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. MARAMI 2011 - Seconde conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique, Oct 2011, Grenoble, France. 12p. hal-00744224

HAL Id: hal-00744224
https://hal.science/hal-00744224v1

Submitted on 22 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

Cedric Lagnier — Eric Gaussier — Francois Kawala

Université Joseph Fourier / Grenoble 1 / CNRS
Laboratoire LIG
Bat. CE4
Allée de la Palestine
38610 GIERES

{cedric.lagnier, eric.gaussier, francois.kawala}@imag.fr

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 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 les modèles standard de diffusion de l'information.*

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 ilustrate their behavior through a comparison with standard information diffusion models on a real dataset.*

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

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

1. Introduction

L'utilisation importante des réseaux sociaux dans notre société est un fait avéré. La communication entre personnes existait bien avant Internet et la création d'outils a permis une plus grande facilité à communiquer et surtout un champ d'action plus large. On peut échanger assez facilement avec une autre personne dans le monde malgré la distance ou les frontières. Nous nous intéresserons ici à la manière dont se diffuse l'information au sein de ces réseaux et plus particulièrement à la façon dont les utilisateurs agissent pour diffuser cette information. Ce domaine de recherche étant très proche de la diffusion d'innovations dans le marketing et de la diffusion de virus au sein d'une population, nous parlerons des différents modèles existant puis présenterons notre approche centrée sur l'utilisateur.

L'étude des réseaux sociaux commence par l'étude de leurs caractéristiques, comme souligné dans [MER 04]. Le nombre d'utilisateurs, la densité moyenne du réseau, la centralité d'un utilisateur, etc constituent un premier jeu d'indicateurs permettant de comprendre les interactions entre personnes ([MAZ 06]). On peut essayer de catégoriser les comportements de chacun afin de mieux comprendre les actions des utilisateurs ([GOL 04]). Une autre caractéristique capitale, qui joue un rôle primordial dans la diffusion de l'information, est bien sûr la topologie du réseau ([LIE 05]). Les utilisateurs étant liés entre eux, la densité plus ou moins grande de ces liens rend certains utilisateurs plus influents que d'autres. Comme dit précédement, les modèles de diffusion ne sont pas réservés à la diffusion de l'information et on retrouvera notemment la famille de modèles épidémiologiqes SI (*Suspected-Infected*) ([TRO 01]).

La diffusion d'information dans les réseaux a été étudiée au cours de plusieurs travaux. [CHA 09] montre que la diffusion ne réussit pas pour tous les contenus. De plus, les utilisateurs étant plus ou moins influents, la diffusion va dépendre des diffuseurs initiaux. L'étude présentée dans [KEM 03] démontre le caractère NP-complet du problème de maximisation de l'influence, qui consiste à trouver les K meilleurs diffuseurs initiaux, K étant un nombre fixé. Des méthodes approchées pour choisir au mieux où commencer une diffusion ont donc vu le jour, comme celle présentée dans [KIM 07]. On peut représenter une diffusion d'information comme un sous-graphe du graphe original en ne gardant que les utilisateurs et les liens touchés par l'information lors de la diffusion. De tels sous-graphes sont appelés cascades. Les études menées dans [LES 07a, LES 07b] analysent ces cascades dans un réseau de diffusion d'innovation ainsi que dans dans un réseau de blogs.

L'étude de la diffusion dans les réseaux se fait principalement par le développement de modèles qui essaient de représenter au mieux le processus de diffusion. Les premiers modèles ont été introduits par [GRA 78] et sont des modèles à seuil dans lesquels un utilisateur s'active si le nombre de ses voisins actifs dépasse un seuil fixé. Le travail présenté dans [ABR 97] étudie les modèles à seuil dans le cadre de la diffusion d'innovations. Le modèle à seuil le plus répandu est le modèle LT (*Linear Threshold*). Un autre type de modèle très répandu est le modèle IC (*Independent Cascades*) : lorsqu'un utilisateur *u* est activé, il tente d'activer chacun de ses voisins *v* avec une

probabilité $k_{u,v}$. Une procédure d'estimation des paramètres du modèle IC, fondée sur la maximisation de la vraisemblance, est présentée dans [SAI 10].

Plusieurs améliorations ont été proposées sur ces modèles de base afin d'en améliorer le fonctionnement et d'essayer de mieux coller à la réalité. Un dérivé du modèle IC qui tient compte du temps que met un utilisateur à activer ses voisins est le modèle ASIC ([SAI 09]). Dans ce cadre, l'utilisateur activé ne tente pas d'activer tout de suite, mais peut le faire avec un temps de latence. Ce phénomène de latence dans la diffusion se retrouve aussi dans le modèle introduit dans [LIB 08]. Dans ce dernier modèle, un paramètre de filtre dans la diffusion est aussi introduit, reposant sur le filtre anti-spam des chaînes de courriers électroniques. Une autre idée intéressante est celle d'homophilie, qui rend compte du fait que des utilisateurs ayant des caractéristiques proches auront plus tendance à communiquer que des utilisateurs ayant des caractéristiques éloignées. On retrouve dans [APO 09] une étude sur les actions efféctuées par les utilisateurs de Flickr. [SAI 11] présente un modèle dérivé de ASIC dans lequel les probabilités dépendent de la similarité entre les utilisateurs. Une autre amélioration, qui conduit à des simulations plus réalistes, consiste à ne pas exprimer la diffusion comme un simple phénomène binaire dont on peut observer les valeurs à un moment donné, mais plutôt comme un phénomène évoluant dans le temps et caractérisé par une probabilité de diffusion en chaque nœud (ou ensemble de nœud). On obtient ainsi dans [YOU 09] et [LOP 08] des modèles fondés sur des équations aux différences. Enfin, dans [YAN 10], la diffusion est vue à plus grande échelle, en ne considérant plus les utlisateurs de manière individuelle mais en tentant de modéliser l'influence de certains sur le reste du réseau.

Dans la section suivante nous présentons un ensemble de nouveaux modèles qui tiennent compte des divers éléments précédents et ajoutent un certain nombre de caractéristiques propres à l'utilisateur. Nous présentons ensuite, section 3, une validation expérimentale de ces modèles sur un jeu de données issus de blogs. La section 5 conclut notre étude et présente les perspectives que nous souhaitons développer par la suite.

2. Modèles centrés utilisateur

Nous considérons dans la suite des graphes orientés $G = (V, L)$ avec V l'ensemble des nœuds (ici des utlisateurs du réseau social) et L l'ensemble des liens entre utilisateurs. Prenons *(u,v)* un lien de *u* vers *v* dans notre réseau : nous dirons que *v* est un voisin sortant de *u* et que *u* est un voisin entrant de *v*. Nous appellerons *B(u)* l'ensemble des voisins entrant de *u* et *F(u)* l'ensemble des voisins sortant de *u*. De la même façon que pour le modèle IC, un utilisateur peut se trouver dans deux états : actif ou inactif (on considèrera qu'un utilisateur est inactif tant qu'il n'a pas diffusé l'information, et actif dès lors qu'il a diffusé l'information). Tout comme dans les travaux récents sur les modèles de diffusion, nous nous inscrivons ici dans un cadre probabiliste, et parlerons donc, pour chaque utilisateur, de « probabilité d'être actif » (la probabilité d'être inactif s'en déduisant directement).

Nous allons dans cette partie prsésenter trois modèles centrés sur l'utilisateur, fondés tous trois sur une fonction de transition *T* correspondant à la probabilité d'un utilisateur de devenir actif, c'est-à-dire de diffuser une information qu'il a reçue. Cette fonction est caractérisée par trois paramètres : i) la topologie du réseau (l'état des utilisateurs voisins de notre utilisateur), ii) la similarité entre le contenu diffusé et le profil utilisateur et iii) la propension d'un utilisateur à diffuser de l'information en général (on fera une différence entre le diffuseur « fou », proche du spammeur, et une personne qui ne diffuse presque rien).

Nous avons choisi pour combiner ces trois paramètres d'utiliser une fonction de seuil sigmoïde :

$$T(u,e,t) = \begin{cases} (1 + exp(-\lambda_1(s(u,e,\theta_s)) - \lambda_2 a(u,e,t) - \lambda_3(w(u,\theta_w)))^{-1} \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{si } a(u,e,t) > 0 \\ 0 \text{ sinon} \end{cases}$$

où :

– *u* est un utilisateur, *e* l'information diffusée[1], *t* le temps auquel on se trouve ;

– $s(u,e,\theta_s)$ correspond à la différence entre, d'une part, la similarité ($cos(u,e)$) entre le profil de l'utilisateur *u* et l'information *e*, et, d'autre part, un seuil θ_s : $s(u,e,\theta_s) = cos(u,e) - \theta_s$. Dans cette étude, profil utilisateur et information sont représentés sous la forme de vecteur de termes (cf. section 3) et le cosinus est utilisé pour calculer la similarité entre ces vecteurs. θ_s a de plus été fixé arbitrairement à 0.5, ce qui signifie qu'une similarité entre profil utilisateur et contenu diffusé inférieure à 0.5 pénalise la probabilité de diffusion de l'utilisateur (elle l'augmente dans le cas contraire) ;

– $a(u,d,t)$ est le nombre de voisins entrant de l'utilisateur *u* qui sont actifs au temps *t* ;

– $w(u,\theta_w)$ est la différence entre la propension de *u* à diffuser de l'information et un seuil θ_w. Nous définissons ici la propension d'un utilisateur u comme le rapport du nombre de messages reçus et rediffusés par u, sur le nombre total de messages reçus par u. En d'autres termes, plus un utilisateur a, dans le passé, rediffusé des messages, plus sa propension à diffuser est élevée. La propension est un nombre qui varie entre 0 et 1 ; plus ce nombre est élevé, plus grande est la probabilité d'un utilisateur de devenir actif et, inversement, plus ce nombre est faible, plus cette probabilité doit être pénalisée. Tout comme pour la similarité, nous avons fixé ici le seuil θ_w à 0.5 pour rendre compte de cet effet ;

– Enfin, λ_1, λ_2 et λ_3 sont 3 paramètres positifs qui sont appris automatiquement pour chaque réseau.

Il est à noter que dans la fonction précédente, les paramètres λ sont communs à tous les utilisateurs. Il est possible bien sûr de considérer des paramètres différents pour

1. On peut considérer que chaque information diffusée correspond à un « épisode » de diffusion, d'où la notation utilisée, que l'on retrouve dans plusieurs travaux.

chaque utilisateur ou groupes d'utilisateurs. Toutefois, cela peut entraîner des problèmes au niveau de l'apprentissage de ces paramètres (surapprentissage). Nous nous contentons ici d'une version plus simple qui, comme nous le verrons plus loin, se comporte bien en pratique. Enfin, il est également possible d'apprendre les deux seuils θ_s et θ_w.

2.1. *Modèle UC*

Le premier des trois modèles que nous allons présenter est le modèle UC (*User Centric model*) car il est fondementalement très proche du modèle IC. De la même manière qu'avec le modèle IC, on considère ici que lorsqu'un utilisateur u s'active au temps t, tous ses voisins sortants v ont une probabilité de s'activer entre le temps t et le temps $t+1$ égale à $T(v, e, t)$. Si plusieurs des voisins entrant de v deviennent actifs au temps t, c'est autant de chances pour v de s'activer, c'est-à-dire que v essaiera de s'activer autant de fois qu'il a de voisins entrant devenus actifs au temps t. La différence fondamentale avec IC réside dans le fait que ce n'est pas l'utilisateur u qui essaie d'activer ses voisins lorsqu'il est devenu actif, mais ses voisins qui ont une probabilité de devenir actifs lorsqu'il y a du changement dans leur entourage. Le système evolue ensuite de la même façon que IC et s'arrête lorsqu'il n'y a plus de changement.

2.2. *Modèle RUC*

Dans le modèle RUC (*Reinforced User Centric model*) un utilisateur u n'est pas actif ou inactif mais a une probabilité $\rho(u, e, t)$ d'être actif au temps *t* (cette probabilité dépend bien sûr du temps t, de l'information diffusée e et de l'utisateur u). De plus, nous permettons à chaque utilisateur de s'activer à tout moment (suivant bien sûr la fonction de transition T) et pas seulement lorsqu'un de ses voisins entrant devient actif. Ce modèle est alors caractérisé, pour chaque utilisateur, par deux états et une probabilité de transition donnée par T (à noter que cette probabilité de transition évolue au cours du temps). Nous avons :

$$\rho(u, e, t+1) = \rho(u, e, t) + (1 - \rho(u, e, t)) \times T(u, e, t)$$

En d'autres termes, un utilisateur actif à l'instant $t+1$ soit était actif à l'instant t, soit ne l'était pas et l'est devenu entre t et $t+1$. On peut alors montrer par récurrence que, pour tout utilisateur qui n'est pas à l'origine de la diffusion, sa probabilité d'être actif à une étape donné vaut :

$$\rho(u, e, t) = \sum_{t'=0}^{t-1} T(u, e, t') \prod_{\tau=0}^{t'-1} (1 - T(u, e, \tau))$$

Rappelons ici que la fonction T dépend du nombre de voisins actifs entrant et que, lorsque celui-ci est nul, elle vaut 0. Un utilisateur n'ayant aucun voisin entrant actif aura donc une probabilité nulle de devenir actif. Dans ce modèle chaque utilisateur a une probabilité d'être actif, il n'est donc pas possible de calculer le nombre de voisins

actifs. On peut en revanche calculer l'espérance de ce nombre, que nous utiliserons à la place de $a(u, e, t)$ dans la fonction T :

$$\begin{aligned} E[a(u,e,t)] &= \sum_{n=0}^{B(u)} n \; P(a(u,e,t) = n) \\ &= \sum_{v \in B(u)} \rho(u,e,t) \end{aligned}$$

où $P(a(u, e, t) = n)$ est la probabilité que $a(u, e, t)$ soit égal à n[2].

Le principal problème de ce modèle réside dans le fait que si $T(u, e, t)$ est positif à un instant donné (ce qui se produit dès lors que l'un des voisins entrant de u a une probabilité d'etre actif non nulle) la probabilité $\rho(u, e, t)$ tend vers 1 (ni ρ ni T ne peuvent diminuer au cours du temps). On voit donc que ce modèle est déficient sur le (très) long terme.

2.3. *Modèle FRUC*

Nous avons pensé le troisième modèle que nous présentons pour pallier au principal problème du modèle précédent. Le modèle FRUC (*Forgetful Reinforced User Centric model*) est très similaire au modèle RUC, si ce n'est que nous y avons ajouté un paramètre d'oubli α de manière à ce qu'un utilisateur tienne peu compte d'une information qui a été diffusée il y a longtemps. En d'autres termes, plus une information est fraîche, plus elle a de chances d'influencer un utilisateur, et plus elle est ancienne, plus elle a de chances d'avoir été oubliée. Nous introduisons dans ce modèle une nouvelle quantité $I(u, e, t)$ qui est l'influence qu'a l'utilisateur u sur ses voisins sortant au temps t pour l'information e. Sa valeur au cours du temps est directement liée à la probablité qu'un utilisateur a de diffuser l'information, mais diminue à travers le paramètre α :

$$I(u, e, t + 1) = \alpha \times I(u, e, t) + (1 - \rho(u, e, t)) \times T(u, e, t)$$

avec $0 \leq \alpha \leq 1$. L'influence d'un utilisateur à l'instant $t + 1$ dépend donc de son influence à l'instant t, qui se trouve diminuée à chaque étape de temps, et de sa probabilité de devenir actif entre les instants t et $t + 1$ ($(1 - \rho(u, e, t)) \times T(u, d, t)$). De la même façon que pour le modèle RUC, on peut montrer par récurrence que l'influence à un instant donné, pour un utilisateur qui n'est pas à l'origine de la diffusion, prend la forme :

$$I(u, e, t) = \sum_{t'=0}^{t-1} \alpha^{t-t'-1} T(u, e, t') \prod_{\tau=0}^{t'-1} (1 - T(u, e, \tau))$$

Les équations pour $\rho(u, e, t)$ restent inchangées par rapport au modèle RUC.

2. Nous ne donnons pas ici la démonstration de la deuxième égalité, qui est purement technique

L'influence sert dans le modèle FRUC au calcul du nombre de voisins actifs. On ne considère en effet plus la probabilité d'un voisin d'être actif, mais directement son influence, ce qui permet de tenir compte du facteur d'oubli dans la fonction de transition de chaque utilisateur :

$$E[a(u,e,t)] = \sum_{v \in B(u)} I(u,e,t)$$

On peut remarquer que lorsque α vaut 1, l'influence se confond avec la probabilité d'être actif, et le modèle FRUC est équivalent au modèle RUC. Dans la suite de l'article, nous avons arbitrairement fixé α à 0.9.

2.4. *Estimation des paramètres*

De façon à pouvoir utiliser nos modèles sur des données réelles afin de pouvoir tester leur efficacité, nous avons besoin d'estimer les valeurs des paramètres λ_1, λ_2 et λ_3. Nous nous reposons pour cela sur un principe du maximum de vraisemblance, c'est-à-dire que nous cherchons, pour chaque modèle, la valeur des paramètres qui maximise la vraisemblance du modèle sur des données réelles. Nous définissons E comme le cardinal de l'ensemble des episodes de diffusion (les épisodes sont donc indicés de 1 à E), $D(t,e)$ comme étant l'ensemble des utilisateurs qui diffusent réellement l'information e au temps t et $C(t,e)$ l'ensemble des utilisateurs l'ayant déjà diffusé au temps t : $C(t,e) = \cup_{\tau=0}^{t} D(\tau, e)$

Pour le modèle UC, la fonction de vraisemblance prend la forme suivante :

$$L(\lambda_1, \lambda_2, \lambda_3) =$$
$$\prod_{e=1}^{E} \prod_{t=1}^{t_{max}} \left(\prod_{u \in D(t,e)} T(u,e,t) \prod_{u \in D(t-1,e)} \prod_{v \in F(u) \setminus C(t,e)} (1 - T(v,e,t)) \right)$$

Autrement dit, un utilisateur qui est réellement actif au temps t doit avoir une probabilité de s'activer avec notre modèle au temps $t-1$ très forte. De la même manière, un utilisateur qui a la possibilité de s'activer mais ne le fait pas dans les données réelles doit avoir une probabilité très faible de s'activer avec notre modèle.

Pour les modèles RUC et FRUC, la vraisemblance dépend directement des états des utlisateurs et s'écrit :

$$L(\lambda_1, \lambda_2, \lambda_3) = \prod_{e=1}^{E} \prod_{t=1}^{t_{max}} \left(\prod_{u \in C(t,e)} \rho(u,e,t) \prod_{u \notin C(t,e)} (1 - \rho(u,e,t)) \right)$$

Il est relativement complexe de calculer le gradient de ces vraisemblances ou de déployer un algorithme EM (à cause du terme $E[a(u,e,t)]$ de la fonction T). En revanche, dans la mesure où les valeurs des paramètres sont limitées (ils ne peuvent être négatifs par définition ni trop grands pour des raisons de calcul), il est possible d'effectuer une recherche par grille, qui présente l'avantage de ne pas rester « coincée »

dans un minimum local. Nous procédons pour cela en deux étapes : nous recherchons tout d'abord les valeurs « grossières » les meilleures (parmi l'ensemble 0, 1, 5, 10, 25 pour chaque paramètre), valeurs que nous notons λ^g, puis affinons ces valeurs par une deuxième recherche par grille en considérant pour chaque paramètre λ l'intervalle de valeurs $[\max(0, \lambda^g - 5); \lambda^g + 5]$. Par exemple, pour une valeur optimale des paramètres de $(\lambda_1 = 10, \lambda_2 = 1, \lambda_3 = 5)$ pour la première étape, nous cherchons, pour la seconde étape, des valeurs des paramètres entre 5 et 15 pour λ_1, entre 0 et 6 pour λ_2 et entre 0 et 10 pour λ_3.

3. Expérimentation

3.1. *Données*

Nous sommes partis du jeu de données brut qui a été utilisé lors du challenge de ICWSM 2009[3]. Il s'agit d'un ensemble de billets de blogs qui ont été récupérés par spinn3r[4]. Nous avons effectué un certain nombre de filtres sur ces données. Tout d'abord nous n'avons gardé que les billets en anglais appartenant à la plus grosse composante connexe (tout nœud est lié à au moins un autre nœud de la communauté, quel que soit le sens de ce lien) et qui ont été postés lors du mois d'août. L'étape suivante a été la récupération des liens et des contenus de chaque billet. Nous avons effectué une racinisation (stemmer de Porter), filtré les mots vides à l'aide d'une liste puis conservé uniquement les mots alpha-numériques qui apparaissaient plus de 5 fois dans le jeu de données. Les documents vides ainsi que les liens vers le futur (dans le jeu brut, soit un certain nombre de billets n'étaient pas étiquetés par la date réelle, soit ils avaient été édités pour ajouter un lien mais la date était toujours celle d'origine) ont été supprimés du jeu de données.
En notant N_u le nombre d'utilisateurs, N_b le nombre de billets et N_t le nombre de termes, notre jeu de donnés se présente sous la forme d'une série de matrices : (i) une matrice M de taille $N_u \times N_b$ avec $M_{ij} = 1$ si le billet j a été posté/diffusé par l'utilisateur i, (ii) une matrice L de taille $N_b \times N_b$ avec $L_{ij} = 1$ si le billet i contient un lien vers le billet j, (iii) et une matrice O de taille $N_b \times N_t$ où O_{ij} représente le nombre d'occurrences du terme j dans le billet i. A chaque billet est de plus associée une date de publication.

L'étape suivante a consisté à calculer les liens entre utilisateurs ainsi que les profils de chaque utilisateur. Les liens sont représentés par une matrice G de taille $N_u \times N_u$ avec $M_{i,j} = n$ si l'utilisateur i a mis n liens vers l'utilisateur j dans ses billets. Le profil P_i d'un utilisateur i est défini comme la moyenne entre tous les vecteurs termes des billets diffusés par cet utilisateur :

$$P_i = \frac{(M_{i.}O)^T}{\sum_{j=1}^{N_b} M_{ij}}$$

Nous avons ensuite calculé les cascades (les groupes de billets tous liés les uns aux autres) qui correspondent aux différentes diffusion d'information ou épisodes, puis

3. http ://icwsm.org/data/
4. http ://spinn3r.com/

	1re itération			2e itération		
Méthode	λ_1	λ_2	λ_3	λ_1	λ_2	λ_3
UC	10	5	10	13	5	7
RUC	10	1	5	10	1	6
FRUC	5	0	5	10	1	6

Tableau 1. *Valeurs des paramètres des modèles (FR)UC après estimation*

avons sélectionné au hasard 100000 cascades. Nous obtenons ainsi 40000 utilisateurs et environ 103000 billets dont la plupart appartiennent à des cascades de taille 1. La taille moyenne d'une cascade est de $1,02$ (ou $2,7$ si on ne prend pas en compte les cascades de taille 1). Il y a une remarque qu'il reste à ajouter : les modèles IC et UC ne fonctionnent pas sur une échelle de temps standard mais sur une chronologie d'évènements de diffusion. Il a donc fallu tranformer les temps de chacun des messages pour qu'ils correspondent à cette chronologie. Nous ne pourrons du coup pas comparer les résultats, entre ces deux modèles et les autres, au cours du processus de diffusion, mais seulement à la fin de la diffusion.

3.2. *Résultats*

Nous avons découpé notre jeu de données en deux étapes de temps distinctes : les deux premiers tiers correspondent à la période d'apprentissage, et le dernier tiers correspond à la période de test. On peut imaginer que les modèles sont entraînés sur le début et qu'ils sont ensuite utilisés pour prédire la suite des évènements. Au début du jeu de test, nous avons considéré que tous les messages diffusés pendant la période d'apprentissage étaient des points de départ des diffusiosn (de la même façon que les réels points de départ, *i.e.* les messages qui n'ont fait aucune citation), ce qui correspond à dire que les modèles ont tous prédit parfaitement la diffusion avant le début du jeu de test. Le tableau 1 montre les valeurs des paramètres obtenus lors des deux itérations de l'estimation pour nos modèles. On remarque très clairement que le paramètre qui a le plus de poids après ces estimations est λ_1, qui définit l'importance de la similarité avec le contenu. On remarque aussi que, pour les modèles avec renforcement temporel (RUC et FRUC), le paramètre agissant sur l'importance du nombre de voisin (λ_2) est très faible, ce qui s'explique par le fait que, dans ces modèles, les utilisateurs peuvent diffuser sur un plus large espace temporel. La similarité de contenu et la propension à diffuser discriminent mieux les diffuseurs des non-diffuseurs dans ce cas. Nous avons fait quelques tests de robustesse sur ces estimations en changeant légèrement les valeurs des paramètres (de 0.1) et les résultats obtenus avec les paramètres modifiés sont très proches de ceux obtenus avec les paramètres estimés. Pour l'estimation des paramètres des modèles IC et ASIC, nous avons directement utilisé les algorithmes EM décrits dans [SAI 10] et [SAI 09].

Nous avons ensuite prédit, sur le jeu de test, les diffusions avec les 5 modèles cités précédemment et comparé les probabilités d'être actif obtenues avec chaque modèle avec les valeurs réelles (1 si l'utilisateur est actif, 0 sinon). Dans un premier temps, nous ne considérons que les résultats après intégralité de la diffusion, de manière à

Méthode	Positifs	Négatifs	Ensemble	Distance euclidienne
IC	0.5975	0.0433	0.2178	14.7803
UC	0.4183	0.0998	0.2001	14.5069
ASIC	0.6843	0.0406	0.2441	15.3948
RUC	**0.2203**	0.0285	0.0892	8.5907
FRUC	0.2217	**0.0277**	**0.0891**	**8.5866**

Tableau 2. *Erreur moyenne obtenue avec les différents modèles à la fin de la diffusion.*

pouvoir comparer les modèles temporels (RUC, FRUC) avec les modèles fondés sur une simple chronologie (IC, ASIC, UC). Le tableau 2 montre les résultats obtenus par les cinq modèles sur le jeu de données de blogs. Nous avons fait deux mesures différentes. Les colonnes *positifs*, *négatifs* et *ensemble* correspondent à l'écart moyen, sur tous les utilisateurs et les épisodes, entre la probabilité d'être actif ou inactif fournie par le modèle et la valeur réelle (1 ou 0). Les utilisateurs positifs sont les utilisateurs qui sont effectivement actifs dans les données réelles, les utilisateurs négatifs sont les utilisateurs non actifs dans les données réelles, et l'ensemble représente l'intégralité de tous les utilisateurs. Plus l'écart moyen est faible, meilleurs sont les résultats. La seconde mesure est une distance euclidienne sur l'ensemble des utilisateurs entre le vecteur des activations réelles des utilisateurs et le vecteur de probabilités prédit par le modèle. D'une manière génerale, les modèles ont tendance à mieux prédire l'absence de diffusion que la diffusion, ceci étant surement dû au fait qu'il y a peu de diffusion dans ce type de réseaux (les blogs). Beaucoup d'utilisateurs postent des billets, mais peu rediffusent, et, dans la mesure où l'on considère les utilisateurs diffuseurs et non-diffuseurs avec un poids équivalent lors du calcul de la vraisemblance, les modèles ont tendance à estimer des paramètres qui favorisent la non-diffusion. On remarque que le modèle UC a un fort taux d'erreur sur les utilisateurs négatifs, alors que les modèles RUC et FRUC ont une très faible erreur avec moins de 3%. En ce qui concerne l'erreur sur la diffusion des utilisateurs positifs, on remarque tout d'abord une nette amélioration entre le modèle IC et le modèle UC, de l'ordre de 20%. Les modèles RUC et FRUC fournissent à leur tour une amélioration non négligeable (encore une fois de l'ordre de 20%). Les résultats du modèle ASIC sont plutot médiocres , mais ceci peut s'expliquer par la durée de la période de temps sur laquelle nous travaillons, qui est relativement courte. Dans ce cas, l'estimation des paramètres r influant sur la latence de la diffusion est assez délicate et l'algorithme EM ne permet pas ici de trouver des valeurs adéquates, ce qui entraîne une forte probabilité de diffusion en dehors de la période de temps retenue. En utilisant les paramètres k estimés sur le modèle ASIC avec le modèle IC standard, on obtient des résultats identiques à ceux obtenus avec les paramètres k estimés pour le modèle IC, ce qui montre que le facteur de latence n'apporte rien ici. On remarque aussi que les résultats du modèle FRUC sont très proches de ceux du modèle RUC et nous pensons que la période de temps assez courte nous empêche de voir l'apport du paramètre d'oubli. En effet, on peut prédire une perte de qualité de la prédiction sur les utilisateurs négatifs en utilisant le modèle RUC sur une période de temps suffisamment longue. Dans un second temps, le ta-

Méthode	Positifs avant	Positifs après	Négatifs	Ensemble	Dist. eucl.
ASIC	0.7398	0.0026	0.0410	0.2357	40.8219
RUC	**0.3750**	0.0061	0.0189	**0.1229**	**29.2404**
FRUC	0.3768	**0.0054**	**0.0180**	**0.1229**	29.2483

Tableau 3. *Erreur moyenne obtenue avec les différents modèles au cours de la diffusion.*

bleau 3 montre l'erreur moyenne et la distance euclidienne pour chacun des modèles sur toutes les étapes de temps du jeu de test. La colonne *positifs avant* correspond à l'erreur avant la diffusion réelle (si la diffusion réelle a lieu au temps *t*, elle correspond à la moyenne des erreurs de diffusion sur les étapes *0* à *t-1*) et *positifs après* correspond à l'erreur après la diffusion réelle. Ces résultats vont dans le même sens que ceux obtenus précédemment à la fin de la diffusion. On remarque néanmoins que l'erreur sur les utilisateurs positifs après est très faible, ce qui signifie que les modèles ont plutôt tendance à diffuser trop tôt que trop tard.

4. Conclusion

Nous avons présenté plusieurs modèles de diffusion basés sur quelques principes de base : un utilisateur a tendance à diffuser une information si plusieurs de ses voisins l'ont déjà diffusé, si l'information lui plait et si il a une tendance à diffuser d'une manière générale. Nous avons exprimé les dynamique de ces modèles au cours du temps et présenté la comparaison avec deux autres modèles dont un très standard. Les apports principalement sur l'utilisation de la similarité entre l'information et le profil de l'utilisateur ainsi que sur le renforcement au cours du temps de la diffusion ont l'air d'apporter un réel plus sur le jeu de données réel de réseau de blogs. Pour la suite du travail nous aimerions dans un premier temps fixer les seuils de manière un peu moins arbitraire pour qu'ils dépendent des distributions des valeurs de similarité et de volonté. Ensuite, nous voudrions aller plus loin dans les tests de nos modèles en les appliquant sur plusieurs jeux de données de type et de contexte différents (forums, mails, réseau social type twitter) ainsi que sur des périodes de temps plus longues et implémenter d'autres méthodes pour avoir une évaluation qui couvre une plus grande partie du domaine. Nos modèles n'étants pas figés sur leurs paramètres, il serait aussi intéressant d'intégrer d'autres paramètres tels qu'un filtre ou une influence extérieure au réseau (par exemple la télévision ou les journaux). A plus long terme nous aimerions aussi adapter notre modèle pour fonctionner sur les domaines frères de la diffusion d'information : la diffusion d'innovations et l'épidémiologie.

5. Bibliographie

[ABR 97] ABRAHAMSON E., ROSENKOPF L., « Social Network Effects on the Extent of Innovation Diffusion : A Computer Simulation », , 1997, p. 289–309.

[APO 09] APOLLONI A., CHANNAKESHAVA K., DURBECK L., KHAN M., KUHLMAN C., LEWIS B., SWARUP S., « A Study of Information Diffusion over a Realistic Social Network Model », *2009 International Conference on Computational Science and Engineering*, , 2009.

[CHA 09] CHA M., MISLOVE A., GUMMADI K. P., « A measurement-driven analysis of information propagation in the flickr social network », *WWW '09 : Proceedings of the 18th international conference on World wide web*, New York, NY, USA, 2009, ACM.

[GOL 04] GOLDER S. A., DONATH J., « Social roles in electronic communities », *in Association of Internet Researchers (AoIR) conference Internet Research 5.0*, 2004.

[GRA 78] GRANOVETTER M., « Threshold Models of Collective Behavior », *The American Journal of Sociology*, vol. 83, n° 6, 1978, p. 1420–1443, The University of Chicago Press.

[KEM 03] KEMPE D., KLEINBERG J., TARDOS E., « Maximizing the Spread of Influence through a Social Network », *In KDD*, ACM Press, 2003, p. 137–146.

[KIM 07] KIMURA M., SAITO K., NAKANO R., « Extracting influential nodes for information diffusion on a social network », *AAAI'07 : Proceedings of the 22nd national conference on Artificial intelligence*, AAAI Press, 2007, p. 1371–1376.

[LES 07a] LESKOVEC J., ADAMIC L. A., HUBERMAN B. A., « The Dynamics of Viral Marketing », , 2007.

[LES 07b] LESKOVEC J., MCGLOHON M., FALOUTSOS C., GLANCE N., HURST M., « Cascading Behavior in Large Blog Graphs », , 2007.

[LIB 08] LIBEN-NOWELL D., KLEINBERG J., « Tracing information flow on a global scale using Internet chain-letter data », *Proceedings of the National Academy of Sciences*, vol. 105, n° 12, 2008.

[LIE 05] LIEBERMAN E., HAUERT C., NOWAK M. A., « Evolutionary dynamics on graphs », *Nature*, vol. 433, n° 7023, 2005, p. 312–316, Nature Publishing Group.

[LOP 08] LOPEZ-PINTADO D., « Diffusion in complex social networks », *Games and Economic Behavior*, vol. 62, n° 2, 2008, p. 573-590.

[MAZ 06] MAZZONI E., « 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 », 2006.

[MER 04] MERCKLÉ P., « Les origines de l'analyse des réseaux sociaux », 2004.

[SAI 09] SAITO K., KIMURA M., OHARA K., MOTODA H., « Learning Continuous-Time Information Diffusion Model for Social Behavioral Data Analysis », *ACML '09 : Proceedings of the 1st Asian Conference on Machine Learning*, Berlin, Heidelberg, 2009, Springer-Verlag, p. 322–337.

[SAI 10] SAITO K., NAKANO R., KIMURA M., « Prediction of Information Diffusion Probabilities for Independent Cascade Model », *Knowledge-Based Intelligent Information and Engineering Systems*, vol. 5179 de *Lecture Notes in Computer Science*, chapitre 9, p. 67–75, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[SAI 11] SAITO K., OHARA K., YAMAGISHI Y., KIMURA M., « Learning Diffusion Probability Based on Node Attributes in Social Networks », *Social Networks*, , n° Ic, 2011.

[TRO 01] TROTTIER H., PHILIPPE P., « Deterministic Modeling Of Infectious Diseases : Theory And Methods », *The Internet Journal of Infectious Diseases*, vol. 1, 2001.

[YAN 10] YANG J., LESKOVEC J., « Modeling Information Diffusion in Implicit Networks », *Most*, , 2010.

[YOU 09] YOUNG H. P., « Innovation Diffusion in Heterogeneous Populations : Contagion, Social Influence, and Social Learning. », Open access publications from university of oxford, 2009, University of Oxford.