohara K., Motoda H. (2009). Learning continuous-time information diffusion model for social behavioral data analysis. *Learning*, Vol. 5828, pp. 322–337.

Saito K., Nakano R., Kimura M. (2008). Prediction of information diffusion probabilities for independent cascade model. In *Proceedings of the 12th international conference on knowledge-based intelligent information and engineering systems, part iii*, p. 67-75. Springer-Verlag.

Saito K., Ohara K., Yamagishi Y., Kimura M., Motoda H. (2011). Learning diffusion probability based on node attributes in social networks. In M. Kryszkiewicz, H. Rybinski, A. Skowron, Z. W. Ras (Eds.), *Ismis*, Vol. 6804, p. 153-162. Springer.

Schelling T. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, Vol. 1.

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

Valente T. (1996). Social network thresholds in the diffusion of innovations. *Social Networks*, Vol. 18, N° 1, pp. 69-89.

Valente T. W. (1995). *Network Models of the Diffusion of Innovations (Quantitative Methods in Communication Subseries)*. Hampton Press (NJ).

Young H. P. (2009). Innovation diffusion in heterogeneous populations : Contagion, social influence, and social learning. *American Economic Review*, Vol. 99, N° 5, pp. 1899-1924.