Diffusion de l'information dans les réseaux sociaux

Cédric Lagnier

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Informatique**

Arrêté ministériel : 7 août 2006

Présentée par

Cédric Lagnier

Thèse dirigée par **Eric Gaussier**
et codirigée par **Gilles Bisson**

préparée au sein **Laboratoire d'Informatique de Grenoble**
et de **Ecole Doctorale Mathématiques, Informatique, Sciences et Technologies de l'Information**

Diffusion de l'information dans les réseaux sociaux

Thèse soutenue publiquement le **2 octobre 2013**,
devant le jury composé de :

Mme, Amer-Yahia Sihem
Directrice de Recherche, Présidente
Mr, Viennet Emmanuel
Professeur, Rapporteur
Mme, Largeron Christine
Professeur, Rapporteur
Mr, Denoyer Ludovic
Maître de Conférences, Examinateur
Mr, Eric Gaussier
Professeur, Directeur de thèse
Mr, Gilles Bisson
Chargé de Recherche, Co-Directeur de thèse

"Everybody is a genius.
But if you judge a fish by its ability to climb a tree,
it will live its whole life believing that it is stupid."
- Albert Einstein.

Résumé

Prédire la diffusion de l'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. La plupart des modèles proposés récemment sont des extensions des modèles à cascades et de seuil. Dans ces modèles, le processus de diffusion est basé sur les interactions entre les utilisateurs du réseau (la pression sociale), et ignore des caractéristiques importantes comme le contenu de l'information diffusé ou le rôle actif/passif des utilisateurs. Nous proposons une nouvelle famille de modèles pour prédire la façon dont le contenu se diffuse dans un réseau en prenant en compte ces nouvelles caractéristiques : le contenu diffusé, le profil des utilisateurs et leur tendance à diffuser. Nous montrons comment combiner ces caractéristiques et proposons une modélisation probabiliste pour résoudre le problème de la diffusion. Ces modèles sont illustrés et comparés avec d'autres approches sur deux jeux de données de blogs. Les résultats obtenus sur ces jeux de données montrent que prendre en compte ces caractéristiques est important pour modéliser le processus de diffusion. Enfin, nous étudions le problème de maximisation de l'influence avec ces modèles et prouvons qu'il est NP-difficile, avant de proposer une adaptation d'un algorithme glouton pour approcher la solution optimale.

Abstract

Predicting the diffusion of information in social networks is a key problem for applications like Opinion Leader Detection, Buzz Detection or Viral Marketing. Many recent diffusion models are direct extensions of the *Cascade* and *Threshold* models, initially proposed for epidemiology and social studies. In such models, the diffusion process is based on the dynamics of interactions between neighbor nodes in the network (the social pressure), and largely ignores important dimensions as the content diffused and the active/passive role users tend to have in social networks. We propose here a new family of models that aims at predicting how a content diffuses in a network by making use of additional dimensions : the content diffused, user's profile and willingness to diffuse. In particular, we show how to integrate these dimensions into simple feature functions, and propose a probabilistic modeling to account for the diffusion process. These models are then illustrated and compared with other approaches on two blog datasets. The experimental results obtained on these datasets show that taking into account these dimensions are important to accurately model the diffusion process. Lastly, we study the influence maximization problem with these models and prove that it is NP-hard, prior to propose an adaptation of the greedy algorithm to approximate the optimal solution.

Remerciements

Je tiens a remercier mon directeur de thèse Eric Gaussier, qui m'a accompagné tout au long de cette thèse. Je remercie également mon co-directeur de thèse Gilles Bisson. Je voudrais remercier les membres de mon jury, et plus particulièrement mes rapporteurs Christine Largeron et Emmanuel Viennet pour le temps passé et les retours donnés sur mon travail.

Je remercie l'ensemble des membres de l'equipe AMA avec qui j'ai prit plaisir à venir travailler tous les jours. Nous avons appris à nous connaître et certains sont même devenus de très bons amis. Il y a dans cette équipe une ambiance à la fois sympathique et studieuse qui permet un travail de qualité. Je n'oublierai pas les soirées coinches, les barbecues, les sorties en équipe, ni non plus les réunions de travail.

Je voudrais remercier les membres de ma famille, dont mes parents qui m'ont permis d'en arriver ou je suis aujourd'hui, toujours à l'écoute et très patients. Je remercie également mon frère pour les bons et les mauvais moments.

Enfin, je voudrais remercier mes amis qui, même s'ils ne s'en rendent pas compte, ont beaucoup participé à l'aboutissement de cette thèse. Dans un travail de longue halène comme celui-ci, il y a des hauts et des bas, et pour surmonter les bas il est important de pouvoir compter sur des gens qui sont toujours la pour nous redonner le sourire. Un remerciement particulier aux gros.

Table des matières

Table des figures

Liste des tableaux

Liste des Algorithmes

Introduction

Définition d'un réseau social

Un réseau social est avant tout un ensemble de personnes qui communiquent les unes avec les autres. Il peut aussi bien s'agir d'une cour d'école dans laquelle les élèves sont regroupés et discutent par classe que d'un site web comme Facebook où les utilisateurs échangent toutes sortes d'informations. Il fait donc intervenir à la fois des êtres humains et des contenus. Les utilisateurs d'un réseau sont reliés par des liens qui peuvent être explicites comme c'est le cas dans les réseaux sociaux numériques : sur Facebook on définit ses amis, sur Twitter on choisit de suivre une personne, etc. Ces liens peuvent aussi être implicites, comme pour les enfants de la cour de récréation qui discutent et jouent avec les autres enfants qui sont à coté d'eux.

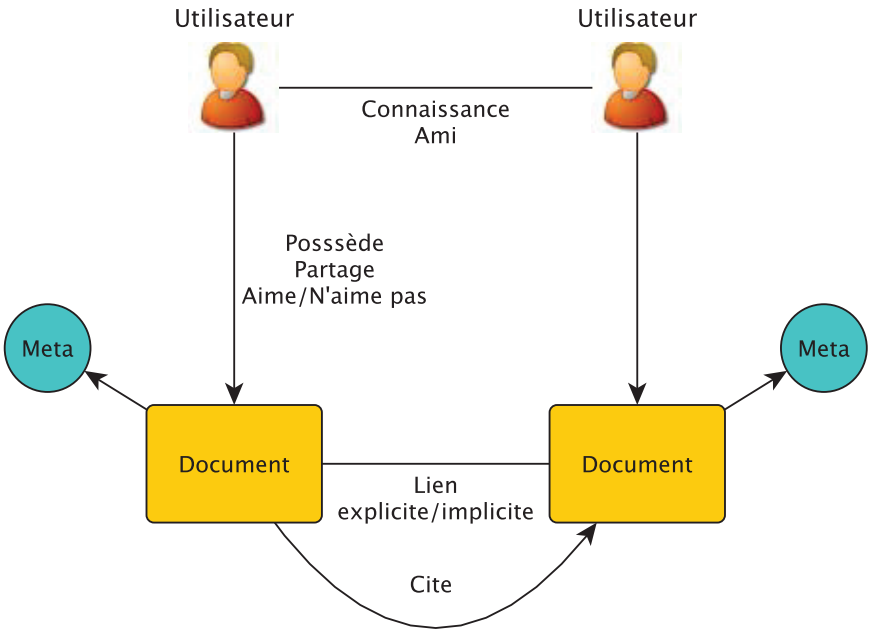


FIGURE 1 – Caractéristiques d'un réseau social.

La figure 1 montre les liens standards que l'on peut trouver dans un réseau social.

On parle de document car on s'intéresse dans cette thèse aux réseaux sociaux numériques dans lesquels les utilisateurs se partagent des contenus textuels, vidéos ou audios. Chaque contenu partagé peut s'accompagner de méta-données, c'est-à-dire des informations supplémentaires sur le contenu. Par exemple, le titre, l'auteur, la date, des annotations sont des méta-données très courantes. Les documents peuvent aussi être reliés entre eux. Un document peut en citer un autre comme c'est le cas dans les billets de blogs qui se citent les uns les autres. Deux documents peuvent avoir un auteur commun. Quand un utilisateur trouve un contenu intéressant, il peut décider de le partager. Les utilisateurs qui lui sont proches et voient ce contenu peuvent à leur tour décider de le repartager. Commence alors un phénomène de diffusion qui va faire qu'un contenu posté par un utilisateur d'un réseau social peut se propager vers un très grand nombre de personnes. Dans les réseaux sociaux numériques, un phénomène de grande propagation est appelé buzz. Certains utilisateurs ont une plus grande influence sur la diffusion comme étudié dans [Anagnostopoulos et al., 2011]. Ils sont souvent appelés "autorités". On peut retrouver des analyses des réseaux sociaux dans [Mercklé, 2004, Hanneman and Riddle, 2005].

Diffusion de l'information

On dit qu'une information se diffuse lorsque plusieurs personnes membres d'un réseau social reliées entre elles la partagent. Au démarrage d'une diffusion, seul les initiateurs sont au fait de l'information. Au fur et à mesure que les autres utilisateurs en prennent connaissance, ils peuvent choisir de la rediffuser, ou de ne pas la repartager. Le chemin ainsi parcouru par l'information forme un sous graphe $\mathcal{G}'$ du réseau social observé $\mathcal{G}$ ($\mathcal{G}' \subseteq \mathcal{G}$). Ce chemin est appelé cascade de diffusion. Un contenu partagé qui n'a jamais été repris consiste en une cascade un peu particulière puisqu'elle n'implique qu'un seul utilisateur.

Dans cette thèse, nous utiliserons indifféremment les termes diffuser et propager. De plus, lorsqu'un utilisateur a diffusé le contenu étudié, nous dirons qu'il est actif ou a été contaminé par le contenu.

Dans les réseaux sociaux, on peut observer deux types principaux de diffusion : un utilisateur peut partager une information pour tous ses voisins qui voudraient la lire, ou cibler explicitement un sous-ensemble de destinataires. Le premier cas correspond aux réseaux sociaux tels que les blogs, les conférences ou Facebook et Twitter dans leur utilisation la plus simple, alors qu'on retrouve le second cas dans un réseau de communications d'e-mails, ou d'appels téléphoniques. Même si les règles de diffusion sont plus ou moins fixées par le réseau social utilisé, les utilisateurs ont tous une approche différente en ce qui concerne quand et pourquoi diffuser un contenu [Boyd et al., 2010, Java et al., 2007]. Dans cette thèse, nous nous sommes focalisés sur la diffusion sans cible particulière, même si nos modèles peuvent être adaptés au second type mentionné ci-dessus.

Le problème de la modélisation de la diffusion dans les réseaux sociaux n'est pas

limité à l'informatique. Les premiers modèles qui ont été proposés avaient pour but de modéliser la diffusion de virus au sein d'une population pour des études en épidémiologie. Les initiateurs sont dans ce cas les premières personnes à être contaminées. Un certain nombre d'études ont ensuite cherché à modéliser l'adoption d'un nouveau produit par un ensemble d'utilisateurs. Les initiateurs de la diffusion sont dans ce cas un ensemble de personnes auxquelles le nouveau produit est offert afin d'en faire la publicité auprès des autres. Le but étant de jouer sur le bouche à oreille afin qu'un maximum de personne adopte le produit. Plus récemment, un certain nombre d'études se sont penchés sur l'étude de l'opinion à partir des discussions dans les réseaux sociaux afin de prédire la réussite de films [Asur and Huberman, 2010] ou d'une marque [Jansen et al., 2009] en utilisant twitter ou encore des forums [Stavrianou et al., 2009]. L'étude que nous en faisons en informatique n'est que très récente, due à l'apparition des réseaux sociaux numériques. Elle fait suite aux études de sociologie qui cherchaient à comprendre la façon dont les individus communiquent entre eux et comment l'information se propage au sein d'une communauté. Les articles [Adar and Adamic, 2005, Li et al., 2012] présentent une structure générale d'étude et de visualisation de la diffusion dans les réseaux sociaux.

Le problème de diffusion de l'information est lié à un certain nombre de problèmes connexes comme l'analyse de réseaux sociaux, l'annotation de réseaux ou de contenus, la détection de communautés, la prédiction de liens, etc. Il est aussi possible de s'intéresser à la prédiction du volume de la diffusion plutôt que tout le détail comme montré dans [Chierichetti et al., 2011].

La plupart des modèles existant aujourd'hui sont basés sur trois grandes familles. La première est la famille de modèles SI (Susceptible-Infected) [Anderson, 1984]. Il s'agit de modèles proposés en épidémiologie dans lesquels les utilisateurs peuvent se trouver dans plusieurs états correspondant à leur état de santé. Les transitions entre ces états sont régies par des lois de probabilités. La seconde famille de modèles est basé sur le modèle IC (Independent Cascade) [Newell and Simon, 1972]. Un utilisateur ne peut se trouver dans ce cas là que dans deux états (dépendant du fait qu'il ait diffusé ou pas le contenu) : actif ou inactif. Les initiateurs d'une diffusion sont automatiquement actifs. Dans ce modèle les utilisateurs qui prennent connaissance d'un contenu, tentent d'influencer leurs voisins afin que ceux-ci le rediffusent. Ce modèle a tout d'abord été utilisé dans des études marketing comme [Wortman, 2008]. La dernière famille de modèles est basée sur le modèle LT (Linear Threshold) [Granovetter, 1978]. Ici, à chaque utilisateur est associé un seuil qui correspond au nombre de voisins qui doivent avoir diffusé le contenu pour qu'il le diffuse à son tour. Ces modèles et leurs dérivés sont décrits dans le chapitre 1. Enfin, certains autres modèles, comme ceux de [Gupte et al., 2009, Jackson and Yariv, 2005], se basent sur des statistiques utilisateur

Ces modèles ont cependant un certain nombre de points faibles. Tout d'abord, ils reposent sur un très grand nombre de paramètres (au moins un par lien). Or dans les réseaux

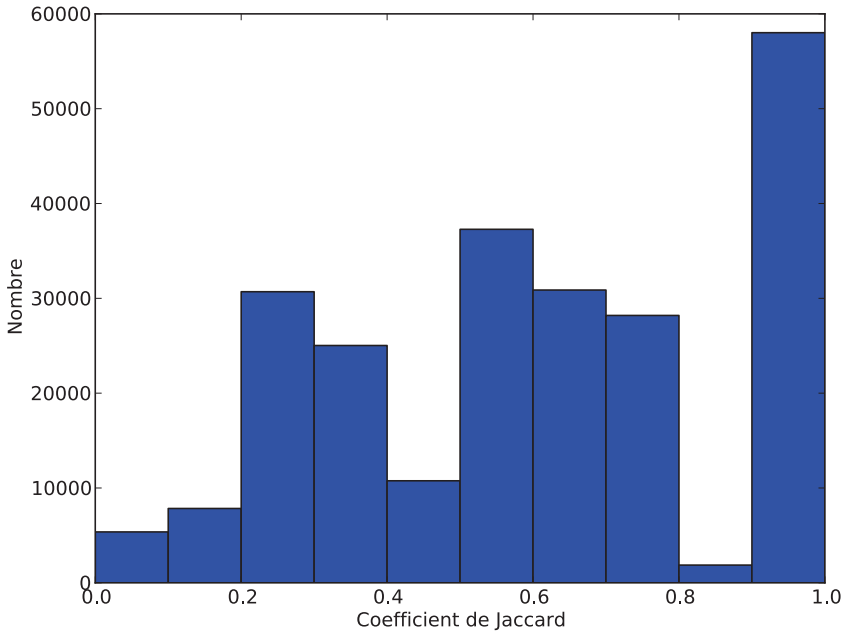


FIGURE 2 – Répartition des coefficients de Jaccard pour tous les couples de contenus issus des mêmes initiateurs.

sociaux la plupart des diffusions de contenus n'impliquent que très peu d'utilisateurs (les cascades sont souvent très courtes), ce qui a pour conséquence que la majorité des liens entre les utilisateurs ne sont utilisés que très rarement pour la diffusion. Par exemple, la plupart des billets de blogs ne sont cités que par leurs auteurs [Leskovec et al., 2007]. Dans le même ordre d'idée, il n'y a que peu de tweets [1] qui sont repris sur Twitter [Romero et al., 2011a]. Se reposer sur un grand nombre de paramètres dans ce cas là peut-être problématique car il est souvent impossible de les estimer avec précision.

Ensuite, la plupart des modèles existant dans la littérature ne tiennent pas compte de plusieurs points importants :

- ils ignorent le contenu de l'information diffusée alors que, dans un réseau social, deux contenus différents ne se propagent pas de la même manière ;
- ils ignorent les caractéristiques des utilisateurs du réseau alors que l'intérêt qu'un utilisateur a pour un contenu joue un rôle majeur dans la diffusion.

Le fait d'ignorer le contenu diffusé correspond à une hypothèse implicite forte qui est que deux contenus différents issus d'un même utilisateur vont se diffuser de la même manière.

Etude des diffusions

Si cette hypothèse était vraie, alors les ensembles d'utilisateurs atteints par la diffusion de deux contenus issus du même initiateur devraient être proches l'un de l'autre. Afin de vérifier cette hypothèse, nous avons calculé le coefficient de Jaccard [Jaccard, 1901] entre tous les couples de contenus dont la diffusion est issue du même utilisateur sur le jeu de données de blog ICWSM [2]. Nous n'avons gardé que les diffusions qui impliquent au minimum deux utilisateurs car nous ne nous intéressons qu'aux cascades qui se propagent. Ce coefficient mesure la similarité entre deux ensembles ; en d'autres mots, le coefficient de Jaccard calculé sur les deux ensembles d'utilisateurs qui ont rediffusé deux contenus devrait être proche de 1 si le processus de diffusion est le même.

La figure 2 montre un histogramme de ces valeurs, c'est à dire le nombre de cascades pour lesquelles la valeur du coefficient de Jaccard est comprise entre α et $\alpha + 0.1$. On peut voir que ce coefficient est dans la plupart des cas différent de 1 [3]. Cela signifie que deux diffusions issus des mêmes utilisateur mais de contenus différent impliquent souvent des utilisateurs différents. Ce résultat est similaire à celui que l'on trouve dans [Romero et al., 2011b] pour la diffusion d'annotations (*hashtag*) appartenant à différents thèmes sur Twitter.

Cela dit, est-ce que la diffusion est réellement liée au contenu diffusé, ou ce phénomène est-il dû à d'autres paramètres (influence extérieure au réseau par exemple) qui expliqueraient la différence dans les diffusions ?

La figure 3 montre la valeur moyenne de similarité entre deux contenus dont la diffusion est issue d'un même utilisateur et le coefficient de Jaccard calculé entre les ensembles d'utilisateurs atteints par la diffusion de ces contenus. Nous avons utilisé la similarité cosinus afin de mesurer la similarité et pour que la moyenne ait un sens, nous n'avons pas pris en compte les valeurs de coefficient de Jaccard pour lesquelles nous avons moins de vingt paires de cascades. On voit clairement sur cette figure une augmentation le long de la diagonale qui montre que des contenus similaires ont tendance à se diffuser de la même manière, alors que des contenus dissimilaires se diffusent de façons différentes. Le coefficient de corrélation de Spearman [Spearman, 1904] dans ce cas est de 0.67, indiquant une corrélation positive et non négligeable entre contenus et diffusions.

Conclusion

Nous venons de voir que la diffusion de contenus dans les réseaux sociaux suit certaines caractéristiques. Tout d'abord, la façon dont une information se diffuse est en par-

1. Un tweet est un message court qui correspond à un contenu dans le réseau social Twitter.
2. Ce jeu de données est décrit dans le chapitre 3.
3. En effet, environ 50% des paires de cascades issus des mêmes utilisateurs ont un cœfficient de Jaccard inférieur à 0.6.

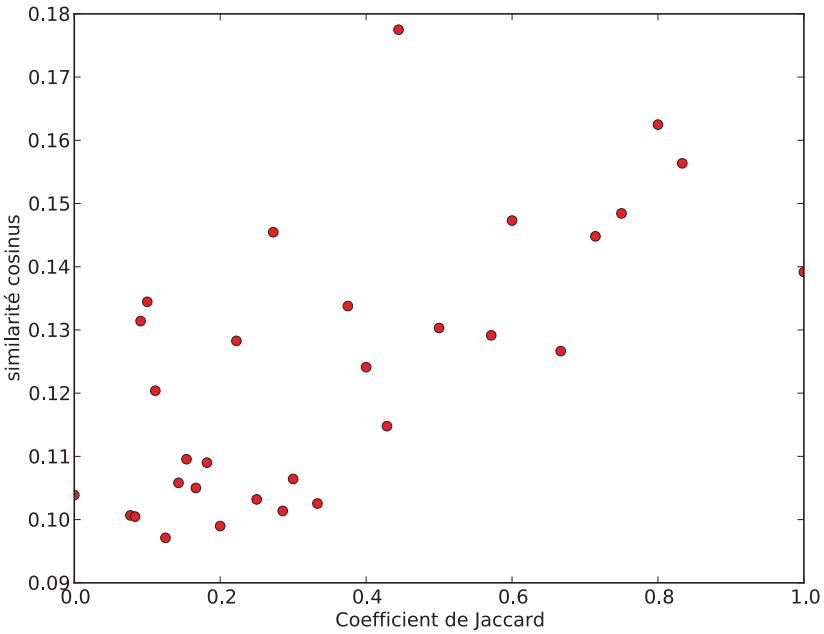


FIGURE 3 – Corrélation entre le coefficient de Jaccard et la similarité cosinus pour les couples de contenus. Chaque point correspond à la moyenne des similarités cosinus de tous les couples ayant le même coefficient de Jaccard.

tie déterminée par son contenu. De plus, deux contenus issus d'un même diffuseur initial ne se diffusent pas exactement de la même manière. L'hypothèse faite par la plupart des modèles du domaine est donc caduque et il importe de prendre en compte le contenu afin de modéliser au mieux la diffusion d'informations dans les réseaux sociaux. Cela ne veut pas dire pour autant qu'il faut en oublier la topologie du réseau. Elle est importante et influence la diffusion comme le montrent [Lieberman et al., 2005].

Notations

Nous considérons ici un graphe social orienté $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ composé d'un ensemble d'utilisateurs $\mathcal{N} = \{n_1, ..., n_N\}$ et d'un ensemble de liens $\mathcal{E}$ entre ces utilisateurs. Un utilisateur n_i est relié à un utilisateur n_j si $(n_i, n_j) \in \mathcal{E}$. Nous utilisons de plus les notations suivantes :

– $\mathcal{B}(n_i)$ dénote l'ensemble des voisins entrant de l'utilisateur n_i (les utilisateurs qui ont un lien vers n_i) :

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

$\mathcal{F}(n_i)$ dénote l'ensemble des voisins sortant de l'utilisateur n_i (les utilisateurs vers lesquels n_i a un lien) :

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

- Tous les utilisateurs ont un profil basé sur ce qu'ils aiment ou n'aiment pas. $\mathcal{P}$ est l'ensemble de tous les profils des utilisateurs et $\forall i, 1 \leq i \leq N, p^i$ dénote le profil de l'utilisateur n_i. Dans cette thèse, p^i est un vecteur de caractéristiques calculé pour chaque utilisateur. F dénote le nombre de caractéristiques : $\forall i, 1 \leq i \leq N, |p^i| = F$.
- $\mathcal{C} = \{c^1, ..., c^K\}$ est l'ensemble de toutes les informations qui se propagent sur le réseau. Un contenu est défini dans le même espace de caractéristiques que les profils des utilisateurs : $\forall k, 1 \leq k \leq K, |c^k| = F$.
- $\mathcal{M} = \{M^1, ..., M^K\}$ est l'ensemble des matrices de diffusion pour chaque contenu c^k. $\forall k$, M^k est de la forme :

$$M^k = \begin{pmatrix} m^k_{1,0} & m^k_{1,1} & m^k_{1,2} & \dots & m^k_{1,T^k} \\ m^k_{2,0} & m^k_{2,1} & m^k_{2,2} & \dots & m^k_{2,T^k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m^k_{N,0} & m^k_{N,1} & m^k_{2,2} & \dots & m^k_{N,T^k} \end{pmatrix}$$

  où $m^k_{i,t} \in \{0, 1\}$; $m^k_{i,t} = 1$ indique que l'utilisateur n_i a diffusé le contenu c^k avant ou au temps t. T^k correspond à la durée de la diffusion du contenu c^k en étapes de temps. Une étape de temps peut correspondre aussi bien à une seconde, une minute, une heure, etc. Dans cette thèse, nous considérons des étapes de temps journalières. Enfin, l'ensemble $\mathcal{M}$ est divisé en deux sous-ensembles disjoints : un ensemble de matrices utilisées pour l'apprentissage des modèles $\mathcal{D} = \{(M^1, c^1), ..., (M^\ell, c^\ell)\}$, et un ensemble de matrices utilisées pour l'évaluation des modèles $\mathcal{T} = \{(M^{\ell+1}, c^{\ell+1}), ..., (M^K, c^K)\}$.
- $D^k(t)$ est l'ensemble des utilisateurs qui ont diffusé le contenu c^k au temps t ($D^k(t) = \{n_i/m^k_{i,t} = 1, m^k_{i,t-1} = 0\}$), et $C^k(t)$ est l'ensemble des utilisateurs qui ont diffusé le contenu c^k avant ou au temps t ($C^k(t) = \cup^t_{t'=0} D^k(t')$).

Afin de rendre le discours plus simple, nous utilisons aussi l'ensemble $C^k(n_i, t)$ de voisins entrant de l'utilisateurs n_i qui ont diffusé le contenu c^k avant ou au temps t.

Contributions

Cette thèse de doctorat se place dans le cadre de la fouille de données dans les réseaux sociaux. Dans l'équipe AMA du LIG, deux autres thèses traitent de sujets proches :

- la thèse de Clément Grimal [Grimal, 2012] qui a été soutenue le 11 octobre 2012 et qui étudie le problème de l'*apprentissage de co-similarités pour la classification automatique de données monovues et multivues.*
- la thèse en cours de François Kawala au sein du groupe *Best of Media* qui traite de *méthodes de détection et prédiction des phénomènes de buzz sur les réseaux sociaux.*

Dans cette thèse, nous traitons le problème de la prédiction de la diffusion de l'information dans les réseaux sociaux en proposant une nouvelle famille de modèles probabilistes. Ces modèles sont appelés *centrés utilisateur* car ils prennent en compte un certain nombre de caractéristiques propres aux utilisateur afin de calculer leur probabilité de diffuser un contenu donné. Ils prennent en compte :

- le contenu de l'information diffusé ;
- le goûts des utilisateurs et leur intérêt pour le contenu diffusé ;
- la tendance des utilisateurs à diffuser, c'est-à-dire s'ils diffusent peu ou beaucoup d'information ;
- la pression sociale subie par les utilisateurs, c'est-à-dire le nombre de leurs voisins ayant diffusé le contenu.

Nous proposons trois modèles qui utilisent ces probabilités de diffusion : le modèle UC qui est une adaptation du modèle à cascades indépendantes pour prendre en compte cette dimension utilisateur et les modèles RUC et DRUC qui intègrent un renforcement sur le temps, permettant une modélisation plus fine, étape par étape, de la diffusion. Les modèles standards peuvent tous s'exprimer comme une percolation de liens. Nous montrons que ce n'est pas le cas pour les modèles centrés utilisateur avec renforcement. Enfin nous illustrons le phénomène de diffusion engendré par ces modèles sur des graphes de réseaux sociaux à la fois artificiels et réels.

Afin de tester la qualité de ces modèles, nous comparons leurs résultats dans une tâche de prédiction de la diffusion de contenus à ceux des modèles standards IC, ASIC et Netrate. Nous utilisons pour nos expériences des jeux de données artificiels ainsi que deux jeux de données réels de billets de blogs : ICWSM et Memetracker. Nous évaluons les résultats d'une part en utilisant une mesure de précision qui nous permet de montrer que les modèles centrés utilisateur parviennent mieux à différencier les acteurs d'une diffusion par rapport aux autres utilisateurs du réseau. D'autre part, une mesure d'erreur de volume nous permet de voir que tous ces modèles, qui modélisent la diffusion étape par étape, ont du mal à prédire la quantité exacte d'utilisateurs touchés par la diffusion d'un contenu. Enfin une étude des paramètres des modèles centrés utilisateur nous permet de montrer l'importance de l'intérêt thématique des utilisateurs dans la modélisation de la diffusion par ces modèles.

Dans un second temps nous étudions le problème de maximisation de l'influence qui consiste, connaissant un réseau social et un modèle de diffusion à trouver les k meilleurs initiateurs pour que la diffusion d'un contenu soit maximale. Ce problème est prouvé NP-difficile pour les modèles standards IC et LT. Nous montrons que c'est aussi le cas pour les modèles centrés utilisateur. L'algorithme glouton *greedy hill climbing* est une bonne approximation du problème dans le cas où la fonction à optimiser est sous-modulaire, ce qui est le cas pour les modèles standards. Nous montrons que pour les modèles RUC et DRUC cette fonction n'est pas sous-modulaire. Cependant nos expériences sur l'utilisation

de l'algorithme et d'une variante de celui-ci montrent que, même si la fonction à optimiser n'est pas sous-modulaire, l'algorithme *greedy hill climbing* donne de bonnes approximations du résultat optimal sur de petits jeux de données artificiels. Nous montrons aussi sur des jeux de données réels que l'algorithme glouton obtient de meilleurs résultats que des heuristiques simples.

Plan de la thèse

Le chapitre 1 présente les différents modèles que l'on trouve dans la littérature : les modèles simples de diffusion ainsi que leurs extensions. Dans le chapitre 2 nous présentons une nouvelle famille de modèles probabilistes qui tiennent compte des caractéristiques des utilisateurs et du contenu diffusé. Ils résolvent aussi le problème du grand nombre de paramètres en n'étant basés que sur quatre paramètres. Ces modèles permettent d'obtenir de meilleurs résultats que les modèles standards du domaine comme nous le montrons dans le chapitre 3 en validant les modèles sur une tâche de prédiction de la diffusion sur deux jeux de données de blogs. Nous abordons le problème de la maximisation de l'influence, qui consiste à chercher les meilleurs initiateurs pour une diffusion maximale dans le chapitre 4. Nous concluons ensuite et présentons un certain nombre de perspectives.

Chapitre 1

Etat de l'art

Sommaire

1.1 Modèles de propagation simples

Les premiers modèles de propagation qui ont étés proposés sont basés sur l'idée principale que les utilisateurs d'un réseau interagissent les uns avec les autres. Dans le cas de la diffusion d'un virus, les personnes qui côtoient une personne contaminée, i.e. sont proches d'elle dans le réseau, ont des chances d'être atteints par le même virus. En marketing, si une personne adopte un nouveau produit, elle a des chances d'influencer les personnes qu'elle connaît, qui adoptent à leur tour ce même produit. Dans un groupe de personnes qui communiquent, l'information circule. Une rumeur va se propager d'un utilisateur à un autre par l'effet du bouche à oreille. Tous ces modèles prennent en compte le temps de manière discrète : l'étape $t+1$ suis directement l'étape t. Ces étapes ne correspondent pas forcément à une unité de temps réel comme la seconde, la minute ou l'heure.

1.1.1 Modèles épidémiques : famille de modèles SI

Cette famille de modèles a été introduite afin de prédire la diffusion de virus au sein d'une population. Chaque modèle définit un ensemble d'état dans lesquels peuvent se trouver les utilisateur. Ils sont définis dans les articles [Trottier and Philippe, 2001, Newman, 2003]. Le modèle le plus simple de cette famille, SI (Susceptible-Infected), définit deux états :

- un utilisateur Susceptible n'est pas atteint par le virus, mais peut le devenir si d'autres personnes dans la population en sont porteuses ;
- un utilisateur Infectieux est atteint par le virus et peut contaminer d'autres personnes au sein de la population.

La figure 1.1 décrit les états et les transitions possibles pour les principaux modèles de cette famille. Dans le modèle SIS, un utilisateur Infectieux peut redevenir sain, contrairement au modèle SI dans lequel les utilisateurs Infectieux le restent *ad vitam æternam*. L'état Rétabli correspond à un état sain dans lequel les utilisateurs ne peuvent pas être à nouveau atteints par le virus. Certains modèles ajoutent un état Exposé : à ce moment là un utilisateur Susceptible ne peut plus directement être contaminé, il doit d'abord être Exposé au virus pour pouvoir être atteint par la diffusion. Un certain nombre de variantes existent avec ces états.

Ces modèles ne représentent pas l'état de chaque utilisateur mais la proportion d'utilisateurs de la population dans chacun des états. Ils modélisent de manière globale le système. Les utilisateurs peuvent changer d'état en fonction des transitions définies par le modèle et ainsi changer ces proportions. La figure 1.2 nous sert d'exemple en utilisant le modèle SIR. Les autres modèles de cette famille ont le même fonctionnement avec des états différents. $S(t)$ correspond à la proportion d'utilisateurs Susceptibles au temps t, $I(t)$ la proportion d'utilisateurs Infectieux et $R(t)$ la proportion d'utilisateurs Rétablis.

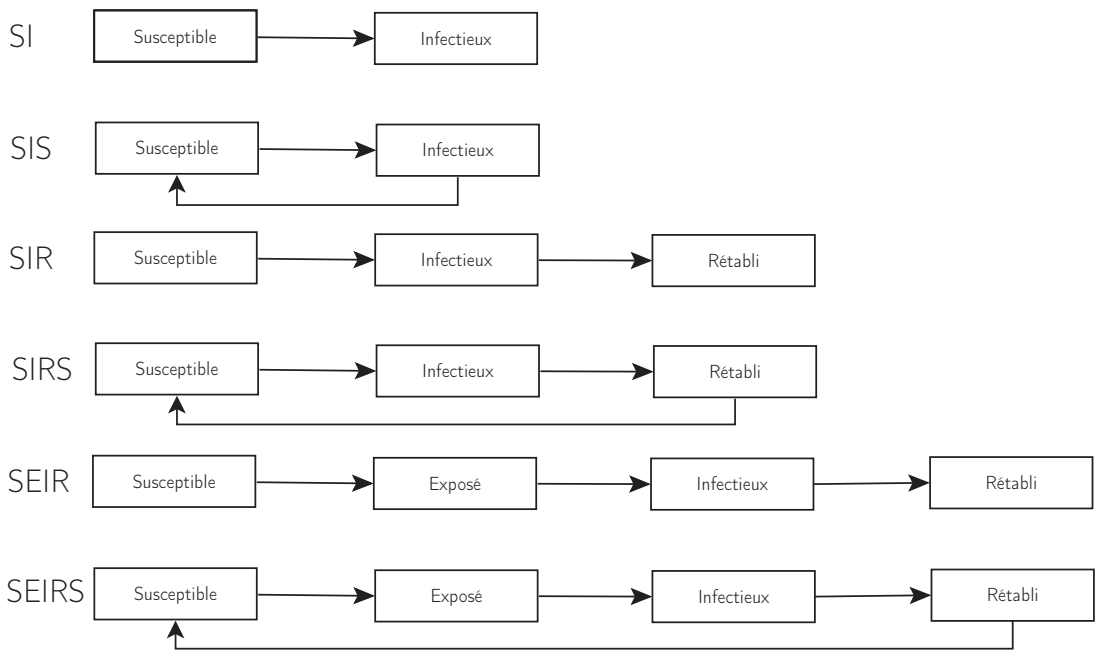


FIGURE 1.1 – Famille de modèles SI.

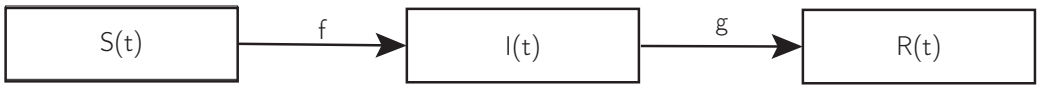


FIGURE 1.2 – Propagation en utilisant le modèle SIR.

Prenons un exemple : dans une population de 6 personnes, à une étape de temps t, deux sont infectieuses et une est rétablie. Le reste de la population n'a encore jamais été atteint par le virus. On aura alors :

$$\begin{aligned} S(t) &= 1/2 \\ I(t) &= 1/3 \\ R(t) &= 1/6 \end{aligned}$$

Pour chacun des trois états, on peut définir la proportion des utilisateurs à une étape par rapport à la précédente :

$$\begin{aligned} S(t+1) &= S(t) + \frac{dS(t)}{dt} \\ I(t+1) &= I(t) + \frac{dI(t)}{dt} \\ R(t+1) &= R(t) + \frac{dR(t)}{dt} \end{aligned}$$

Comme indiqué en figure 1.2, des taux de transitions sont fixés entre les états. f est la proportion d'utilisateurs qui sont susceptibles et deviennent infectieux d'une étape à l'autre. Si aucun utilisateur n'était Susceptible, il n'y aurait pas de nouveau utilisateurs infectieux. g correspond à la proportion d'utilisateurs qui étaient infectieux et deviennent rétablis. On a ainsi la définition des dérivées suivante :

$$\begin{aligned} \frac{dS(t)}{dt} &= -f.S(t) \\ \frac{dI(t)}{dt} &= f.S(t) - g.I(t) \\ \frac{dR(t)}{dt} &= g.I(t) \end{aligned}$$

L'article [Schaffer and Bronnikovab, 2007] présente une extension du modèles SEIR dans lequel les auteurs incluent la mort des utilisateurs. Celle-ci peut se produire pour tous les utilisateurs, quel que soit leur état. Afin que le système comporte un nombre de personnes constant, à chaque fois qu'une personne meurt, une nouvelle naît et se trouve dans l'état Susceptible.

Des variantes de tous ces modèles existent en prenant en compte la topologie du réseau. Seuls les voisins entrant infectieux d'un utilisateur peuvent le contaminer. L'article [López-Pintado, 2008] présente une variante du modèle SIS dans laquelle la probabilité de diffusion des utilisateurs dépend de leur connectivité. L'auteur étudie les contraintes qui entraînent soit une forte diffusion, soit très peu de diffusion. Dans [Prakash et al., 2010], les auteurs utilisent un modèle dans lequel les probabilités évoluent au cours du temps. Enfin, il est possible de montrer certaines propriétés liés à ces modèles comme un seuil de diffusion au-delà duquel on ne peut stopper la propagation [Prakash et al., 2011].

1.1.2 Modèle à cascades indépendantes : IC

Le concept du modèle IC (Independent Cascade) est proche de celui du modèle SI. Proposé dans les articles [Saito et al., 2008, Newell and Simon, 1972], comme pour la variante locale de SI, il modélise l'évolution de chaque utilisateur et non pas l'état global du système. Chaque utilisateur peut se trouver dans deux états : inactif et actif. Lors de la diffusion d'un produit ou d'une information, un utilisateur dans l'état inactif n'est pas encore atteint par la diffusion, il n'a pas encore diffusé le contenu ou le produit. A partir du moment où il l'a reçu par un de ses voisins dans le réseau et l'a diffusé à son tour, il devient actif. La figure 1.3 montre les états définis par ce modèle. A chaque lien du réseau entre deux utilisateurs n_i et n_j est associé une probabilité $p_{i,j}$. Un utilisateur n_i qui devient actif (choisit de diffuser) à une étape de temps t a une unique chance d'influencer chacun de ses voisins pour qu'ils diffusent à leur tour. Un de ses voisins n_j a une probabilité $p_{i,j}$ de diffuser à l'étape $t+1$. Si n_i réussit à influencer n_j, ce dernier devient actif. Si au contraire il échoue, il ne pourra plus l'influencer à nouveau dans le futur. D'une certaine façon, il passe en arrière plan, toujours actif, et ne peut plus intervenir.

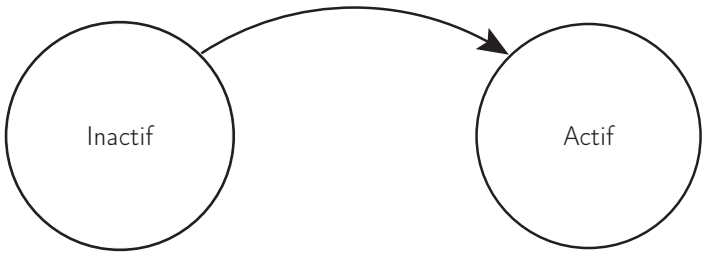


FIGURE 1.3 – Etats du modèle IC.

La figure 1.4 montre un exemple d'exécution du modèle IC sur un réseau jouet. Le réseau comprend six utilisateurs et sept liens entre eux qui définissent les influences. A chaque lien est associé une probabilité. A l'étape 0, l'utilisateur n_5 est actif et tous les autres sont inactifs. Il est considéré comme le diffuseur initial. Comme il vient de devenir actif, il tente d'influencer ses trois voisins n_1, n_3 et n_6. Il échoue pour n_1 et n_6 mais réussit pour n_3, donc, à l'étape 1, n_3 est actif en plus de n_1. Comme n_3 est nouvellement actif, il va tenter à son tour d'influencer ses voisins. Ce processus se poursuit jusqu'à une étape de temps à laquelle aucun utilisateur n'est devenu actif. On considère à ce moment là la diffusion terminée.

Vu qu'à chaque lien est associé un paramètre, ce modèle en possède autant que de liens. Leur apprentissage peut être effectué par maximum de vraisemblance comme indiqué dans [Saito et al., 2008].

1.1.3 Modèles de seuil

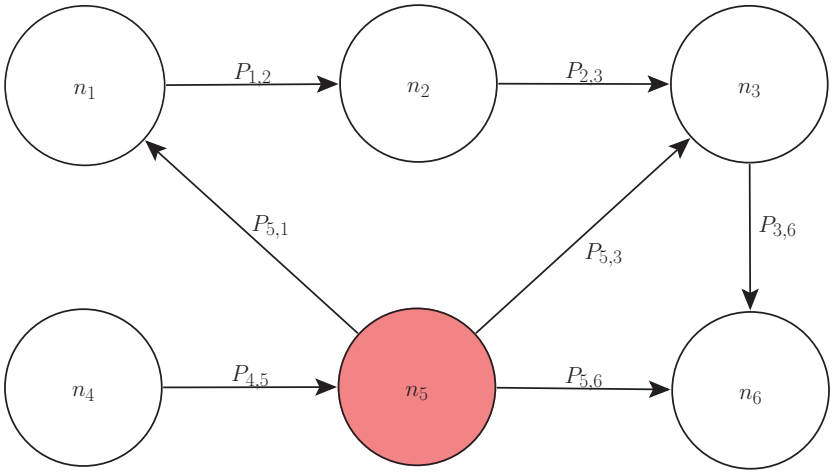
Dans les modèles précédents, un utilisateur avait une probabilité de prendre la décision d'adopter un produit. Dans le cas des modèles de seuil, chaque utilisateur a un élément déclencheur qui va faire qu'il adopte le produit. L'élément le plus simple étant représenté par un seuil sur le nombre de voisins. Cela peut être par exemple : si un utilisateur a au moins trois de ses voisins dans le réseau qui ont adopté le produit, alors il choisit de l'adopter lui aussi.

Modèle de seuil linéaire

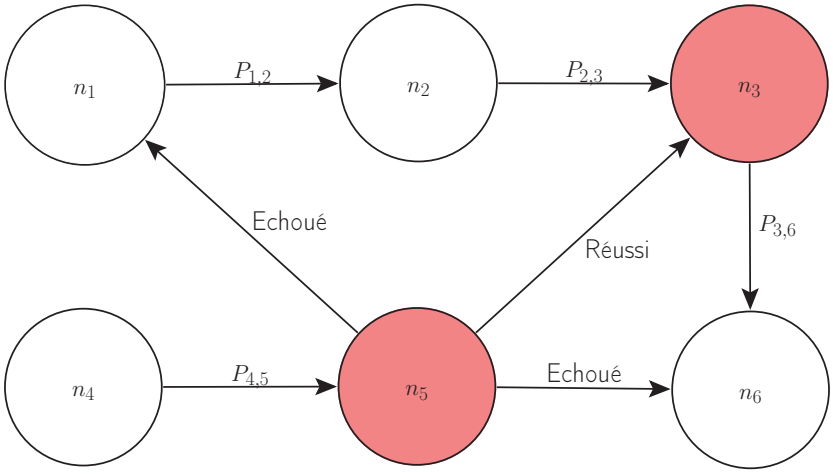
Dans le modèle de seuil linéaire [Granovetter, 1978, Kempe et al., 2003], un poids non négatif est associé à chacun des liens du graphe. Il représente la "force" du lien entre les deux utilisateurs. Le poids sur le lien entre les utilisateurs n_i est n_j est dénoté $w_{i,j}$ et est contraint par :

$$\sum_{i \in B(j)} w_{i,j} \leq 1$$

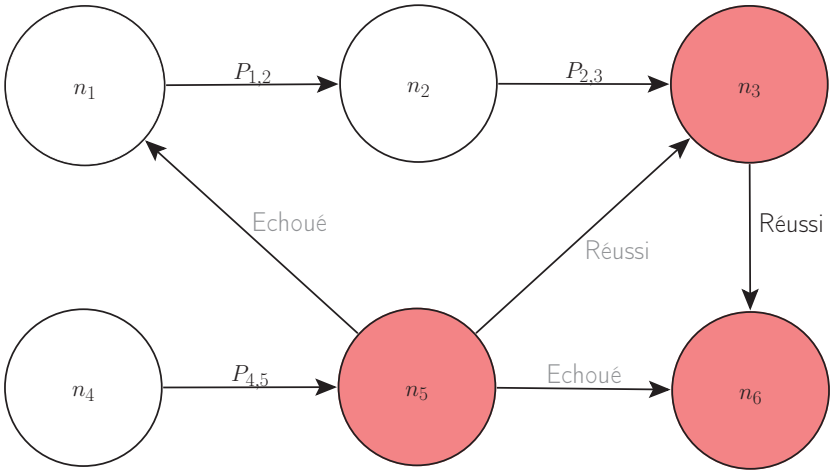
De plus, chaque utilisateur n_j choisit un seuil d'activation θ_j. Il représente la proportion pondérée de ses voisins entrant qui doivent déjà être actifs pour qu'il s'active à son tour.



(a) Etape 0



(b) Etape 1



(c) Etape 2

FIGURE 1.4 – Propagation en utilisant le modèle IC.

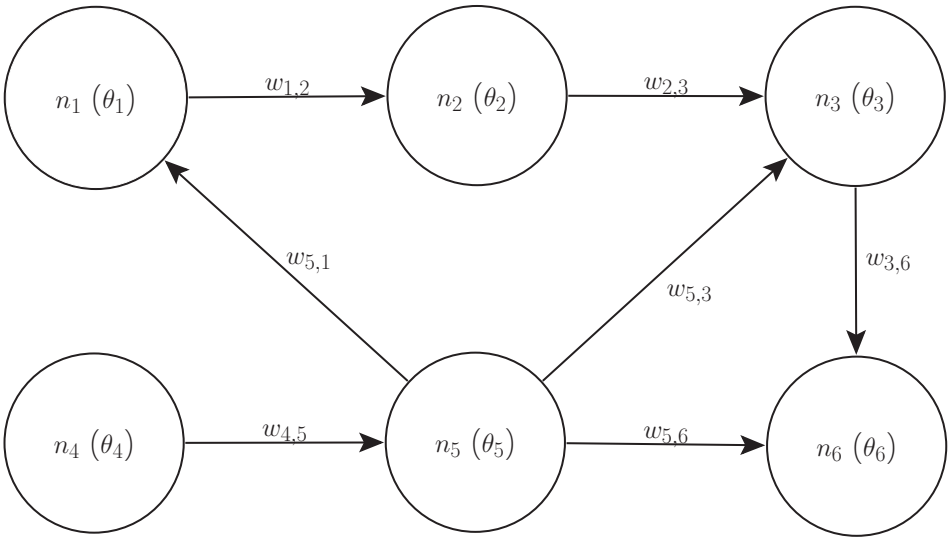


FIGURE 1.5 – Exemple d'utilisation du modèle LT

La figure 1.5 montre un exemple de réseau social pour le modèle LT. Avec un ensemble de seuils donné et un ensemble d'initiateurs de la diffusion, le processus de diffusion est déterministe. Comme pour les modèles précédents, les utilisateurs actifs à une étape de temps t le restent à l'étape $t+1$. De plus, tous les utilisateurs n_j dont la proportion pondérée de voisins actifs est supérieure à leur seuil θ_j s'activent :

$$\sum_{i \in \phi(t), i \in B(j)} w_{i,j} \geq \theta_j$$

où $\phi(t)$ est l'ensemble des utilisateurs actifs à l'étape de temps t. Le seuil d'un utilisateur représente sa propension à adopter un produit quand ses voisins le font. Les valeurs de seuil sont choisies de manière aléatoire afin de modéliser le manque de connaissance sur les utilisateurs.

Dans [Granovetter and Soong, 1988] les auteurs présentent un modèle de seuil dans lequel les utilisateurs sont tous connectés les uns aux autres dans une communauté. Ce modèle est en quelque sorte une version globale de LT, semblable au modèle SI avec des seuils au lieu de probabilités définissant les choix des utilisateurs.

Version stochastique du modèle de seuil linéaire

Le modèle de seuil linéaire décide si un utilisateur va adopter un produit ou non en fonction d'un seuil précis : en dessous de la valeur de seuil, il n'adopte pas, au dessus il adopte. Dans [Macy, 1991] est proposé une variante de modèle LT avec un seuil flou. Soit

$\pi_j(t)$ le taux de participation des voisins de l'utilisateur n_j au temps t :

$$\pi_j(t) = \sum_{i \in \phi(t), i \in B(j)} w_{i,j}$$

La probabilité que l'utilisateur n_j adopte le produit est alors définit par une fonction de seuil logistique de la distance entre $\pi_j(t)$ et θ_j :

$$P_j(t) = \frac{1}{1 + e^{\alpha(\theta_j - \pi_j(t))}}$$

où α est un paramètre de la fonction logistique ($\alpha > 1$). La décision de l'utilisateur n_j d'adopter le produit dépend donc de la probabilité $P_j(t)$. Cette probabilité augmente plus le taux de participation des voisins est important. Si $\pi_j(t) = \theta_j$, la probabilité que n_j adopte le produit est de 0.5.

Diffusion de deux produits en concurrence

Dans la réalité, il n'y a rarement qu'un seul produit, ou qu'une seule information qui se diffuse en même temps, et souvent il y a de la concurrence. Une adaptation du modèle LT est proposée dans l'article [Borodin et al., 2010] pour tenir compte de la diffusion de plusieurs produits en parallèle dans un réseau. On observe la diffusion de deux produits A et B dans une communauté. $\phi_A(t)$ dénote l'ensemble des utilisateurs ayant adopté le produit A au temps t. De la même manière, $\phi_B(t)$ dénote l'ensemble des utilisateurs ayant adopté le produit B au temps t. Chaque utilisateur n_j possède un unique seuil d'activation θ_j et il adopte un produit si le taux de participation de ses voisins pour les deux produits est supérieur à ce seuil :

$$\sum_{i \in \phi_A(t) \cup \phi_B(t), i \in B(j)} w_{i,j} \geq \theta_j$$

Il choisit le produit qu'il adopte en fonction de ceux qu'ont adoptés ses voisins entrant. Il choisit le produit A avec la probabilité $P_j^A(t)$ égale au rapport du taux de participation de ses voisins pour le produit A sur le taux de participation de ses voisins pour les deux produits :

$$P_j^A(t) = \frac{\sum_{i \in \phi_A(t), i \in B(j)} w_{i,j}}{\sum_{i \in \phi_B(t) \cup \phi_B(t), i \in B(j)} w_{i,j}}$$

S'il n'adopte pas le produit A, il adopte alors le produit B. Dans tous les cas il adopte un des deux produits si le taux de participation de ses voisins aux deux produits est supérieur à son seuil. Il est possible d'exprimer avec ce modèle un problème de diffusion sur un modèle de seuil linéaire simple en choisissant un ensemble d'adopteurs initial vide pour l'un des deux produits.

1.2 Equivalence des modèles de propagation simples avec une percolation de lien

Les modèles que nous avons présentés, SI, IC et LT, sont différents de par leur exécution. Le processus qui régit la diffusion n'est pas le même. La percolation de lien est une méthode initialement utilisée pour étudier l'écoulement de liquides en physique. Pour chaque problème de diffusion qui suit l'un des modèles standards, il existe une percolation de lien correspondante. L'inverse n'est pas vrai. Ces modèles sont des sous-problèmes de percolation de lien.

1.2.1 Percolation de lien

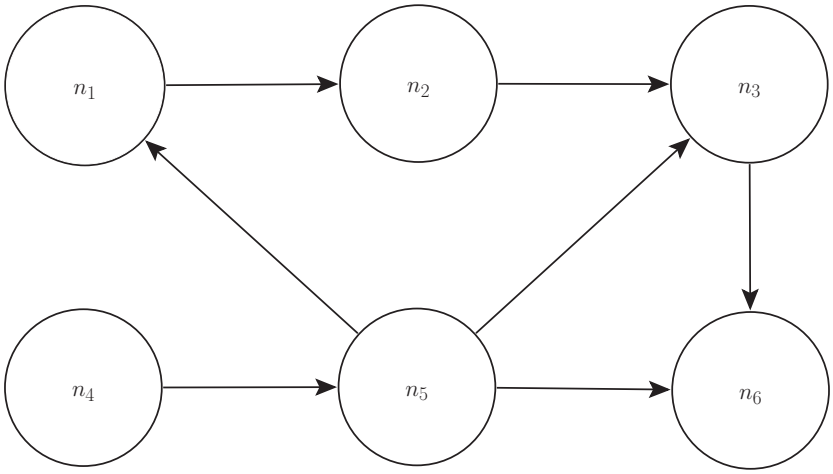
La percolation de lien est une méthode qui a été proposée pour étudier la propagation d'un liquide dans un corps poreux. En versant initialement le liquide au sommet d'un corps poreux, la question est de savoir avec quelle probabilité il atteindra le bas du corps. Le corps poreux est représenté par un graphe dans lequel les liens indiquent les chemins que peut emprunter le liquide. Les principes de percolation sont expliqués en détail dans [Stauffer and Aharony, 1992].

Dans notre cas, le graphe correspond au réseau social $\mathcal{G}$ sur lequel nous étudions la diffusion, les liens correspondant aux chemins que l'information peut emprunter pour se propager. L'idée est de représenter la diffusion en supprimant toute probabilité : on considère un sous-graphe $\mathcal{G}'$ de $\mathcal{G}$ dans lequel une partie des liens sont considérés comme ne laissant pas filtrer l'information (ceux qui ne sont pas présents dans le graphe $\mathcal{G}'$) et l'autre partie laissent passer l'information sans aucune restriction. On peut dire qu'il y a une probabilité de diffusion de 1 à travers les liens de $\mathcal{G}'$. Le processus de diffusion est ainsi complètement déterministe et simple à mettre en œuvre. Nous appelons $R_{\mathcal{G}}$ l'ensemble des instances possibles de percolation de lien sur le graphe $\mathcal{G} = (\mathcal{N}, \mathcal{E})$:

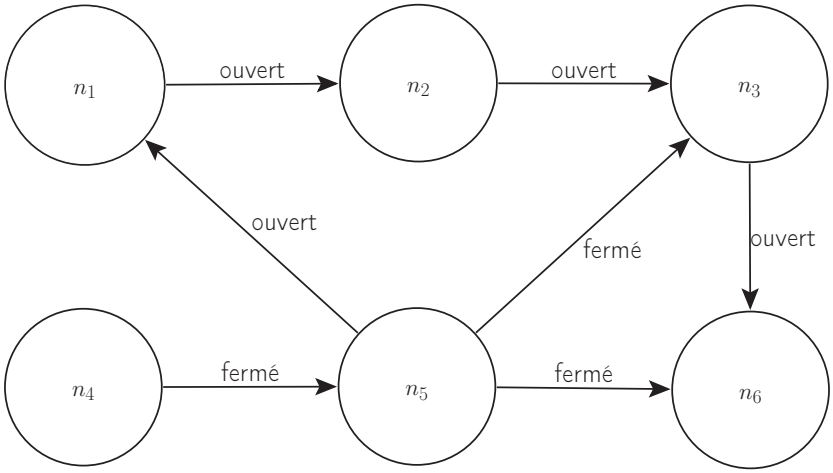
$$R_{\mathcal{G}} = \{r = (r_{i,j})_{(n_i,n_j)\in\mathcal{E}} \in \{0,1\}^{|\mathcal{E}|}\}$$

$r_{i,j} = 1$ si le lien entre les utilisateurs n_i et n_j est présent dans l'instance de percolation r, il vaut 0 dans le cas contraire. Une distribution de probabilité $q(r)$ est définie sur l'ensemble $R_{\mathcal{G}}$ afin de déterminer la probabilité de chaque instance r.

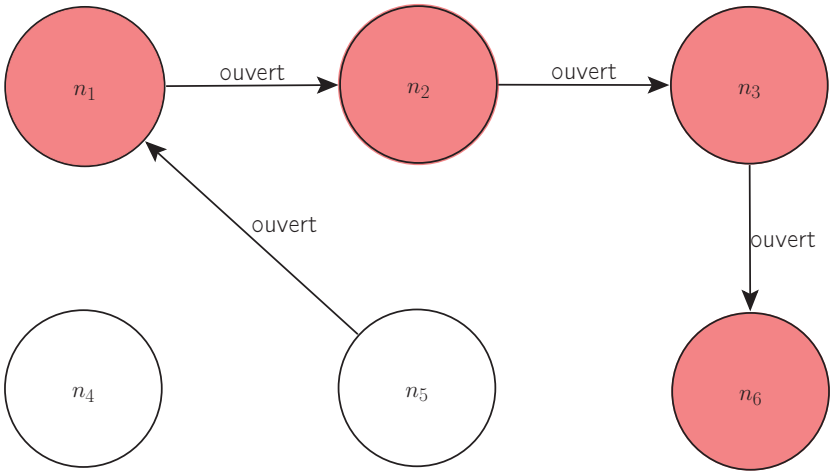
Nous considérons un graphe social avec six utilisateurs et sept liens entre eux représenté sur le figure 1.6(a). Une instance de percolation correspond à un sous-graphe du réseau social : pour chaque lien, il peut soit être considéré comme ouvert et laisser passer toutes les informations, soit être fermé et filtrer tous les contenus. Il existe ainsi 2^7 instances de percolation possibles pour ce réseau social. La figure 1.6(b) montre une instance de percolation pour laquelle quatre liens sont ouverts et trois sont fermés. Le processus de diffusion sur ce graphe se fait ensuite en suivant simplement les liens ouverts. La figure



(a) Graphe social



(b) Une instance de percolation de lien



(c) Exemple de diffusion sur une instance de percolation de lien

FIGURE 1.6 – Exemple de percolation de lien.

1.6(c) montre la diffusion d'un contenu que l'utilisateur n_1 a initialement diffusé. Les nœuds en rouge sont les utilisateurs contaminés à la fin de la diffusion.

Nous avons dit qu'à chaque instance de percolation correspond une probabilité d'apparition. Il est ensuite possible de tirer de ces probabilités la probabilité qu'un utilisateur donné soit contaminé lors une diffusion. Prenons l'exemple d'une diffusion partant de l'utilisateur n_1. La probabilité que l'utilisateur n_2 soit atteint par la diffusion correspond à la probabilité que le lien entre n_1 et n_2 soit ouvert. Dans la configuration que nous avons donnée précédemment, c'est le cas. Mais c'est aussi le cas dans la configuration où tous les liens sont fermés sauf celui-là. On peut ainsi sommer les probabilités des instances pour lesquelles ce lien est ouvert afin d'obtenir la probabilité que l'utilisateur n_2 rediffuse le contenu provenant de l'utilisateur n_1.

Enfin, une percolation de lien nécessite une distribution particulière sur les probabilités des instances de la percolation. Elle implique que pour un paramètre donné p on observe un phénomène de changement de phase lorsque p dépasse un certain seuil critique p_0. Dans le cadre de la diffusion d'un liquide à travers un corps poreux, il s'agit de la probabilité que le liquide puisse traverser le corps. Dans le cadre de la diffusion de contenu, il s'agit de la probabilité que le contenu soit diffusé par un grand nombre d'utilisateurs. C'est le cas lorsque la distribution de probabilités sur les instances de la percolation favorise les instances pour lesquelles l'ensemble des utilisateurs du graphes sont atteignables par une diffusion.

La famille de modèles SI ne tenant pas compte du graphe, il n'est pas possible de définir de distribution sur des instances de percolation de liens lui correspondant. Il est néanmoins facile de voir qu'il existe un seuil critique sur les probabilités de transition entre les états au delà duquel l'ensemble de la communauté observé est contaminé. Ces modèles ont donc une propriété de percolation. On peut trouver plus de détails sur la correspondance entre ces modèles et un processus de percolation dans l'article [Grassberger, 1983].

1.2.2 Equivalence entre IC et une percolation de liens

Cette équivalence a été montré dans [Kempe et al., 2003] pour étudier la maximisation de l'influence en utilisant les modèles IC et LT. Une explication plus détaillée est présente dans [Kimura et al., 2007].

Une instance du modèle à cascades indépendantes sur un réseau social $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ est définie par un ensemble de probabilités pour chacun des liens du graphe :

$$\forall (n_i, n_j) \in \mathcal{E}, p_{i,j} \in [0, 1]$$

où chaque probabilité est indépendante des autres.

Exprimer ce problème sous la forme d'une percolation de lien se fait en suivant ces probabilités. Pour une instance donnée $r \in R_{\mathcal{G}}$ de percolation de lien, on peut faire un

tirage pour chacun de ses liens. Sa probabilité d'apparition est donc la probabilité que le tirage des liens ouverts dans l'instance ait réussi et que le tirage des liens fermés ait échoué :

$$q(r) = \prod_{(n_i,n_j)\in\mathcal{E}} \left((p_{i,j})^{r_{i,j}}(1-p_{i,j})^{1-r_{i,j}}\right)$$

Dans le cas du modèle IC, le paramètre de changement de phase est l'ensemble des probabilités associées aux liens du graphe. Il existe un seuil critique pour ces probabilités (qui vaut 1 si tous les liens doivent être ouverts) au delà duquel l'ensemble des utilisateurs atteignables depuis les initiateurs sera atteint par une diffusion.

1.2.3 Equivalence entre LT et une percolation de liens

Comme pour la démonstration de l'equivalence entre IC et une percolation de lien, des explications de cette démonstration se trouvent dans [Kempe et al., 2003, Kimura et al., 2007].

Une instance du modèle de seuil linéaire est définie par un ensemble de poids sur chacun des liens du graphe :

$$\forall(n_i,n_j)\in\mathcal{E}, w_{i,j}\in[0,1]$$

La somme des poids des liens entrant vers un utilisateur doit être inférieure à 1. On choisit ensuite aléatoirement un seuil pour chacun des utilisateurs du réseau. Un utilisateur inactif à l'étape t s'active à l'étape $t+1$ si l'influence de ses voisins devenus actifs au temps t fait passer l'influence globale de ses voisins au dessus de son seuil. Au final, un seul lien déclenche le fait que l'utilisateur s'active.

Le processus de choix des liens pour la percolation de lien est le suivant : comme les seuils sont choisis aléatoirement, pour chaque utilisateur n_j, on choisit au plus un lien entrant (n_i,n_j) avec une probabilité $w_{i,j}$ et aucun lien avec une probabilité $1-\sum_{n_i\in B(n_j)} w_{i,j}$. Les liens choisis son ouverts et les autres restent fermés dans l'instance de la percolation de liens. Ainsi, la probabilité de cette instance dépend de la probabilité de choisir ces liens :

$$q(r) = \prod_{n_j=1}^{N} \prod_{n_i\in B(n_j)} \left((w_{i,j})^{r_{i,j}} \left(1 - \sum_{n_i\in B(n_j)} w_{i,j}\right)^{\left(1-\sum\limits_{n_i\in B(n_j)} r_{i,j}\right)} \right)$$

Le paramètre de changement de phase est l'ensemble des poids associés aux liens du graphe. Il existe un seuil critique au delà duquel tous les utilisateurs atteignables depuis les initiateurs seront actifs à la fin d'une diffusion. Il s'agit d'un chemin partant des initiateurs et allant vers tous les utilisateurs atteignables avec des poids de 1 sur les liens parcourus. Les autres liens auront donc des poids de 0.

1.3 Modèles généralisés

Dans le modèle standard à cascades indépendantes IC, un utilisateur n_i qui tente d'en influencer un autre n_j ne tient pas compte de tout ce qui a pu se passer par le passé. Qu'il soit le premier à tenter d'influencer n_j ou que trois autres personnes aient tenté de le faire avant lui ne change rien à la probabilité que n_j aura de se faire contaminer. D'autre part, le modèle de seuil linéaire LT ne permet de modéliser la fonction de seuil que de manière linéaire. En d'autres mots, chaque voisin entrant de l'utilisateur n_j l'influencera avec la même force, quels que soient les autres utilisateurs impliqués dans la diffusion.

Afin de pallier ce problème et de permettre de modéliser une dépendance entre les voisins entrant d'un utilisateur lors de son choix, l'auteur de [Kleinberg, 2007] en propose deux extensions : le modèle à cascades généralisé et le modèle de seuil généralisé.

1.3.1 Modèle à cascades généralisé

Comme pour les modèles IC, un ensemble de nœuds initialement actifs est désigné. A chaque étape de temps, t, tous les utilisateurs qui sont devenus actifs à l'étape de temps $t-1$ tentent d'influencer leurs voisins sortant afin que ceux-ci deviennent actifs à leur tour. Contrairement au modèle IC, lorsqu'un utilisateur n_i tente d'influencer l'un de ses voisins n_j, il ne le fait plus avec la probabilité $p_{i,j}$ mais avec la probabilité $p_i(j, X)$ qui ne dépend plus seulement des utilisateurs n_i et n_j, X étant l'ensemble des voisins entrant de n_j qui ont déjà tenté de l'influencer, et échoué, par le passé. L'ensemble X ne contient pas l'utilisateur n_i. Il est ainsi possible de définir des fonctions de probabilités qui prennent en compte les influences passées de plusieurs manières. Les deux principales sont :

- les utilisateurs passés peuvent influencer positivement la diffusion. En d'autres mots, plus un utilisateur aura été influencé par ses voisins, plus la probabilité qu'il diffuse à son tour sera forte. Si un utilisateur n_i tente d'influencer un autre utilisateur n_j en étant le premier à le faire, il aura une probabilité plus faible de réussir que s'il est le cinquième. La probabilité de diffusion est incrémentale :

$$\forall n_i, n_j, \forall S \subseteq T, n_i \notin S, \quad p_j(i, S) \leq p_j(i, T)$$

- au contraire, la probabilité de diffusion peut être décrémentale, auquel cas elle diminue plus le nombre de tentatives d'influence d'un utilisateur augmente. Cela permet de modéliser la lassitude d'un utilisateur. A force de voir le même contenu, il finit par ne plus y faire attention.

Enfin, le modèle IC est inclus dans le modèle à cascades généralisé. Il est possible de définir la probabilité de diffusion entre les utilisateurs n_i et n_j de façon à ce qu'elle ne dépende que de ces utilisateurs :

$$\forall n_i, n_j, \forall X, n_i \notin X, \quad p_j(i, S) = p_{i,j}$$

1.3.2 Modèle de seuil généralisé

Ce modèle n'est pas dirigé comme le modèle de seuil linéaire par les poids sur les liens du graphe. A chaque utilisateur n_j est associé une fonction g_j définie sur un sous-ensemble des voisins entrant de n_j, $B(n_j)$. Pour tout sous-ensemble $X \subseteq B(n_j)$, la fonction $g_j(X)$ a une valeur dans $[0, 1]$. La fonction g_j est croissante est vérifie donc que si $X \subseteq Y$, alors $g_j(X) \leq g_j(Y)$. Malgré le fait qu'elle n'est pas obligatoirement la somme des poids des voisins entrant actifs, elle reste une fonction de seuil. Chaque utilisateur n_j choisit aléatoirement un seuil θ_j. Un ensemble de diffuseurs initiaux est défini puis à chaque étape t, un utilisateur n_j devient actif si la valeur de sa fonction de seuil est supérieure à son seuil :

$$g_j(X) \geq \theta_j$$

Comme pour le modèle à cascades généralisé, ce modèle étant une généralisation du modèle qu'il étend, le modèle de seuil linéaire LT, il est possible d'en définir une instance équivalente à un modèle linéaire :

$$g_j(X) = \sum_{n_i \in X} w_{i,j}$$

1.3.3 Equivalence entre les modèles généralisés

Les deux modèles généralisés que nous avons présentés ont une forme et une exécution différentes. Néanmoins dans le cas où les probabilités du modèle à cascades sont indépendantes de l'ordre dans lequel les voisins entrant d'un utilisateur tentent de l'influencer, on peut prouver qu'une instance du modèle à cascades généralisé est équivalente à une instance du modèle de seuil généralisé [Kempe et al., 2005].

Ecrire une instance du modèle à cascades généralisé à partir d'une instance du modèle de seuil généralisé

Nous connaissons pour l'ensemble des utilisateurs les fonctions de seuil qui leur sont associées g_j. Si n_j n'est pas actif, c'est que l'ensemble de ses voisins entrant qui sont déjà actifs X n'ont pas réussi à l'influencer. Autrement dit, sa fonction de seuil a une valeur plus faible que son seuil : $g_j(X) < \theta_j \leq 1$. Sachant que le seuil θ_j est choisi aléatoirement, la probabilité qu'un nouveau voisin diffuseur n_i réussisse à influencer n_j est la probabilité que $g_j(X \cup \{i\}) \geq \theta_j$. Cette probabilité est uniformément répartie et s'écrit :

$$p_j(i, X) = \frac{g_j(X \cup \{i\}) - g_j(X)}{1 - g_j(X)}$$

Ecrire une instance du modèle de seuil généralisé à partir d'une instance du modèle à cascades généralisé

Nous connaissons pour l'ensemble des couples d'utilisateurs (n_i, n_j) les probabilités de transition qui leurs sont associées : $p_j(i, X)$. On choisit la valeur de seuil de chaque utilisateur aléatoirement. La fonction de seuil $g_j(X)$ correspond donc à la probabilité que l'ensemble des voisins entrant actifs X de n_j aient réussi à l'influencer pour qu'il s'active. Cette probabilité vaut $1 - \overline{p}$ où $\overline{p}$ est la probabilité qu'aucun voisin de n_j n'ait réussi à l'influencer. En écrivant $X = \{n_1, n_2, ..., n_k\}$, on peut définir la probabilité d'activation de la façon suivante :

$$g_j(X) = 1 - \prod_{i=1}^{k}(1 - p_j(i, X_{i-1}))$$

où $X_{i-1} = \{n_1, ..., n_{i-1}\}$. La notion d'indépendance dans l'ordre des tentatives d'influence de la part des voisins pour le calcul de la probabilité de diffusion est importante ici.

1.4 Intégration du temps

Les modèles standards, autant les versions simples que les versions généralisées, considèrent la diffusion étape de temps après étape de temps. Pour ces modèles, un utilisateur qui diffuse un contenu au temps t ne peut influencer ses voisins pour qu'ils rediffusent qu'à l'étape $t+1$. Afin de remédier à ce problème, un certain nombre d'adaptations de ces modèles ont été proposées dans lesquelles est inclus un phénomène de latence dans le choix de la diffusion. Une question se pose cependant : un utilisateur a-t-il autant de chance de diffuser un contenu le lendemain de sa création et trois ans plus tard ? Des études sur les réseaux d'e-mails et d'envois de lettres dans l'article [Vázquez et al., 2006] ainsi que sur un réseau de messagerie instantanée dans le rapport [Leskovec and Horvitz, 2006] montrent que, d'une manière générale, les utilisateurs ont tendance à rediffuser peu de temps après la première diffusion. Ce phénomène n'empêche pas quelques utilisateurs de prendre connaissance du contenu après une longue période, mais ces utilisateurs sont très peu nombreux. Afin de prendre en compte ce phénomène, la plupart des modèles qui incluent une latence dans la diffusion, incluent aussi un paramètre de décroissance diminuant la probabilité de diffuser plus le temps passe.

1.4.1 Adaptation des modèles épidémiques

Ajout d'un paramètre de délai

L'article [Liben-Nowell and Kleinberg, 2008] montre une étude des chaînes de diffusion dans un réseau d'e-mails. Les auteurs proposent un modèle simple dans lequel chaque utilisateur qui reçoit un mail à une étape t_0 a une probabilité f de le rediffuser. Chaque

utilisateur qui choisit de rediffuser un mail le fait à une étape $t = t_0 + \tau$, où τ est distribué selon une fonction de puissance. La probabilité pour un utilisateur de rediffuser un mail au temps t lorsqu'il l'a vu au temps t_0 est donc :

$$P(t_0 + \tau) = f\tau^{-\alpha}$$

Une adaptation similaire est présente dans l'article [Bailey et al., 2004] dans lequel les auteurs modélisent la diffusion d'une maladie au sein d'une population de plantes. Ils utilisent une loi exponentielle pour modéliser le phénomène de décroissance de la probabilité de diffuser. Ainsi une plante en contact avec une autre plante contaminée au temps t_0 a la probabilité suivante d'être contaminée au temps $t > t_0$:

$$P(t_0 + \tau) = fe^{\tau}$$

Ces deux manières de représenter le délai ont été reprises dans l'article [Gomez-Rodriguez et al., 2010] dans lequel les auteurs modélisent la communication entre bloggeurs. Ils utilisent leur modèle afin d'inférer les liens entre utilisateurs au sein d'un réseau social.

Seuil cumulatif sur les étapes

Dans les modèles épidémiques, les utilisateurs en contact avec un autre utilisateur contaminé ont une chance d'être contaminés à leur tour. Dans l'article [Dodds and Watts, 2004] les auteurs incluent un seuil de contamination. Ils considèrent qu'à chaque étape de temps où un utilisateur n_i est exposé au virus, il a une probabilité f de recevoir une *dose* positive du virus $d_i(t)$. S'il n'est pas exposé au virus, $d_i(t) = 0$. Chaque individu cumule les doses qu'il a reçu lors des T dernières étapes de temps. On dénote le nombre de doses qu'il a reçues par :

$$D_i(t) = \sum_{t'=t-T+1}^{t} d_i(t')$$

A l'étape de temps t, l'utilisateur n_i est contaminé si la quantité de doses qu'il a reçues est supérieure à son seuil de résistance : $D_i(t) \geq d_i^*$ où d_i^* est le seuil de l'individu n_i qui est choisi aléatoirement avant la diffusion.

Couplage avec un modèle à cascades : Connie

Les modèles standards de la famille SI ne permettent pas de modéliser le délai dans la diffusion. Les auteurs de l'article [Myers and Leskovec, 2010] proposent une adaptation de ces modèles dans lesquels ils ajoutent deux distributions exprimant le temps que met un utilisateur à devenir *Infectieux* et le temps qu'il met à devenir *Rétabli* (modèle SIR) ou de nouveau *Susceptible* (modèle SIS). Contrairement aux modèles standards, dans le modèle qu'ils proposent, un individu n_j peut être contaminé par un de ses voisins n_i avec une

probabilité dépendant de ces deux utilisateurs. Dans un sens, ce modèle prend en compte une influence différente en fonction de la source, comme c'est le cas pour le modèle à cascades indépendantes IC. Lorsque l'utilisateur n_i est contaminé à l'étape t_0, il a une probabilité $p_{i,j}$ de contaminer chacun de ses voisins n_j *Susceptibles*. S'il réussit alors n_j sera contaminé à l'étape $t > t_0$ avec l'intervalle $\delta_j = t - t_0$ qui suit une distribution $\delta(\tau)$. Un utilisateur qui est contaminé le reste pour un certain temps avant de redevenir sain (sans possibilité d'être contaminé à nouveau s'il s'agit du modèle SIR). L'intervalle de temps pendant lequel un utilisateur reste contaminé, s_j, suit une distribution $s(\tau)$.

Dans cet article, les auteurs se focalisent sur des réseaux de partage de contenu et utilisent donc la variante du modèle SI tel que présenté ci-dessus ($\forall n_j, \quad s_j = \infty$). Ils utilisent ce modèle pour expliquer la diffusion au sein d'une communauté et trouver le réseau sous-jacent qui explique le mieux les diffusions observées. Les paramètres $p_{i,j}$ et la distribution $\delta(\tau)$ sont appris pour maximiser la vraisemblance avec les données d'entraînement. La distribution $\delta(\tau)$ est libre et ne suit aucune loi.

1.4.2 Adaptation des modèles de seuil

Délai avec décroissance exponentielle : ASLT

Dans la définition du modèle de seuil linéaire, un utilisateur n_j devient actif aussitôt qu'un de ses voisins s'active en rendant l'influence totale des voisins de n_j supérieure à son seuil θ_j. Les auteurs de l'article [Saito et al., 2010a] proposent deux méthodes pour appliquer un délai dans la diffusion pour ce modèle.

La première méthode est appelée **délai de lien**. Lorsqu'un utilisateur n_i devient actif à l'étape t_0, il applique son influence $w_{i,j}$ à chacun de ses voisins n_j après un délai $\delta_{i,j}$. Ce délai suit une distribution exponentielle de paramètre $r_{i,j}$ afin de prendre en compte le phénomène de décroissance dont nous avons parlé précédemment. Le paramètre $r_{i,j}$ est associé aux utilisateurs n_i et n_j :

$$P(\delta_{i,j} = \tau) = r_{i,j} e^{-r_{i,j}\tau}$$

où $P(\delta_{i,j} = \tau)$ est la probabilité que le délai $\delta_{i,j}$ soit égal à τ. Dans cette adaptation du modèle LT, c'est la source d'une diffusion qui apporte un délai lors de la propagation et ce délai peut être différent pour chaque couple d'utilisateurs.

Les auteurs proposent une seconde modélisation de ce délai du coté de l'utilisateur destination de la diffusion : **le délai de nœud**. Cela revient à dire que le délai dépend du temps que l'utilisateur cible va mettre à regarder et rediffuser le contenu. De plus, le nombre de paramètres de délai est grandement diminué puisqu'il y en a un par utilisateur et non un par lien du réseau. Lorsqu'un utilisateur n_i devient actif, il influence instantanément ses voisins n_j. Par contre, lorsqu'un utilisateur n_j subit plus d'influence que son seuil θ_j, il attend un certain temps avant de diffuser à son tour. Ce délai δ_j suit une distribution

de paramètre r_j :

$$P(\delta_j = \tau) = r_j e^{-r_j \tau}$$

1.4.3 Adaptation des modèles à cascades

Prise en compte de la probabilité de lecture d'un contenu

Dans tous les modèles présentés jusque là, le fait de voir un contenu et de le rediffuser sont couplés dans l'action de diffuser. Dans l'article [Gruhl and Liben-nowell, 2004], les auteurs proposent de modéliser la lecture du contenu par les utilisateurs avant de le diffuser. Un utilisateur ne peut lire un contenu que s'il a été diffusé par un de ses voisins. Si un utilisateur n_i diffuse un contenu au temps t_0, à chaque étape de temps qui suit, tous ses voisins n_j qui n'ont pas encore lu le contenu ont une probabilité $r_{i,j}$ de le lire. Ainsi, la probabilité qu'un utilisateur n_j lise le contenu diffusé par n_i à une étape donnée $t = t_0 + \tau$ est la suivante :

$$R_{i,j}(t_0 + \tau) = r_{i,j}(1 - r_{i,j})^{(\tau - 1)}$$

Cela correspond à la probabilité de ne pas l'avoir lu pendant $\tau - 1$ étapes de temps pour enfin le lire à l'étape $t_0 + \tau$. Lorsqu'un utilisateur a lu le contenu, il a une probabilité de le diffuser $p_{i,j}$ puis reste passif lors de la suite de la diffusion. S'il a diffusé le contenu, il donne la possibilité à ses voisins de le lire. Sinon il aura la possibilité de diffuser le contenu seulement si un autre de ses voisins le diffuse à son tour, mais l'utilisateur n_i ne l'influencera plus par la suite.

Délai avec décroissance exponentielle : ASIC

Comme pour le modèle de seuil linéaire pour lequel l'extension ASLT a été proposée, les mêmes auteurs ont proposé dans l'article [Saito et al., 2009] une extension ASIC pour le modèle à cascades indépendantes. Lorsqu'un utilisateur n_i devient actif au temps t_0, il a une probabilité $p_{i,j}$ d'influencer chacun de ses voisins n_j pour qu'il diffuse à son tour le contenu. Contrairement au modèle IC, si n_j décide de diffuser le contenu, il le fera avec un délai qui suit une loi exponentielle de paramètre $r_{i,j}$. L'utilisateur n_j a ainsi une probabilité de diffuser le contenu à l'étape de temps $t = t_0 + \tau$ en étant influencé par n_i qui dépend de ces deux utilisateurs :

$$P_{i,j}(t_0 + \tau) = p_{i,j} r_{i,j} e^{-r_{i,j} \tau}$$

Les auteurs estiment les paramètres $p_{i,j}$ et $r_{i,j}$ de ce modèle en utilisant un algorithme EM pour maximiser la vraisemblance entre les prédictions du modèle et des données réelles. Des variantes de ce modèle sont proposées dans l'article [Saito et al., 2010a]. On peut trouver une autre variante de ce modèle dans [Koide et al., 2011] dans laquelle les paramètres ne sont plus liés aux utilisateurs, ils sont globaux.

Probabilités de diffusion à partir d'heuristiques

Les auteurs de l'article [Goyal et al., 2010] proposent tout d'abord un modèle de diffusion proche de celui du modèle à cascades indépendantes IC pour lequel ils estiment les paramètres $p_{i,j}$ de diffusion entre les utilisateurs n_i et n_j à partir d'heuristiques sur les informations qu'ils ont des diffusions passées entre ces utilisateurs. Afin de simplifier les explications suivantes, nous introduisons de nouvelles notations :

- A_i est le nombre de contenus que l'utilisateur n_i a diffusé ;
- A_{i2j} est le nombre de contenus qui se sont propagés de l'utilisateur n_i vers l'utilisateur n_j ;
- $A_{i|j}$ est le nombre de contenus qu'au moins un des deux utilisateurs n_i ou n_j a diffusé.

Les auteurs proposent trois méthodes pour définir les probabilités de diffusion :

- **Distribution de Bernoulli.** Dans ce modèle, chaque utilisateur n_i qui tente d'influencer l'un de ses voisins n_j a une probabilité fixe de réussir. Chaque essai correspondant à un essai suivant la loi de Bernoulli. Afin de maximiser la vraisemblance avec les données d'entraînement, la valeur du paramètre $p_{i,j}$ correspond au ratio du nombre de contenus diffusés par n_i et rediffusés par n_j sur le nombre de contenu diffusés par n_i.
$$p_{i,j} = \frac{A_{i2j}}{A_i}$$
- **Coefficient de Jaccard.** Au lieu de comparer le nombre de contenus qui se sont propagés avec le nombre de contenus émis par l'utilisateur source, les auteurs proposent de comparer le nombre de contenus qui se sont propagés par le lien avec le nombre de diffusions dans lesquelles ont participé indifféremment les deux utilisateurs.
$$p_{i,j} = \frac{A_{i2j}}{A_{i|j}}$$
- **Crédit partiel.** Lors d'une diffusion, un utilisateur peut être influencé par plusieurs de ses voisins entrant. Les auteurs proposent que le crédit de l'influence réalisée par ses voisins lors de la diffusion d'un contenu par un utilisateur soit réparti équitablement entre eux. Ils définissent ainsi le crédit de l'influence d'un utilisateur n_i sur un utilisateur n_j pour la diffusion d'un contenu c^k :
$$credit_{i,j}(k) = \frac{1}{|C^k(t_j(k)-1) \cap B(j)|}$$
où $t_j(k)$ est le temps auquel l'utilisateur n_j a diffusé le contenu c^k et $C^k(t)$ est l'ensemble des utilisateurs ayant diffusé le contenu c^k au temps t ou avant. Ils proposent ensuite deux manières de prendre en compte ce crédit qui sont les deux méthodes

précédemment présentées :

$$p_{i,j} = \frac{\sum_{k=1}^{K} credit_{i,j}(k)}{A_i} \quad ou \quad p_{i,j} = \frac{\sum_{k=1}^{K} credit_{i,j}(k)}{A_{i|j}} \tag{1.1}$$

Dans un second temps, les auteurs adaptent leur modèle pour une prise en compte d'un délai de diffusion. Comme pour le modèle ASIC, ils définissent un délai dans la diffusion qui suit une distribution de paramètre $r_{i,j}$. Ainsi la probabilité que l'utilisateur n_j diffuse un contenu provenant de l'utilisateur n_i τ étapes de temps après que n_i l'ait diffusé est :

$$P_{i,j}(t_0 + \tau) = p_{i,j} e^{-\tau / r_{i,j}}$$

où t_0 est le temps auquel n_i a diffusé le contenu. Le paramètre $r_{i,j}$ correspond au temps moyen que n_j met à rediffuser un contenu provenant de n_i.

Plusieurs modélisations du délai : Netrate

Dans l'article [Gomez-Rodriguez et al., 2011] les auteurs proposent une méthode de prédiction qui cherche à calculer le temps que vont mettre les utilisateurs à rediffuser un contenu. Si un utilisateur ne rediffuse pas un contenu, on considère qu'il a mis un temps infini à le rediffuser. Comme pour le modèle à cascades indépendantes IC, lorsqu'un utilisateur n_i devient actif au temps t_0, il tente d'influencer chacun de ses voisins n_j. On définit par $\delta_{i,j}$ le temps que va mettre l'utilisateur n_j qui n'est pas déjà actif à diffuser le contenu. Comme dit précédemment, si $\delta_{i,j} = \infty$ on considère que l'utilisateur n_i n'a pas réussi à influencer l'utilisateur n_j pour qu'il diffuse le contenu. Trois modèles basés sur cette idée sont proposés pour définir le délai de diffusion. Chacun de ces modèles caractérise $P_{i,j}(t)$ qui est la probabilité que l'utilisateur n_i influence l'utilisateur n_j pour que celui-ci diffuse le contenu à l'étape $t = t_0 + \tau$:

– un modèle exponentiel défini comme suit :

$$P_{i,j}(t_0 + \tau) = r_{i,j} e^{-r_{i,j}\tau}$$

Ce modèle est un cas particulier du modèle ASIC présenté dans la section 1.4.3 qui ne prend pas en compte la probabilité de diffuser ;

– un modèle suivant une loi de puissance :

$$P_{i,j}(t_0 + \tau) = \frac{r_{i,j}}{\theta} \left(\frac{\tau}{\theta}\right)^{(-1-r_{i,j})}$$

– un modèle de Rayleigh :

$$P_{i,j}(t_0 + \tau) = r_{i,j}\tau e^{-\frac{1}{2}r_{i,j}\tau^2}$$

Les auteurs estiment les paramètres $r_{i,j}$ de leurs modèles en utilisant une méthode de descente de gradient pour maximiser la vraisemblance avec les données d'entraînement. Le choix de leurs trois modèles n'est pas un hasard. Les distributions pour le délai ainsi définies entraînent un problème d'optimisation convexe évitant le problème des maximums locaux.

1.5 Prise en compte de caractéristiques utilisateur

Nous venons de voir une série de modèles qui prennent en compte le temps comme une donnée da la diffusion. Plusieurs modèles ajoutent à ceci un certain nombre de paramètres utilisateurs afin de décrire la diffusion dans un réseau. Nous en présentons deux qui sont des extensions du modèle ASIC.

1.5.1 Similarité entre utilisateurs

Dans un premier temps, nous présentons ici une extension du modèle ASIC (présenté dans la section 1.4.3) proposée dans [Saito et al., 2011] qui prend en compte les attributs des utilisateurs afin de définir les probabilités de diffusion. Chaque utilisateur n_i est représenté par un profil p^i. Ce profil est ici considéré comme étant un vecteur d'attributs de taille F. Ces attributs peuvent être par exemple la ville dans laquelle vit l'utilisateur, son sexe, son âge, ou tout autre caractéristique qu'il a pu renseigner dans le réseau social. L'idée est la suivante : si deux utilisateur ont des profils similaires, c'est qu'ils se ressemblent, partagent les mêmes centres d'intérêt et ont donc de fortes chances de partager le même type de contenus. Des études sur l'homophilie [1] sont présentes dans les articles [Apolloni et al., 2009, Anderson et al., 2012].

Pour chaque couple d'utilisateurs liés entre eux (n_i, n_j), le modèle défini une probabilité de diffusion à travers le lien $p_{i,j}$ ainsi qu'un délai d'activation $r_{i,j}$. Dans le modèle ASIC standard, il y a autant de paramètres que de liens entre les utilisateurs, ce qui pose d'une part le problème du grand nombre de paramètres à estimer, et d'autre part la non prise en compte d'informations concernant les utilisateurs. Ici, la probabilité de diffusion est définie en fonction des profils des utilisateurs :

$$p_{i,j} = \frac{1}{1 + exp\left(-\theta_0 + \sum_{m=1}^{F} -\theta_m f_m(p_m^i, p_m^j)\right)}$$

1. l'homophilie est la tendance des individus à s'associer avec des individus similaires.

où $f_m(p_m^i, p_m^j)$ est une fonction définissant la proximité entre les m$^{\text{ième}}$ caractéristiques des utilisateurs n_i et n_j. Le paramètre θ_0 est un paramètre de biais, et les θ_m sont des paramètres pour réguler l'influence que prennent les différentes caractéristiques dans le calcul de la probabilité. D'une manière générale, la probabilité sera haute si les valeurs de $f_m(p_m^i, p_m^j)$ sont faibles. Dans cette étude, les auteurs ont utilisé une fonction simple pour calculer la proximité entre les caractéristiques utilisateur. Si la m$^{\text{ième}}$ caractéristique est nominale :

$$f_m(p_m^i, p_m^j) = \begin{cases} 1 & \text{si } p_m^i = p_m^j \\ \\ 0 & \text{sinon} \end{cases}$$

Si au contraire elle est numérique :

$$f_m(p_m^i, p_m^j) = exp(-|p_m^i - p_m^j|)$$

Pour chacune de ces deux façons de calculer la similarité entre les utilisateurs, plus les caractéristiques sont proches les unes des autres, plus les utilisateurs sont considérés comme proches.

Le délai est calculé de la même manière que dans le modèle ASIC en utilisant une décroissance exponentielle de paramètre $r_{i,j}$. En outre, le calcul des paramètres $r_{i,j}$ se fait en utilisant les caractéristiques utilisateur définies précédemment :

$$r_{i,j} = exp\left(-\phi_0 + \sum_{m=1}^{F} -\phi_m f_m(p_m^i, p_m^j)\right)$$

Comme ils le font pour les modèles IC et ASIC, les auteurs définissent la vraisemblance et apprennent les valeurs des paramètres θ_m et ϕ_m en maximisant la vraisemblance grâce à un algorithme EM.

Un modéle très similaire basé sur Netrate a été proposé dans [Wang et al., 2012b].

1.5.2 Utilisation de propriétés utilisateur

Une autre extension du modèle ASIC se trouve dans l'article [Guille and Hacid, 2012]. Les auteurs ne considèrent pas seulement la proximité entre les utilisateurs mais définissent un certain nombre de propriétés des utilisateurs pour calculer les probabilités de diffusion. Pour chaque lien entre deux utilisateurs (n_i, n_j) et chaque contenu c^k, ils forment un vecteur de propriétés $V^{i,j,k}$ de taille F, calculé à partir des activités passées des utilisateurs. Les valeurs de ces propriétés, qui varient entre 0 et 1, sont les suivantes :

- l'activité de chacun des utilisateurs correspond au nombre de contenus partagés par heure, borné par 1 ;
- l'homogénéité entre deux utilisateurs est la proportion de leurs voisins qu'ils ont en

commun ;
- le ratio de contenus partagés à des destinataires particuliers. Il s'agit là du nombre de contenus partagés avec seulement un autre utilisateur ou une communauté par rapport au nombre total de contenus partagés. Cette mesure ne peut pas être évaluée dans tous les réseaux sociaux, car certains ne permettent pas de choisir les destinataires lors du partage d'information ;
- l'existence d'une relation sociale entre deux utilisateurs ;
- le taux de citations d'un utilisateur correspond à sa popularité. Plus cette propriété est grande, plus l'utilisateur a été cité dans les contenus d'autres utilisateurs du réseau ;
- l'attrait de l'utilisateur pour le contenu. Il s'agit d'une valeur booléenne qui indique si au moins un mot clé du contenu fait parti du vocabulaire de l'utilisateur ;
- enfin, la prise en compte de la dimension temporelle. Selon l'heure à laquelle un contenu est partagé, il n'aura pas le même impact sur les autres utilisateurs.

Les auteurs utilisent ensuite ce vecteur de propriétés pour définir les probabilités de diffusion d'un manière similaire au modèle que nous avons présenté en section 1.5.1 :

$$p_{i,j} = \frac{1}{1 + exp\left(-\theta_0 + \sum_{m=1}^{F} -\theta_k V_m^{i,j,k}\right)}$$

L'estimation des paramètres se fait ensuite en utilisant les algorithmes de *perceptron linéaire* et de *régression logistique bayésienne* pour résoudre la tâche de classification supervisée suivante : $P(Y|V)$ où Y={diffusion, non diffusion} et V est l'ensemble de propriétés qui ont été définies.

L'apprentissage des paramètres de délai est un peu différent du modèle précédent. En effet, les auteurs ne définissent qu'un seul paramètre à estimer et calculent le délai en fonction de ce paramètre et de l'activité de l'utilisateur cible d'une diffusion :

$$r_{i,j} = (1 - A_j)\phi$$

où A_j est l'activité (telle que définie au dessus) de l'utilisateur n_j. Le paramètre ϕ est calculé selon un algorithme de recherche par grille. L'apprentissage se fait pour un ensemble de valeurs fixées et la meilleure est gardée comme valeur du paramètre ϕ.

L'exécution du modèle se déroule ensuite exactement comme celle du modèle ASIC en utilisant les paramètres $p_{i,j}$ et $r_{i,j}$ ainsi définis.

1.6 Modélisation non discrète de la diffusion

Les modèles présentés jusque-là modélisent tous le processus de diffusion de manière discrète : à chaque étape, les utilisateurs influencent leurs voisins, qui pourront eux-mêmes par la suite influencer leurs voisins. Cette façon de modéliser la diffusion nécessite un certain nombre de connaissances *à priori* sur le réseau social, comme par exemple la topologie du réseau (qui peut communiquer avec qui). Un modèle faisant évoluer la topologie du réseau au cours du temps a été proposé dans [Heaukulani and Ghahramani, 2013]. De plus, on suppose toujours que tous les utilisateurs du réseau sont connus, ce qui n'est pas souvent le cas. Comme exemple très simple observons un réseau social sur Internet, certains profils sont privés et donc inaccessibles. Certains méthodes ont pour but de modéliser la diffusion de manière plus globale afin d'éviter ces travers.

1.6.1 Modélisation de l'état final de la diffusion

Les auteurs de [Najar et al., 2012] proposent un modèle de la diffusion de contenus robuste au manque d'information sur l'ensemble des utilisateurs du réseau. Leur méthode consiste à calculer directement l'état final d'une diffusion en fonction de l'état initial. Pour un contenu c^k et la matrice de diffusion correspondante M^k, ils définissent une fonction de prédiction :

$$f_\theta(\mathcal{G}, M_0^k) \rightarrow M_{T^k}^k$$

où M_t^k est le vecteur de diffusion donnant l'état (diffuseur ou non diffuseur) de tous les utilisateurs au temps t.

Les paramètres de ce modèle peuvent ensuite être appris en comparant pour un ensemble de cascades d'entraînement les valeurs réelles de l'état des utilisateurs avec les valeurs prédites par le modèle. Cette comparaison se fait à l'aide d'une fonction de coût qui détermine la pénalité due à une mauvaise prédiction.

La fonction de prédiction ainsi définie est très générique et les auteurs en proposent quatre instantiations particulières qu'ils utilisent dans leurs expériences :

– Le premier modèle LM (Linear Model) est une simple régression linéaire qui s'exprime, pour un utilisateur n_j, de la façon suivante :

$$f_{\theta,j}^{LM}(\mathcal{G}, M_0^k) = \sum_{n_i \in \mathcal{N}} \theta_{j,i} m_{i,0}^k$$

où $m_{i,0}^k$ est l'état de l'utilisateur n_j au temps 0 pour la cascade c^k. $\theta_{j,i}$ est un paramètre d'influence entre les utilisateurs n_j et n_i qui peut prendre n'importe quelle valeur réelle. Ainsi un utilisateur peut en influencer un autre soit négativement, soit positivement. Cette fonction prédit ainsi un score de diffusion dans $\mathbb{R}$.

– Le second modèle LoM (Logistic Model) est une régression logistique :

$$f_{\theta,j}^{LoM}(\mathcal{G}, M_0^k) = logit(\sum_{n_i \in \mathcal{N}} \theta_{j,i} m_{i,0}^k)$$

où *logit* est une fonction logistique. Par rapport au modèle précédent, la fonction logistique force les valeurs prédites par le modèle dans l'espace $[0, 1]$.

– Le troisième modèle PLM (Positive Linear Model) est une version contrainte du modèle linéaire pour lequel les poids d'influence entre les utilisateurs sont positifs :

$$f_{\theta,j}^{PLM}(\mathcal{G}, M_0^k) = \sum_{n_i \in \mathcal{N}} \theta_{j,i}^2 m_{i,0}^k$$

$\theta_{j,i}$ sont des valeurs réelles mais le poids de l'influence de chaque couple d'utilisateurs est lui positif.

– Enfin, le quatrième modèle GPLM (Graph Based Positive Linear Model) prend en compte la topologie du graphe et ne prédit de diffusion qu'entre les utilisateurs connectés entre eux :

$$f_{\theta,j}^{PLM}(\mathcal{G}, M_0^k) = \sum_{n_i \in \mathcal{N}} w_{j,i} \theta_{j,i}^2 m_{i,0}^k$$

où $w_{j,i}$ est un poids associé au lien entre les utilisateurs n_j et n_i. Dans le cas ou aucun lien n'existe entre ces utilisateurs, alors $w_{j,i} = 0$. Il est ensuite possible de définir des poids en fonction des communications observées entre les utilisateurs.
Un avantage de ce modèle est que le nombre de paramètres $\theta_{j,i}$ à estimer correspond au nombre de liens dans le graphe, contrairement aux autres modèles pour lesquels il faut estimer N^2 paramètres.

1.6.2 Modèle global : modèle linéaire d'influence LIM

Dans l'article [Yang and Leskovec, 2010], les auteurs présentent un modèle qui tend à prédire le volume d'utilisateurs touchés par une diffusion en fonction de l'influence qu'ont un petit nombre d'utilisateurs d'intérêt, qui sont par exemple des journaux ou un ensemble d'utilisateurs très actifs.

Pour chacun des utilisateurs d'intérêt n_i, le modèle définit une influence $I_i(t)$. Il s'agit d'une fonction qui prédit le nombre d'utilisateurs touchés par une diffusion par l'intermédiaire de l'utilisateur n_i t étapes de temps après que celui-ci ait diffusé le contenu. Le modèle donne donc le volume d'utilisateurs diffuseurs à une étape de temps t comme la somme des influences des utilisateurs d'intérêt actifs au temps t :

$$V(t+1) = \sum_{n_i \in \phi(t)} I_i(t - t_i)$$

où $\phi(t)$ est l'ensemble d'utilisateurs d'intérêt actifs au temps t et t_i est l'étape à laquelle l'utilisateur n_i a diffusé le contenu.

Les auteurs proposent deux manières de définir les fonctions d'influence :

- en les faisant suivre des formes paramétriques choisies telles qu'une exponentielle $I_i(t) = c_i e^{-\lambda_i t}$ ou une loi de puissance $I_i(t) = c_i t^{-\alpha_i}$. Le problème de cette méthode est que l'on suppose que tous les utilisateurs ont une influence de forme similaire ;
- en ne faisant aucune hypothèses sur la forme de la fonction d'influence. La fonction d'influence est donc définie sur T étapes de temps et on estime sa valeur pour les T étapes qui suivent la contamination de l'utilisateur n_i.

La seconde méthode est utilisée pour pouvoir définir des fonctions d'influence sans aucune contrainte.

Afin de tenir compte de propriétés supplémentaires sur la diffusion, les auteurs proposent deux variantes de leur modèle. Dans les réseaux sociaux, un utilisateur est plus enclin à rediffuser un contenu récent plutôt qu'un contenu trop vieux, qui est possiblement devenu obsolète. Pour prendre en compte ce paramètre ils ajoutent un paramètre d'influence générale qu'une information a en fonction du temps :

$$V(t+1) = \alpha(t) \sum_{n_i \in \phi(t)} I_i(t - t_i)$$

où $\alpha(t)$ est ce nouveau paramètre qui modélise la nouveauté d'un contenu qui est le même pour tous les utilisateurs.

La seconde variante a pour but de prendre en compte un autre aspect de la diffusion de contenus : l'effet d'imitation. Cet effet intervient quand un utilisateur en imite un autre en diffusant un contenu juste parce qu'il est populaire (pas obligatoirement au sein du réseau social). Ils modélisent cette quantité comme un volume latent qui n'est pas relié à l'influence des utilisateurs du réseau :

$$V(t+1) = b(t) + \sum_{n_i \in \phi(t)} I_i(t - t_i)$$

où $b(t)$ est le paramètre de volume latent qui s'ajoute à l'influence que les utilisateurs d'intérêt ont induit.

Cette étude n'est pas la seule, et d'autres ont cherché à résoudre le problème de définir un volume de diffusion plutôt que de modéliser chaque utilisateur en utilisant des modèles simples [Petrovic et al., 2011, Hong et al., 2011].

1.7 Synthèse

Nous avons présenté dans ce chapitre l'ensemble des modèles les plus communs pour la modélisation de la diffusion. Certains ont été créé dans le but de prédire la diffusion

de virus ou maladies (SI), alors que d'autres ont été pensé pour modéliser la propagation d'innovations et de contenus au sein d'une population (IC et LT). Les modèles les plus simples sont basés sur l'idée que les individus s'influencent les uns les autres : pour la diffusion de contenus, le fait qu'un utilisateur décide ou non de rediffuser une information qu'il a reçu dépend fortement de ses voisins qui la lui ont transmis. Tous ces modèles sont équivalents à des percolations de liens et sont donc similaires dans leur expression. Nous avons présenté un certain nombre de variantes de ces modèles :

- des versions généralisées dans lesquelles les voisins qui ont influencés un utilisateur par le passé continuent de le faire ;
- des versions qui intègrent la temporalité, ce qui permet aux utilisateurs de diffuser un contenus plusieurs étapes de temps après l'avoir reçu et pas seulement à l'étape suivante.

Enfin, nous avons présenté des modèles qui mettent en avant des idées différentes. Les modèles précédents ne prennent en compte que la topologie du réseau pour expliquer la diffusion. Certains modèles ajoutent la prise en compte de caractéristiques utilisateurs afin d'expliquer la diffusion. D'autres modèles, contrairement à tous ceux présentés cherchent à prédire un état final de la diffusion plutôt que d'expliquer le phénomènes étape par étape.

Il existe un certain nombre d'autres idées mises en avant pour des modèles encore ici basés sur les trois modèles principaux SI, IC et LT. Certains ont réutilisé des notions de physique comme la diffusion de la chaleur [Ma et al., 2008] ou la théorie des champs moyens [Nekovee et al., 2007]. Du au manque de diffusion dans les données réelles, [Wang et al., 2012a] groupent les utilisateurs en fonction de leur distance dans le graphe au point d'origine de la diffusion. Chaque modèle a ses forces et ses faiblesses et ne permettent pas tous de prédire toutes les diffusion. C'est en suivant cette idée que [Saito et al., 2010b] proposent un algorithme qui choisit entre plusieurs modèles celui qui prédit le mieux en fonction du contexte. Enfin, plusieurs articles se sont penchés sur le problème de multiples diffusions en concurrence [Pathak et al., 2010, Budak et al., 2011, Myers and Leskovec, 2012, Beutel et al., 2012, Wei et al., 2012]. Ce problème est aussi bien traité pour la diffusion de contenus que la propagation de virus.

Malgré le nombre de modèles proposés pour la tâche de diffusion, la plupart de ces modèles ont cependant de grosses lacunes :

- ils ne prennent pas en compte le contenu diffusé ;
- très peu prennent en compte les goûts, les intérêts des utilisateurs ;
- ils reposent presque intégralement sur la topologie du réseau social qui, même si il a été montré qu'elle a un rôle dans la diffusion [Bakshy et al., 2012, Chierichetti et al., 2010, Zaman et al., 2010], n'est souvent pas ou mal connue dans un cadre réel.

Chapitre 2

Diffusion de contenu : une nouvelle modélisation probabiliste

Sommaire

2.1 Idée principale : Utilisation de caractéristiques utilisateur

La plupart des modèles que nous avons présentés dans le chapitre précédent expliquent la diffusion de l'information au sein d'un réseau social en tenant uniquement compte de la topologie du réseau : un utilisateur aura tendance à rediffuser une information s'il est proche dans le réseau d'un grand nombre de diffuseurs. Ces modèles passent outre un certain nombre d'autres éléments qui jouent un rôle dans la prise de décision des utilisateurs quant à la diffusion ou pas d'un contenu. Par exemple, il est montré dans [Suh et al., 2010] que la diffusion sur Twitter est très dépendante du contenu, des *hashtags* et des *URLs* incluses dans les tweets. Nous proposons ici l'utilisation de trois caractéristiques propres aux utilisateurs ayant une influence sur la décision de rediffuser une information qu'ils ont vue :

- l'intérêt thématique de l'utilisateur pour le contenu
- l'activité de l'utilisateur (son rôle passif ou actif)
- la pression sociale subit par l'utilisateur

2.1.1 L'Intérêt thématique

Le sujet d'un contenu à un rôle dans la propagation d'un contenu [Cha et al., 2009]. De plus, de manière assez évidente, un utilisateur sera plus intéressé par un contenu faisant parti de ses centres d'intérêt et aura donc plus tendance à le lire puis le rediffuser. Nous représentons l'intérêt thématique d'un utilisateur pour un contenu comme la proximité entre son profil et le dit contenu. Le profil d'un utilisateur n_i et la description d'un contenu c^k sont définis dans le même espace : $p^i \in \mathbb{R}^F$ et , $c^k \in \mathbb{R}^F$.

La proximité est alors définie de la forme suivante :

$$S(n_i, \mathcal{P}, c^k, \theta_s) = sim(p^i, c^k) - \theta_s$$

où $sim(p^i, c^k)$ est une mesure de similarité entre le profil de l'utilisateur et la description du contenu. Nous avons utilisé dans cette thèse la similarité cosinus, mais on peut très bien imaginer utiliser une autre mesure de similarité. θ_s est un seuil permettant de définir si les deux objets sont suffisamment proches. Intuitivement, si la similarité est plus grande que le seuil, la proximité sera positive. Elle sera négative dans le cas inverse. Nous pourrons par la suite utiliser cette proximité pour qu'elle joue de manière positive ou négative sur la probabilité d'un utilisateur de diffuser un contenu. En d'autres mots, un utilisateur aura une forte probabilité de diffuser un contenu proche de ses centres d'intérêt et une faible probabilité si le contenu ne l'intéresse pas.

2.1.2 L'Activité

Cette caractéristique représente le caractère actif ou passif d'un utilisateur. Elle peut être directement calculée à partir des observations faites sur chaque utilisateur (sur le jeu de données d'entraînement [1]). Elle correspond au ratio entre le nombre de contenus qu'un utilisateur a vu et repartagé et le nombre de contenus qu'il a vu et n'a pas repartagé :

$$Act(n_i, \mathcal{G}, \mathcal{D}) = \frac{\sum_{k=1}^{l} I(|\mathcal{C}^k(n_i, T^k - 1)| > 0) m^k_{i,T^k}}{\sum_{k=1}^{l} I(|\mathcal{C}^k(n_i, T^k - 1)| > 0)}$$

où $I(X)$ est la fonction indicateur qui vaut 1 si X est vrai et 0 sinon. Nous introduisons un seuil θ_w similaire au seuil de similarité θ_s :

$$W(n_i, \mathcal{G}, \mathcal{D}, \theta_w) = Act(n_i, \mathcal{G}, \mathcal{D}) - \theta_w$$

Comme la précédente, cette caractéristique utilisateur sera donc positive si elle doit améliorer la probabilité de l'utilisateur de diffuser, nulle si elle ne doit pas l'influencer et négative sinon.

Dans [Romero et al., 2011a], une mesure similaire est utilisée afin de déterminer l'activité des utilisateurs et leur influence sur les autres lors de la diffusion d'un contenu.

2.1.3 La Pression sociale

Cette dernière caractéristique utilisateur représente le fait que la confiance qu'un utilisateur a sur l'intérêt d'une information est croissante avec le nombre de sources différentes. Si une personne nous parle d'un événement, on pourra ne pas y faire attention, alors que si dix personnes nous parlent du même événement on sera forcément au courant et on voudra le repartager. Selon le modèle choisi, la pression sociale est représentée de manière différente. Dans le modèle IC par exemple elle est présente par le fait que si l'activation d'un utilisateur a échoué, un autre essai pourra être effectué plus tard par un autre voisin entrant.

Nous avons modélisé cette pression sociale plus explicitement pour tenir compte du fait qu'une information partagée par deux voisins peut avoir beaucoup plus d'impact qu'une information partagée par un seul voisin. Voici la mesure associée :

$$SP(n_i, \mathcal{G}, M^k, t)$$

Dans l'idée générale elle correspond au nombre de sources (voisins entrant d'un utilisateur) ayant diffusé un contenu. Nous verrons par la suite que ce n'est pas toujours aussi simple car il n'est pas toujours possible de connaître ce nombre.

1. il s'agit de l'ensemble des données qui vont être utilisées pour estimer les valeurs des paramètres des modèles

2.1.4 Combinaison des caractéristiques utilisateur

Nous avons définit un certain nombre de caractéristiques à utiliser afin de prédire si un utilisateur va, ou non, diffuser un contenu. Chaque utilisateur est donc représenté par un vecteur qui évolue au cours du temps pour chaque contenu :

$$\Phi^{n_i,c^k,t} = \begin{pmatrix} S(n_i, \mathcal{P}, c^k, \theta_s) \\ W(n_i, \mathcal{G}, \mathcal{D}, \theta_w) \\ SP(n_i, \mathcal{G}, M^k, t) \end{pmatrix}$$

Il est possible et facile de définir de nouvelles caractéristiques pour les intégrer dans les modèles que nous présentons ici. Nous avons sélectionné ces trois caractéristiques car elles nous paraissent, au vu de la littérature sur l'analyse des réseaux sociaux les plus importantes.

En utilisant ces caractéristiques utilisateur, il est possible de définir des fonctions d'agrégation pour chaque utilisateur, contenu et étape de temps qui serviront à construire les fonctions de probabilité de diffusion. Nous avons opté dans cette thèse pour une simple combinaison linéaire des caractéristiques utilisateur due à leur indépendance :

$$f_\lambda(n_i, c^k, t) = \lambda_0 + \lambda_1 \Phi_1^{n_i,c^k,t} + \lambda_2 \Phi_2^{n_i,c^k,t} + \lambda_3 \Phi_3^{n_i,c^k,t} \tag{2.1}$$

Nous omettons pour des raisons de lisibilité les autres arguments $(\mathcal{P}, c^k, \mathcal{G}, M^k_{.,T^k-1}, \theta_s, \theta_w)$. Les paramètres λ_0, λ_1, λ_2 et λ_3 contrôlent l'influence de chaque dimension de la diffusion. Ces paramètres sont globaux (les mêmes pour tous les utilisateurs). Cela veut dire que l'influence de chaque caractéristique par rapport aux autres sera la même pour tous les utilisateurs. Ce n'est pas pour autant que les valeurs de ces caractéristiques seront les mêmes pour deux utilisateurs différents. Cela a pour effet d'une part de diminuer la complexité de l'apprentissage, et surtout de pouvoir reporter la connaissance que l'on a d'un utilisateur sur les autres. Cette dernière propriété est très intéressante dans un contexte où l'on ne possède que peu d'information ou dans le cas de l'arrivée d'un nouvel utilisateur dans le système.

Nous expliquons dans la section 2.6 comment sont appris ces paramètres ainsi que θ_s et θ_w.

2.2 Probabilité de diffusion d'un utilisateur

De part les définitions précédentes, la probabilité d'un utilisateur de diffuser une information doit être forte quand $f_\lambda(n_i, c^k, t)$ est grand, c.a.d. quand :

- l'intérêt thématique de l'utilisateur pour le contenu est grand
- l'activité de l'utilisateur est grande
- la pression sociale subie par l'utilisateur est forte

Ces contraintes sont naturellement obtenues en utilisant une fonction logistique qui agit comme une fonction de seuil continu. De plus, un utilisateur ne peut rediffuser un contenu que s'il l'a vu, autrement dit si un de ses voisins entrant a déjà partagé le contenu. La probabilité de diffusion d'un contenu c^k par un utilisateur n_i à l'étape de temps t est donc la suivante :

$$P(n_i, c^k, t) = \begin{cases} (1 + e^{-f_\lambda(n_i,t,c^k)})^{-1} \text{ si } SP(n_i, \mathcal{G}, M^k, t) > 0 \\ \\ 0 \text{ sinon} \end{cases} \tag{2.2}$$

avec $\lambda_1 \geq 0$, $\lambda_2 \geq 0$ et $\lambda_3 \geq 0$. Si un paramètre vaut 0, la caractéristique utilisateur correspondante n'aura aucun impact sur la probabilité de diffusion. Si le paramètre est positif, la caractéristique utilisateur correspondante aura un apport positif si elle est positive et négatif si elle est négative.

2.2.1 Remarque sur l'influence des voisins

Dans la plupart des modèles présentés dans le chapitre 1, l'influence des voisins nouvellement diffuseurs sur la probabilité d'un utilisateur de diffuser un contenu augmente de manière continue. Dans un cas simple où tous les voisins sont considérés comme ayant individuellement la même influence, le fait d'avoir deux voisins actifs donnera à un utilisateur une probabilité de diffuser deux fois plus élevée que s'il n'avait qu'un seul voisin diffuseur. Nous montrons ici que ce n'est pas le cas pour les modèles centrés utilisateur.

La figure 2.1 montre la probabilité de diffuser d'un utilisateur en fonction du nombre de ses voisins ayant déjà diffusé le contenu. Pour simplifier cet exemple nous avons choisi de ne pas tenir compte des paramètres de similarité et de volonté de diffuser en les fixant à 0, mais la forme de la probabilité reste la même pour d'autres choix de paramètres dû à la fonction de seuil que nous avons choisie.

On voit clairement que les premiers voisins diffuseurs influencent beaucoup la probabilité de diffuser d'un utilisateur, le second influence plus que le premier et ainsi de suite jusqu'a atteindre un seuil au delà duquel la probabilité est proche de 1. Dans cet exemple, ce seuil est atteint pour trois voisins actifs.

Selon les valeurs du paramètre de biais, de la volonté de l'utilisateur de diffuser et de la similarité entre son profil et le contenu, la probabilité augmentera plus ou moins vite. En fait, si tous les paramètres autres que la pression sociale sont négatifs, il faudra beaucoup de voisins actifs pour pouvoir compenser et la courbe sera très aplatie. Si en outre ces paramètres sont positifs, un seul voisin suffira pour obtenir une forte probabilité de diffuser.

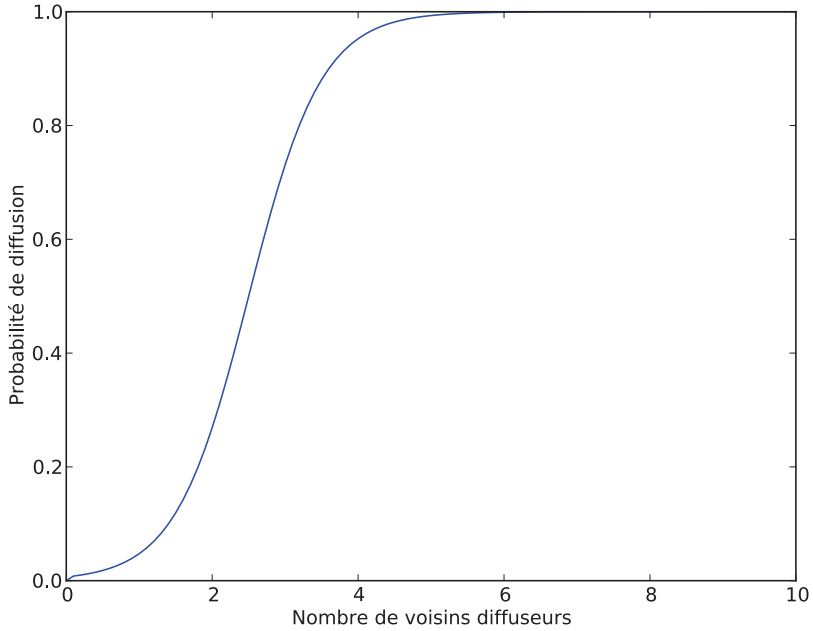


FIGURE 2.1 – Influence des voisins sur la probabilité de diffuser. Nous avons fixé $\lambda_0 = -5$, $\lambda_1 = \lambda_2 = 0$ et $\lambda_3 = 2$.

2.3 Modèle centré utilisateur : UC

Le premier modèle que nous proposons est basé sur les mêmes principes fondamentaux que le modèle à Cascades Indépendantes (IC). Le temps est découpé de manière discrète et à chaque étape t, les utilisateurs qui viennent d'être activés à l'étape précédente $t-1$ tentent d'activer leurs voisins avec une certaine probabilité. La différence avec le modèle IC se trouve dans cette probabilité. Dans le modèle original, cette probabilité est fixe au cours du temps et ne prend en compte au mieux (si l'apprentissages des probabilités est fait pour maximiser la vraisemblance) qu'une seule caractéristique utilisateur : la proportion de messages échangés entre les deux utilisateurs. Dans ce modèle, la pression sociale se caractérise par le nombre de voisins entrant actifs :

$$SP(n_i, \mathcal{G}, M^k, t) = |\mathcal{C}^k(n_i, t)|$$

L'algorithme 1 montre le détail de l'exécution du modèle UC. Ce modèle utilise les caractéristiques que nous avons introduites précédemment, mais garde cependant une grosse lacune : il ne tient compte d'aucune dynamique temporelle de diffusion. En particulier, un utilisateur ne peut s'activer qu'à l'étape qui suit celle où l'un de ses voisins s'est lui-même activé. Nous proposons ensuite le modèle RUC plus axé sur cette dynamique de diffusion.

Algorithme 1: Exécution du modèle UC

Données : Un graphe $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
Un contenu c^k
Un ensemble d'utilisateurs diffuseurs $Initiaux$
Résultat : Ensemble des utilisateurs $Actifs$ après processus de diffusion
$Actifs \leftarrow Initiaux$;
$JusteActifs \leftarrow Initiaux$;
$t \leftarrow 0$;
répéter
  $ProchainsActifs \leftarrow \emptyset$;
  pour tous les $u \in JusteActifs$ **faire**
    pour tous les $v \in Voisinssortant(u)$ **faire**
      si $v \notin Actifs$ **alors**
        si *hasard()* $\leq P(n_v, c^k, t)$ **alors**
          $Actifs \leftarrow Actifs \cup \{\text{v}\}$;
          $ProchainsActifs \leftarrow ProchainsActifs \cup \{\text{v}\}$;
        fin
      fin
    fin
  fin
  JusteActifs ← ProchainsActifs;
  t ← t + 1;
jusqu'à $JusteActifs \neq \emptyset$;

2.4 Intégration du renforcement : RUC

Dans le modèle RUC (Reinforced User-Centric model) nous prenons en compte la dimension temporelle pour la diffusion. Une nouvelle particularité est qu'un utilisateur n'est plus dans un système binaire dans lequel il est soit actif soit inactif : on considère sa probabilité d'être actif (et par conséquent sa probabilité d'être inactif). Le temps est toujours représenté de façon discrète, mais un utilisateur pourra maintenant diffuser un contenu à n'importe quelle étape qui suit sa première prise de connaissance du contenu. Un utilisateur qui voit un contenu pour la première fois chez un de ses voisins à l'étape t aura une probabilité de le rediffuser à l'étape $t+1$, mais aussi à l'étape $t+2$ ainsi que toutes les suivantes, amenant un renforcement dans sa probabilité d'avoir diffusé le contenu. Il s'agit là d'un représentation plus fine du phénomène de diffusion. On définit par $P(n_i, c^k, \leq t)$ la probabilité que l'utilisateur n_i ai diffusé le contenu c^k avant l'étape de temps t.

La dynamique du système de diffusion évolue étape après étape, la probabilité d'être actif d'un utilisateur à une étape $t+1$ étant sa probabilité d'avoir été actif à l'étape de temps t à laquelle s'ajoute sa probabilité de s'être activé entre les étapes t et $t+1$:

$$P(n_i, c^k, \leq t+1) = \quad P(n_i, c^k, \leq t) + (1 - P(n_i, c^k, \leq t))P(n_i, c^k, t) \tag{2.3}$$

Par définition :

- $P(n_i, c^k, \leq 0) = 1$ si n_i est un diffuseur initial
- $P(n_i, c^k, \leq 0) = 0$ sinon

Nous venons de donner l'équation permettant de passer d'une étape à la suivante. Il est possible de calculer directement la probabilité d'un utilisateur d'être actif à l'étape de temps t en déroulant le processus précédent. On peut aussi montrer par récurrence :

Théorème 1. *La probabilité d'un utilisateur d'être actif s'exprime comme suit :*

$$P(n_i, c^k, \leq t) = \sum_{t'=0}^{t-1} P(n_i, c^k, t') \prod_{\tau=0}^{t'-1} (1 - P(n_i, c^k, \tau))$$

La démonstration du théorème 1 se trouve en annexe A.

Ceci étant, il n'est plus possible d'obtenir le nombre de voisins entrant actifs d'un utilisateur comme dans le cas du modèle UC. La pression sociale se caractérise ainsi par l'espérance du nombre de voisins entrant actifs :

$$SP(n_i, \mathcal{G}, M^k, t) = E[|\mathcal{C}^k(n_i, t)|]$$

Cette espérance est définit comme :

$$E[|\mathcal{C}^k(n_i, t)|] = \sum_{m=0}^{|\mathcal{B}(n_i)|} m \; P(|\mathcal{C}^k(n_i, t)| = m)$$

où $P(|\mathcal{C}^k(n_i, t)| = m)$ est la probabilité que le nombre de voisins entrant qui ont diffusé le contenu soit m. On peut montrer que cette espérance se calcule aussi comme la somme des probabilités d'être actifs des voisins entrant :

Théorème 2. *L'espérance du nombre de voisins actifs est :*

$$E[|\mathcal{C}^k(n_i, t)|] = \sum_{n_j \in \mathcal{B}(n_i)} P(n_j, c_k, \leq t)$$

La démonstration du théorème 2 se trouve en annexe B.

L'algorithme 2 montre le détail de l'exécution du modèle RUC étape par étape. Il est possible d'optimiser l'algorithme en ne mettant à jour que les probabilités d'être actif des utilisateurs atteignables à chaque étape de temps. En effet, si un utilisateur n'a pas vu un contenu, il aura une probabilité nulle de le rediffuser.

Ce modèle, contrairement aux modèles standards et au modèle UC possède une propriété intéressante mais peu intuitive. Prenons comme exemple le graphe de la figure 2.2 dans lequel trois utilisateurs sont reliés par deux liens. Supposons que l'utilisateur A diffuse un contenu. Le modèle RUC tel que nous l'avons défini permet après un nombre

Algorithme 2: Exécution du modèle RUC

Données : Un graphe $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
Un contenu c^k
Un ensemble d'utilisateurs diffuseurs $Initiaux$
$Tmax$: le nombre d'étapes de la diffusion
Résultat : Probabilité des utilisateurs d'être actifs après processus de diffusion
$Etat \leftarrow$ Tableau($|\mathcal{V}|$);
pour tous les $u \in \mathcal{V}$ **faire**
  si $u \in Initiaux$ **alors**
    $Etat[u] = 1$;
  sinon
    $Etat[u] = 0$;
  fin
fin
pour $t = 0 \rightarrow Tmax$ **faire**
  $PrecedentEtat \leftarrow Etat$;
  pour tous les $u \in \mathcal{V}$ **faire**
    $Etat[u] = PrecedentEtat[u] + (1 - PrecedentEtat[u]) \times P(n_u, c^k, t)$;
  fin
fin

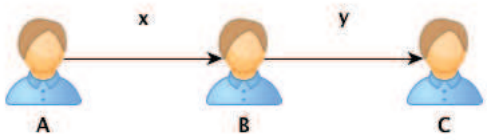


FIGURE 2.2 – Exemple de graphe social.

quelconque d'étapes de temps t que l'utilisateur C ait une probabilité d'avoir (re)diffusé le contenu plus forte que l'utilisateur B alors même que pour le rediffuser il faut auparavant que B le lui ai partagé. Ceci est dû au fait que la probabilité d'être actif des voisins entrant d'un utilisateur n'est qu'un paramètre dans le calcul de la probabilité que l'utilisateur diffuse le contenu. Contrairement aux autres modèles dans lesquels c'est la caractéristique principale. Ce cas de figure se produira principalement quand le contenu en question fait parti des centres d'intérêts de l'utilisateur C et pas de ceux de l'utilisateur B. On peut voir cette caractéristique comme la prise en compte de bruit dans les données. En effet, cela revient à prendre en compte le même graphe et à ajouter un lien entre les utilisateurs A et C, qui peut ne pas être visible dans les données mais exister dans la réalité.

Le principal défaut de ce modèle repose sur le long terme. A partir du moment où la probabilité de diffuser d'un utilisateur est positive à une étape t, elle ne diminuera jamais par la suite. Dû au phénomène de renforcement, sa probabilité d'être actif va de ce fait augmenter au fur et à mesure des étapes. On va donc avoir comme effet lors d'une diffusion

l'activation de tous les utilisateurs atteignables après un nombre d'étapes suffisamment grand. Pour palier ce problème, nous avons proposé le modèle DRUC dans lequel nous intégrons un paramètre d'oubli.

2.5 Ajout d'un paramètre d'oubli : DRUC

2.5.1 Définition du modèle

L'idée derrière l'ajout de ce paramètre d'oubli est de diminuer la probabilité qu'un utilisateur diffuse un contenu ancien. Si un utilisateur voit un contenu partagé par l'un de ses voisins entrant, il aura une plus forte probabilité de le rediffuser le lendemain que 2 ans plus tard. Ce phénomène est étudié dans l'article [Wu and Huberman, 2007] dans lequel les auteurs montrent que les utilisateurs sont affectés par la nouveauté des informations et qu'un phénomène de lassitude apparaît quand celle-ci n'est plus assez récente. Ce paramètre va donc diminuer l'impact sur la volonté de diffuser qu'ont les voisins plus le temps passe. Afin d'introduire ce paramètre dans le modèle, nous définissons une nouvelle grandeur qui caractérise tous les utilisateurs, pour un contenu donné à chaque étape de temps : l'influence $\rho(n_i, c^k, t)$. L'influence de chaque utilisateur sur ses voisins évolue au cours du temps de la façon suivante :

$$\rho(n_i, c^k, t+1) = \quad \delta \times \rho(n_i, c^k, t) + (1 - P(n_i, c^k, \leq t))P(n_i, c^k, t) \tag{2.4}$$

où δ est le paramètre d'oubli qui régule l'influence. Ce paramètre d'oubli étant là pour diminuer l'influence au cours du temps, on a $\delta \in [0, 1]$. Par définition, l'initialisation est la suivante :

- $\rho(n_i, c^k, 0) = 1$ si n_i est un diffuseur initial
- $\rho(n_i, c^k, 0) = 0$ sinon

Cette influence sert à définir la pression sociale. Au lieu de prendre en compte le nombre de voisins entrant ayant déjà diffusé le contenu (UC) ou l'espérance de ce nombre (RUC), dans le modèle DRUC on tient compte de l'influence globale des voisins entrant :

$$SP(n_i, \mathcal{G}, M^k, t) = \sum_{n_j \in \mathcal{B}(n_i)} \rho(n_j, c^k, t) \tag{2.5}$$

L'exécution du modèle reste ensuite la même que celle du modèle RUC et est décrite dans l'algorithme 2.

On voit deux cas particuliers ressortir. Si $\delta = 0$, le modèle ne tient compte que des utilisateurs qui ont eu une probabilité de diffuser le contenu à l'étape précédente et se rapproche à ce moment là du concept du modèle UC. Si $\delta = 1$, le modèle DRUC est équivalent au modèle RUC.

2.5.2 Comparaison entre RUC et DRUC

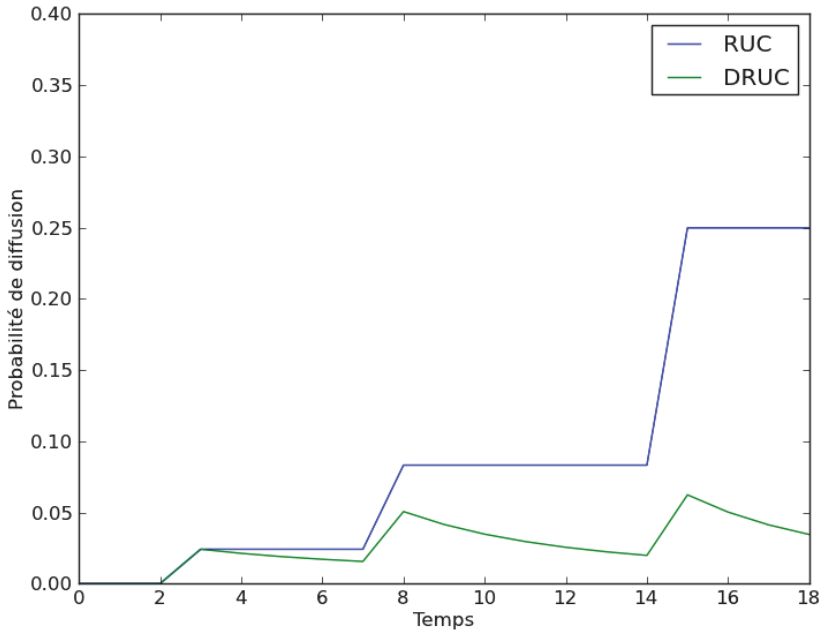


FIGURE 2.3 – Apport du paramètre d'oubli.

Afin de montrer l'apport du paramètre d'oubli, nous proposons un petit exemple pour comparer les modèles RUC et DRUC. On se place ici dans un réseau quelconque et on observe l'évolution d'un utilisateur au cours du temps lors de la diffusion d'un contenu. Au départ (étape 0) il n'a aucune connaissance du contenu. A $t = 2$ le premier de ses voisins entrant lui diffuse le contenu. Aux étapes de temps 7 et 14 deux nouveaux voisins lui diffusent le même contenu.

La figure 2.3 représente la probabilité que l'utilisateur diffuse le contenu à chaque étape de temps en utilisant les deux modèles. On y voit clairement le partage du contenu par chacun des trois voisins, correspondant à une forte hausse dans la probabilité de diffusion. Dans cet exemple l'impact du paramètre d'oubli se caractérise par une baisse constante de la probabilité de diffusion. La probabilité de diffusion dictée par le modèle DRUC a les mêmes hausses que celle dictée par le modèle RUC lorsque le nombre de sources augmente, mais celle donnée par le modèle DRUC baisse rapidement lors des étapes qui suivent l'arrivée d'une nouvelle source.

Nous avons choisi un exemple dans lequel le paramètre d'oubli instaure une faible décroissance de l'influence ($\delta = 0.9$). Les autres paramètres ont été fixés de telle sorte que $\lambda_0 + \lambda_1 \Phi_1^{n_i,c^k,t} + \lambda_2 \Phi_2^{n_i,c^k,t} = -5$ avec $\lambda_3 = 1.3$. Cette situation correspond à un utilisateur qui diffuse peu d'information et/ou qui n'est pas spécialement intéressé par le contenu

diffusé.

Si l'on prend un exemple dans lequel le contenu est dans les centres d'intérêts de l'utilisateur, le paramètre d'oubli aura beaucoup moins d'impact. En effet, de part la définition du modèle, le paramètre d'oubli a de l'impact si la pression sociale compense le non intérêt pour le contenu ou le manque de volonté de diffuser d'un utilisateur, c'est-à-dire si $\lambda_0 + \lambda_1 \Phi_1^{n_i,c^k,t} + \lambda_2 \Phi_2^{n_i,c^k,t} < 0$. Il permet d'éviter qu'avec le temps le modèle donne à un utilisateur qui a une faible probabilité de diffuser un contenu, une forte probabilité de l'avoir diffusé. Ce paramètre sera par contre peu influent sur un utilisateur ayant une forte probabilité de diffuser le contenu.

2.6 Estimation des paramètres

Nous partageons notre jeu de données en deux parties : un partie qui sert pour l'entraînement et une partie qui sert pour le test. Toutes les estimations des paramètres sont effectuées sur le jeu d'entraînement. Le jeu de test sert quant à lui à valider la qualité des modèles. Une description plus précise des jeux de données est donnée dans le chapitre 3.

2.6.1 Valeurs des paramètres de seuil

Nous estimons les paramètres de seuil de telle sorte que la composante dans laquelle ils interviennent (pour le paramètre θ_s il s'agit de $S(n_i, \mathcal{P}, c^k, \theta_s)$) soit positive si l'utilisateur a une probabilité plus forte de diffuser le contenu plutôt que de ne pas le diffuser, et négative si la probabilité est plus forte qu'il ne diffuse pas le contenu. Pour l'estimation du paramètre de seuil de similarité θ_s, nous calculons, sur le jeu d'entraînement, la similarité entre tous les utilisateurs et tous les contenus. Nous créons ensuite un histogramme (avec un pas de 0.05) du nombre d'utilisateurs ayant diffusé les contenus et ne les ayant pas diffusé. A titre d'exemple, la figure 2.4 représente cet histogramme pour le jeu de données **dense-icwsm** que nous présentons dans le chapitre 3.

Nous choisissons ensuite comme seuil de similarité la valeur de similarité au delà de laquelle le nombre d'utilisateurs ayant diffusé les contenus est plus important que le nombre d'utilisateurs ne les ayant pas diffusé. Dans l'exemple que nous donnons il s'agit du seuil $\theta_s = 0.4$.

Fixer la valeur du paramètre θ_w est plus simple. En effet, l'activité d'un utilisateur est calculée comme sa probabilité de rediffuser un contenu qu'il a reçu. En suivant la même logique que pour le paramètre θ_s nous fixons $\theta_w = 0.5$.

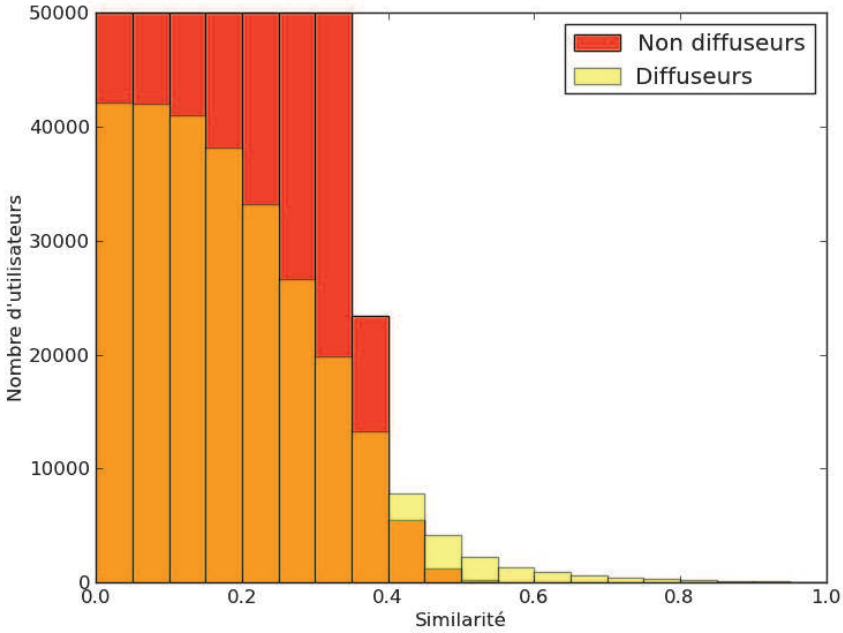


FIGURE 2.4 – Histogramme de répartition des similarités. Le nombre d'utilisateurs a été seuillé à 50000.

2.6.2 Estimation des paramètres λ

Algorithme de montée de gradient

L'algorithme de montée de gradient est une méthode d'optimisation pour trouver un maximum local d'une fonction dérivable. A chaque étape, on calcule le gradient de la fonction au point courant puis on met à jour les coordonnées du point courant en suivant le gradient. La figure 2.5 image une étape de l'algorithme pour une fonction définie sur une seule dimension.

Pour ce qui est de l'estimation des paramètres de nos modèles, la fonction que l'on va chercher à maximiser est la vraisemblance $\mathcal{L}(\lambda_0, \lambda_1, \lambda_2, \lambda_3)$ (ou plus particulièrement la log-vraisemblance $\mathcal{LL}$) entre les prédictions des modèles et les données réelles. Par rapport à une montée de gradient standard, la fonction P n'est définie que pour $\lambda_1 < 0$, $\lambda_2 < 0$ et $\lambda_3 < 0$, ce qui impose des contraintes d'intervalles. Le problème d'apprentissage s'exprime donc comme suit :

$$\begin{cases} \text{argmax}_{\lambda_0,\lambda_1,\lambda_2,\lambda_3} \mathcal{LL}(\lambda_0, \lambda_1, \lambda_2, \lambda_3) \\ \text{sous contraintes : } \lambda_1 \geq 0,\ \lambda_2 \geq 0,\ \lambda_3 \geq 0 \end{cases}$$

La méthode de gradient projeté permet de résoudre ce problème d'optimisation sous contraintes. Le principe est le même que la méthode du gradient standard sauf que chaque

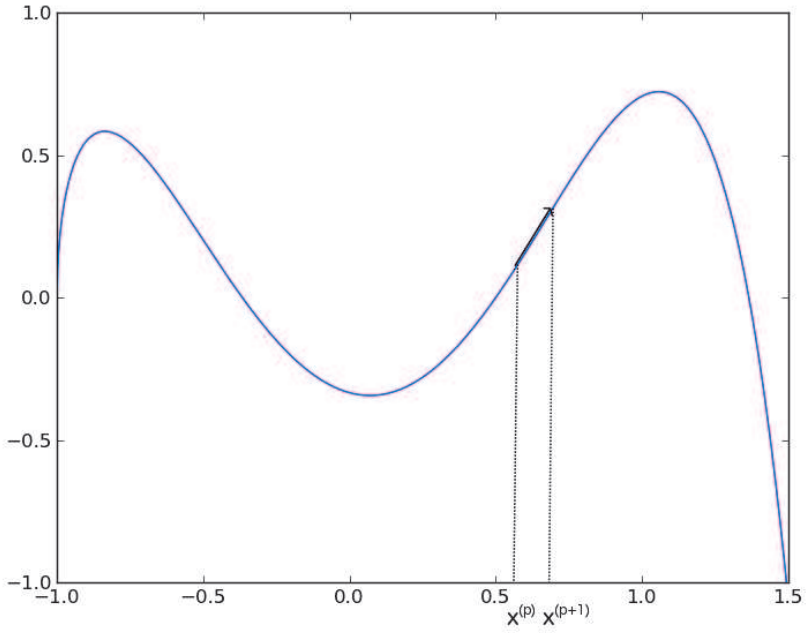


FIGURE 2.5 – Algorithme du gradient : passage de l'étape p à l'étape $p+1$.

étape de montée de gradient est suivie par une projection sur l'espace d'intervalles admissibles. On obtient les formules de mise à jour suivantes (entre les étapes $(p+1)$ et p) :

$$\forall d \in \{0,1,2,3\} : \begin{cases} \lambda_d^{(p+1)} = \lambda_d^{(p)} + \alpha \frac{\partial \mathcal{LL}(\lambda_0^{(p)}, \lambda_1^{(p)}, \lambda_2^{(p)}, \lambda_3^{(p)})}{\partial \lambda_d} \\ \text{Si } \lambda_{d(d\neq 0)}^{(p+1)} < 0, \text{ alors } \lambda_{d(d\neq 0)}^{(p+1)} = 0 \end{cases}$$

Le paramètre α permet de contrôler la vitesse de montée en suivant le gradient de $\mathcal{LL}$.

Dans notre cas, les dérivées partielles de la vraisemblance ont des valeurs qui dépendent beaucoup du jeu de données ainsi que du modèle. De plus, les échelles des valeurs varient beaucoup, ce qui pose des problèmes pour trouver une valeur de α qui permette de converger rapidement. Nous avons observé que dans tous les cas, les valeurs des paramètres sont entre 0 et 20. C'est pour cela que nous utilisons une variante de la méthode de montée de gradient dans laquelle nous n'utilisons pas la valeur de la dérivée mais seulement son signe :

$$\forall d \in \{0,1,2,3\} : \begin{cases} \lambda_d^{(p+1)} = \lambda_d^{(p)} + (\alpha^{-p}) sign(\frac{\partial \mathcal{LL}(\lambda_0^{(p)}, \lambda_1^{(p)}, \lambda_2^{(p)}, \lambda_3^{(p)})}{\partial \lambda_d}) \\ \text{Si } \lambda_{d(d\neq 0)}^{(p+1)} < 0, \text{ alors } \lambda_{d(d\neq 0)}^{(p+1)} = 0 \end{cases}$$

Nous avons choisi une valeur de paramètre $\alpha = 5$. Le contrôle de la vitesse ne se fait plus par la dérivée mais en fonction de l'étape de temps. De plus, notre fonction n'étant pas concave, nous risquons de trouver des maximums locaux. Afin d'éviter ce biais, nous estimons plusieurs fois les paramètres en choisissant des valeurs au temps $(p) = 0$ différentes. Nous avons choisi les valeurs de départ :

$$\lambda^{(0)}_{d(d\neq 0)} \in \{0, 5, 10\}$$
$$\lambda^{(0)}_{0} \in \{-10, -5, 0, 5, 10\}$$

Comme les valeurs des paramètres évoluent d'une étape sur l'autre en fonction du signe de la dérivée mais avec la même valeur, il arrive souvent que plusieurs paramètres aient la même valeur.

Nous allons maintenant donner la forme de la vraisemblance pour les trois modèles centrés utilisateur que nous avons proposé.

Maximum de vraisemblance pour le modèle à cascades UC

Dans le jeu d'entraînement, on connaît pour chaque contenu les utilisateurs qui l'ont diffusé et à quel moment. La vraisemblance du modèle UC est optimale :

- quand un utilisateur est diffuseur du contenu à une étape donnée, le modèle prédit avec une forte probabilité que l'utilisateur diffuse le contenu
- quand un utilisateur est diffuseur du contenu à une étape t, le modèle prédit une faible probabilité de diffusion pour ses voisins sortant qui ne sont pas diffuseurs du contenu à l'étape $t + 1$.

L'inverse est également vrai. L'équation suivante donne la vraisemblance pour le modèle UC :

$$\mathcal{L}(\lambda_1, \lambda_2, \lambda_3) = \prod_{k=1}^{\ell} \prod_{t=1}^{T^k} [\prod_{n_i \in D^k(t)} P(n_i, c^k, t-1) \prod_{n_i \in D^k(t-1)} \prod_{n_j \in F(n_i) \setminus C^k(t)} (1 - P(n_j, c^k, t-1))] \tag{2.6}$$

Dans les données que l'on recueille, à chaque fois qu'un utilisateur partage un contenu, on connaît la date (seconde, minute et heure) à laquelle il l'a fait. Le modèle UC, comme le modèle IC, ne considère pas les événements d'une diffusion au moment où ils se produisent mais les uns par rapport aux autres. En d'autres mots, si un utilisateur diffuse un contenu au temps t dans la réalité et que l'un de ses voisins le reprend au temps $t + 5$, pour le modèle UC il s'agira de deux étapes de temps qui se suivent. Pour cela, pour l'estimation des paramètres des modèles UC et IC, nous ramenons tous les temps de rediffusion d'un contenu sur une échelle de temps relative à la première diffusion du contenu.

Maximum de vraisemblance pour les modèles avec renforcement RUC et DRUC

Dû au fait que l'on calcule directement la probabilité d'être actif de tous les utilisateurs à chaque étape de temps pour les modèles avec renforcement, le calcul de leur vraisemblance est assez intuitif. Si un utilisateur est diffuseur d'un contenu à une étape donnée, le modèle doit prédire avec forte probabilité l'utilisateur comme diffuseur à cette même étape. Inversement, si un utilisateur est non diffuseur d'un contenu à une étape donnée, le modèle doit le prédire non diffuseur. L'équation suivante donne la vraisemblance pour les modèles RUC et DRUC :

$$\mathcal{L}(\lambda_0,\lambda_1,\lambda_2,\lambda_3) = \prod_{k=1}^{\ell}\prod_{t=1}^{T^k}[\prod_{n_i\in C^k(t)} P(n_i,c^k,\leq t) \prod_{n_i\notin C^k(t)} (1-P(n_i,c^k,\leq t))] \tag{2.7}$$

La différence pour le calcul de cette vraisemblance entre les deux modèles réside dans le calcul de la pression sociale $SP(n_i,\mathcal{G},M^k,t)$.

Afin de réduire le coût du calcul du gradient de cette vraisemblance, nous utilisons l'équation 2.3 pour calculer les dérivés partielles, puis stockons, pour chaque utilisateur, les valeurs des probabilités $P(n_i,c^k,\leq t)$ et de leurs dérivées à chaque étape de temps. La dérivée de l'équation 2.3 est la suivante :

$$\begin{aligned}\frac{\partial P(n_i,c^k,\leq t+1)}{\partial\lambda_d} = \quad & \tfrac{\partial P(n_i,c^k,\leq t)}{\partial\lambda_d}(1-P(n_i,c^k,t)) \\ & +\tfrac{\partial P(n_i,c^k,t)}{\partial\lambda_d}(1-P(n_i,c^k,\leq t))\end{aligned} \tag{2.8}$$

Calcul du gradient de la fonction de probabilité de diffusion

Pour les trois modèles, les dérivées des équations pour la mise à jour des probabilités de diffusion à chaque étape sont données ci-dessous.
Si $SP(n_i,\mathcal{G},M^k,t)>0$:

$$\frac{\partial P(n_i,c^k,t)}{\partial\lambda_1} = \frac{(S(n_i,\mathcal{P},c^k,\theta_s))(e^{-\lambda_1 S(n_i,\mathcal{P},c^k,\theta_s)-\lambda_2 W(n_i,\mathcal{G},\mathcal{D},\theta_w)-\lambda_3 SP(n_i,\mathcal{G},M^k,t)})}{(1+e^{-\lambda_1 S(n_i,\mathcal{P},c^k,\theta_s)-\lambda_2 W(n_i,\mathcal{G},\mathcal{D},\theta_w)-\lambda_3 SP(n_i,\mathcal{G},M^k,t)})^2}$$

$$\frac{\partial P(n_i,c^k,t)}{\partial\lambda_2} = \frac{(W(n_i;\theta_w))(e^{-\lambda_1 S(n_i,\mathcal{P},c^k,\theta_s)-\lambda_2 W(n_i,\mathcal{G},\mathcal{D},\theta_w)-\lambda_3 SP(n_i,\mathcal{G},M^k,t)})}{(1+e^{-\lambda_1 S(n_i,\mathcal{P},c^k,\theta_s)-\lambda_2 W(n_i,\mathcal{G},\mathcal{D},\theta_w)-\lambda_3 SP(n_i,\mathcal{G},M^k,t)})^2}$$

$$\frac{\partial P(n_i,c^k,t)}{\partial\lambda_3} = \frac{(SP(n_i,\mathcal{G},M^k,t))(e^{-\lambda_1 S(n_i,\mathcal{P},c^k,\theta_s)-\lambda_2 W(n_i,\mathcal{G},\mathcal{D},\theta_w)-\lambda_3 SP(n_i,\mathcal{G},M^k,t)})}{(1+e^{-\lambda_1 S(n_i,\mathcal{P},c^k,\theta_s)-\lambda_2 W(n_i,\mathcal{G},\mathcal{D},\theta_w)-\lambda_3 SP(n_i,\mathcal{G},M^k,t)})^2}$$

sinon $\frac{\partial P(n_i,c^k,t)}{\partial\lambda_d}=0$.

Il es à noter que lors de l'estimation des paramètres, nous utilisons un jeu de données

pour lequel nous connaissons à l'avance les temps d'activation réels des utilisateurs, contrairement à la phase de test où elles sont calculées par le modèle. Ceci a comme conséquence que les valeurs de pressions sociale $SP(n_i, \mathcal{G}, M^k, t)$ sont des constantes.

2.7 Illustration

Nous proposons comme première approche empirique des modèles centrés utilisateur une simulation sur des graphes sociaux. En utilisant le modèle RUC, le but de cette expérience est de visualiser le schéma général de diffusion qu'entraînent ces modèles. Les paramètres du modèle sont choisis arbitrairement de manière à obtenir une diffusion relativement importante : quelles que soient les valeurs choisies, le schéma de diffusion globale reste similaire. Les profils utilisateur, le contenu et la volonté de diffuser des utilisateurs sont générés aléatoirement. Nous observons la diffusion d'un contenu qui a été diffusé pour la première fois par un unique utilisateur en utilisant à la fois des graphes sociaux générés afin de pouvoir voir le résultat de la diffusion dans des cas particuliers ainsi que deux graphes sociaux réels tirés de Enron [Klimt and Yang, 2004] et ICWSM [Burton et al., 2009].

La figure 2.6 montre l'évolution de la proportion des utilisateurs diffuseurs et non diffuseurs dans un graphe social de 1000 utilisateurs et 5000 liens générés aléatoirement. Au début de la diffusion, il n'y a que peu d'utilisateurs touchés par l'information et la diffusion peine à démarrer. On voit cependant que la diffusion s'accélère avec le nombre d'utilisateurs ayant diffusé le contenu pour ensuite ralentir car de moins en moins d'utilisateurs du réseau ne connaissent pas l'information.

Dans la figure 2.7 nous présentons une diffusion dans un contexte un peu particulier. Nous avons généré ici un graphe social dans lequel on retrouve deux communautés bien distinctes qui sont connectées entre elles par un unique utilisateur. On repère le même schéma de diffusion que précédemment mais sur deux étapes de temps. Tout d'abord, l'information se propage au sein de la première communauté, puis l'utilisateur faisant le lien en prend connaissance et la rediffuse. Démarre ensuite une nouvelle diffusion au sein de la seconde communauté.

Un autre cas particulier de diffusion intéressant est le cas d'un graphe en étoile. Prenons un nœud central qui possède un certain nombre de nf filles et des liens bidirectionnels vers celles-ci. Nous avons ainsi un graphe en étoile à un niveau. Pour chacune des filles du nœud central, on crée de nouveau nf filles. On obtient un graphe en étoile à deux niveaux (voir la figure 2.9). On peut ainsi définir un graphe en étoile de nn niveaux et nf filles.

La figure 2.8(a) montre la diffusion d'un contenu dans un graphe en étoile de cinq niveaux et quatre filles avec comme diffuseur initial le nœud central. On obtient de nouveau le même schéma de diffusion que précédemment. La figure 2.8(b) montre la diffusion d'un contenu dans le même graphe avec comme diffuseur initial un nœud sur la bordure (le

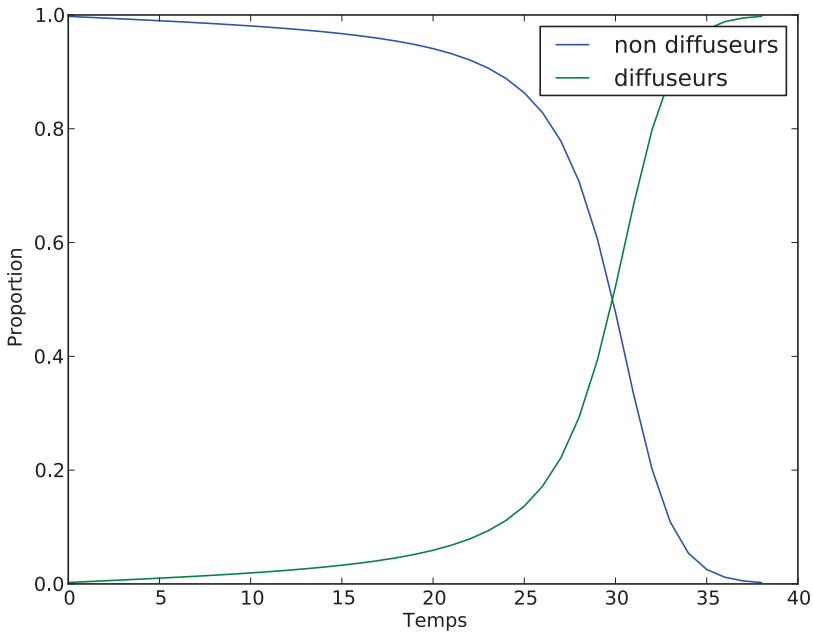


FIGURE 2.6 – Diffusion d'un contenu dans un graphe artificiel de 1000 utilisateurs et 5000 liens.

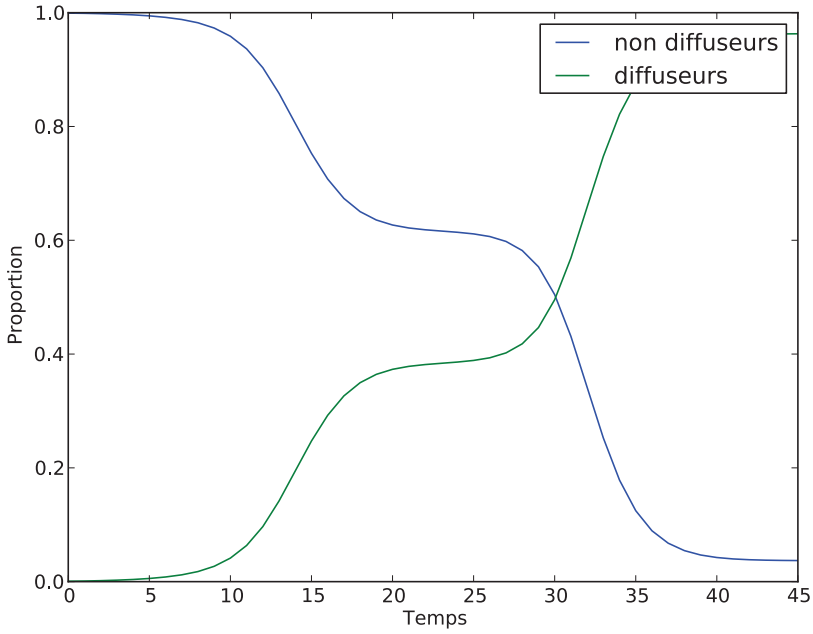
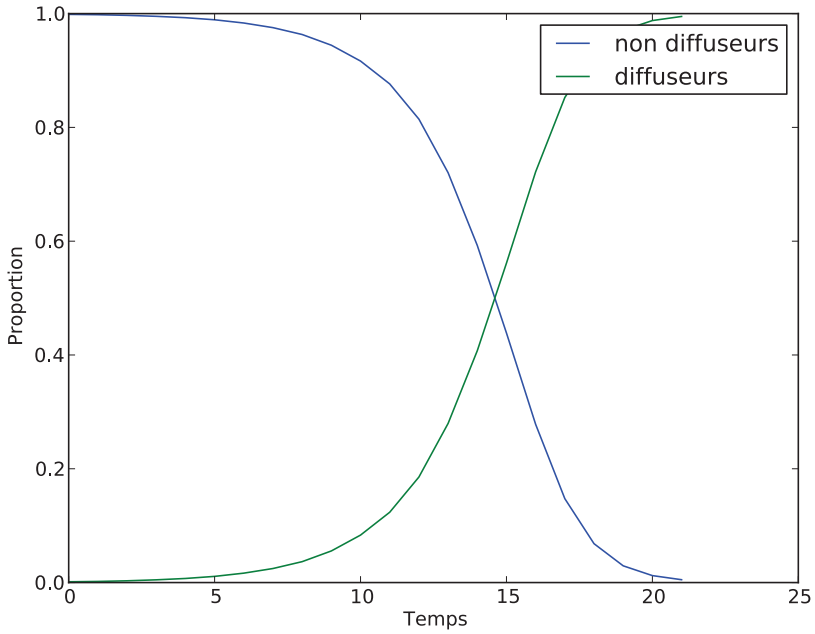
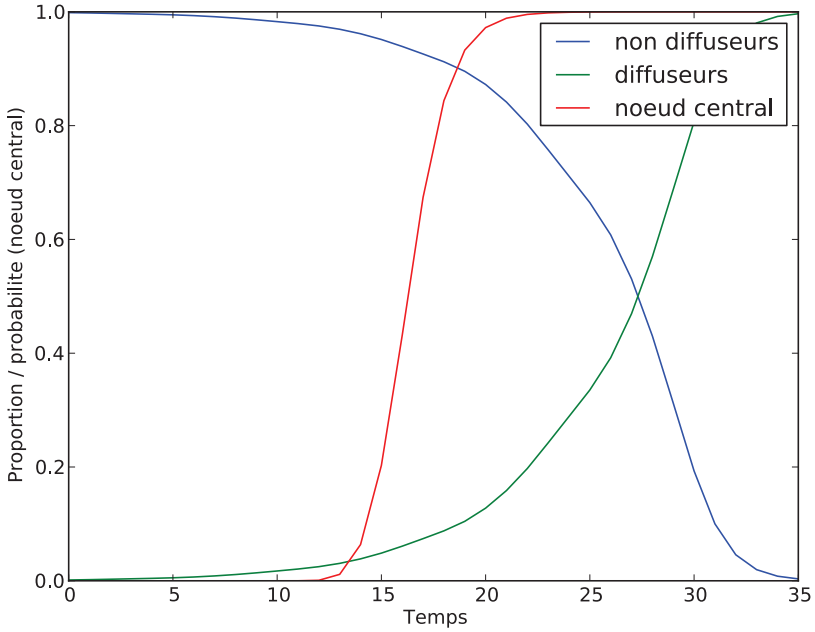


FIGURE 2.7 – Diffusion d'un contenu dans un graphe artificiel regroupant deux communautés reliées entre elles par un unique utilisateur.



(a) Diffusion à partir du nœud central



(b) Diffusion à partir d'un nœud en bordure du graphe

FIGURE 2.8 – Diffusion d'un contenu dans un graphe artificiel en étoile (5 niveaux, 4 filles).

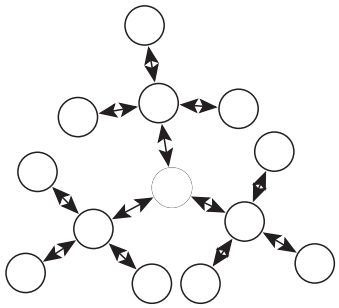


FIGURE 2.9 – Exemple de graphe en étoile avec deux niveaux et trois filles.

dernier niveau le plus loin du centre) du graphe. On remarque tout d'abord que la diffusion qui part de la bordure du graphe est beaucoup plus lente que celle qui part du centre. En effet, le centre est un passage obligatoire pour atteindre l'ensemble du réseau. Comme on peut le voir sur la figure 2.8(b), la diffusion ne commence réellement à devenir importante que lorsque le nœud central est touché.

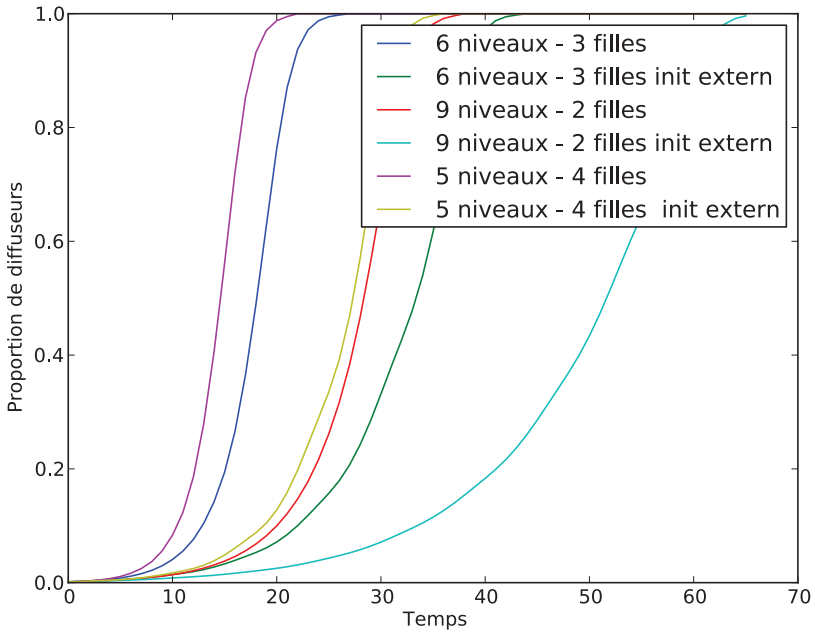
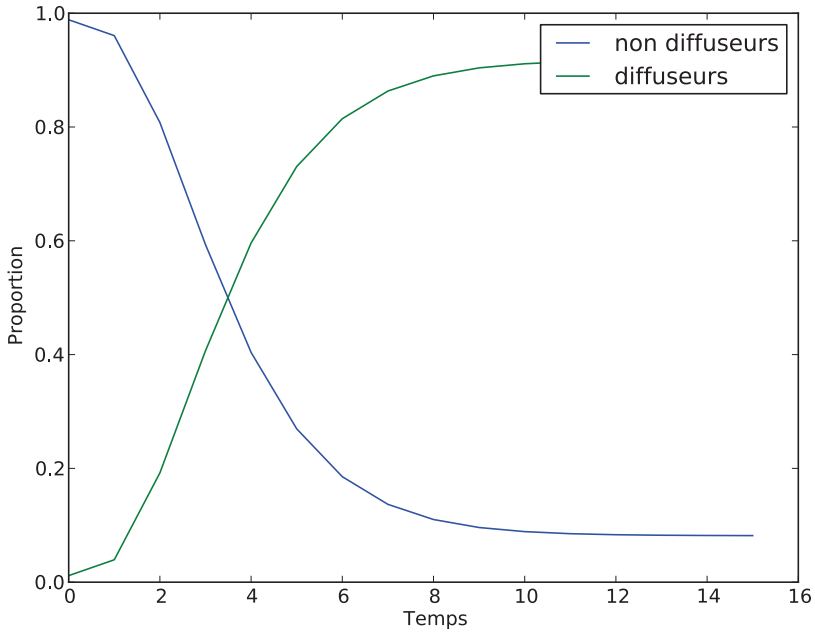


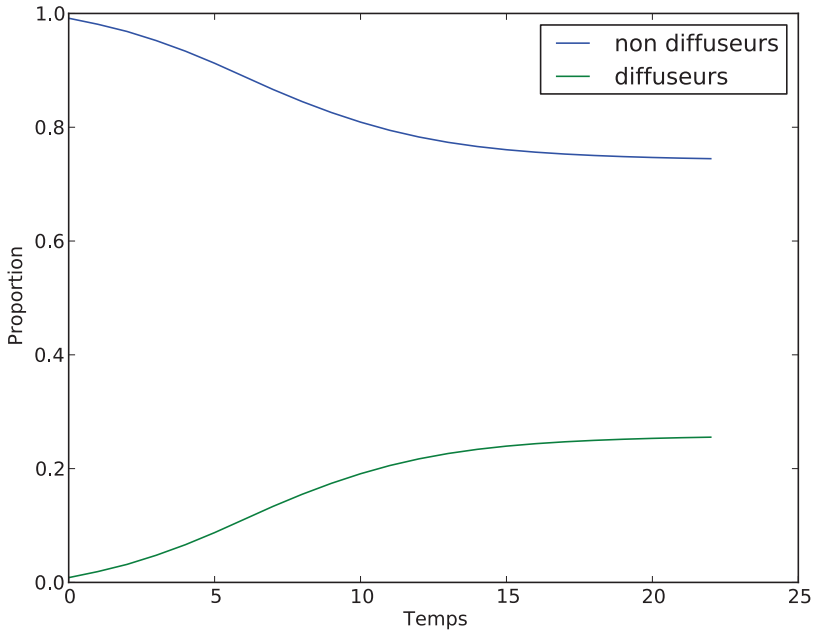
FIGURE 2.10 – Comparaison de la vitesse de diffusion sur plusieurs graphes en étoile.

Afin de comparer la vitesse de diffusion en fonction des propriétés du graphe en étoile, la figure 2.10 montre la proportion de diffuseurs dans différents graphes en étoile. On voit que plus le graphe est aplati (en d'autres mots plus le nombre de niveaux est faible et le nombre de filles par niveau est grand), plus la propagation d'un contenu est rapide.

Enfin, la figure 2.11 montre l'évolution de la diffusion d'un contenu dans deux graphes



(a) Graphe social *Enron*



(b) Graphe social *ICWSM09*

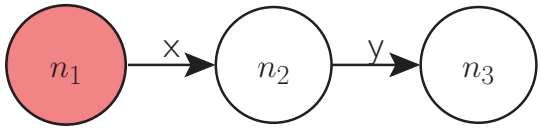
FIGURE 2.11 – Diffusion d'un contenu dans un graphe social réel.

sociaux réels. Le graphe social *Enron* est tiré d'un ensemble d'échanges de mails au sein d'une entreprise, et le graphe *ICWSM09* est un ensemble de billets de blogs que nous utilisons par la suite dans le chapitre 3 pour la validation des modèles. Le schéma de diffusion est similaire à celui obtenu sur les graphes artificiels. On remarque cependant que tous les utilisateurs ne sont pas atteints par la diffusion. Ceci est dû au fait que dans les réseaux sociaux réels, tous les utilisateurs ne sont pas forcément atteignables depuis n'importe quel diffuseur initial.

2.8 Non-équivalence des modèles avec renforcement et des modèles standards

Comme dit précédemment dans le chapitre 1, les modèles standards de seuil linéaire (LT) et à cascades indépendantes (IC) ont étés prouvés équivalents à une percolation de lien. Nous montrons que ce n'est pas le cas des modèles RUC et DRUC.

Théorème 3. *Le modèle RUC (et par extension le modèle DRUC) défini sur un graphe* $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ *n'est pas équivalent à un processus de percolation de lien sur* $\mathcal{G}$*.*



(a) Graphe social

probabilité	x	y
P_1	fermé	fermé
P_2	fermé	ouvert
P_3	ouvert	fermé
P_4	ouvert	ouvert

(b) table des probabilités

FIGURE 2.12 – Exemple de percolation de liens

Démonstration. Afin de prouver le théorème 3, il suffit d'exhiber un graphe sur lequel une instance du modèle RUC ne peut pas être équivalente à une percolation de liens. La figure 2.12(a) définit un graphe pour lequel ce n'est pas possible. Il contient trois utilisateurs connectés entre eux par deux liens, et le diffuseur initial du contenu c^k est l'utilisateur n_1. Après deux étapes de temps, tous les utilisateurs du réseau auront donc une probabilité non nulle d'avoir reçu le contenu initialement diffusé par n_1. L'ensemble $\mathcal{V}$ des vecteurs définissant tous les états possibles du graphe $\mathcal{G}$ contient quatre vecteurs, et la percolation de liens sur ce graphe est caractérisée par trois probabilités (P_1, P_2, P_3), la quatrième étant définie par : $P_1 + P_2 + P_3 + P_4 = 1$. La figure 2.12(b) donne la table des probabilités associée.

Si le modèle RUC est équivalent à une percolation de lien, les probabilités que les utilisateurs n_2 et n_3 soient actifs après deux étapes de temps avec le modèle RUC sont les suivantes :

- $P(n_2, c^k, \leq 2) = P_3 + P_4$
- $P(n_3, c^k, \leq 2) = P_4$

ce qui nous amène à :

$$\begin{aligned} P_3 &= P(n_2, c^k, \leq 2) - P(n_3, c^k, \leq 2) \\ &= P(n_2, c^k, 0) + (1 - P(n_2, c^k, 0)) \times P(n_2, c^k, 1) - P(n_3, c^k, 1) \end{aligned}$$

où la dernière équation est basée sur les définitions de $P(n_2, c^k, \leq 2)$ et $P(n_3, c^k, \leq 2)$.

Cette valeur peut être négative si le contenu est plus dans les centres d'intérêts de l'utilisateur n_3 que de l'utilisateur n_2. Prenons le contexte suivant :

- $\lambda_0 = \lambda_2 = \lambda_3 = 0$, $\lambda_1 = 1$
- $Act(n_2, \mathcal{G}, \mathcal{D}) = Act(n_3, \mathcal{G}, \mathcal{D}) = \theta_w = 0$
- $sim(p^2, c^k) = 0$
- $sim(p^3, c^k) = 1$
- $\theta_s = 0.5$

alors $P_c(n_2, c^k, 0) = P_c(n_2, c^k, 1) \approx 0.38$ et $P_c(n_3, c^k, 1) \approx 0.62$
ce qui a pour conséquence : $P_3 \approx -0.01$, ce qui est impossible car P_3 est une probabilité.

On obtient donc une contradiction, prouvant que le modèle RUC n'est pas équivalent à une percolation de lien, et n'est donc équivalent ni au modèle à cascades indépendantes IC, ni au modèle de seuil linéaire LT. □

2.9 Conclusion

Nous avons proposé dans ce chapitre une nouvelle famille de modèles de diffusion de contenus dans les réseaux sociaux. Ces modèles sont probabilistes et basés sur l'utilisation de caractéristiques utilisateur. Ils prennent en compte :

- le contenu de l'information diffusé ;
- le goûts des utilisateurs et leur intérêt pour le contenu diffusé ;
- la tendance des utilisateurs à diffuser, c'est-à-dire s'ils diffusent peu ou beaucoup d'information ;
- la pression sociale subie par les utilisateurs, c'est-à-dire le nombre de leurs voisins ayant diffusé le contenu.

Cette famille comprend trois modèles. Une adaptation du modèle à cascades indépendantes qui définit la probabilité d'un utilisateur de diffuser un contenu en fonction des caractéristiques utilisateur. Deux autres modèles qui intègrent un phénomène de renforcement permettant une prise en compte temporelle plus fine (un utilisateur peut diffuser un contenu longtemps après l'avoir reçu). Ces modèles reposant sur peu de paramètres, ils souffrent peu du manque de diffusion que l'on retrouve dans les données réelles. Ces

paramètres sont appris par maximum de vraisemblance en utilisant un algorithme de montée de gradient. Enfin, nous avons montré que les modèles avec renforcement ne sont pas équivalents à une percolation de lien, ce qui les rend par conséquence différents des modèles standards du domaine.

Chapitre 3

Validation et comparaison

Sommaire

Nous présentons dans cette partie une étude des modèles de diffusion que nous avons proposés en comparaison à des modèles standards. Tout d'abord, nous montrons une étude de la qualité de la prédiction de la diffusion en utilisant des jeux de données réels. Nous étudions ces résultats dans des contextes de diffusion différents. Les modèles que nous utilisons pour prédire les diffusions sont les suivants :

- Le modèle à cascades indépendantes (IC). Ses paramètres sont appris en utilisant un algorithme EM proposé dans [Saito et al., 2008] ;
- Le modèle asynchrone à cascades indépendantes (ASIC) qui est décrit dans [Saito et al., 2009] qui est une version asynchrone du modèle IC. Ses paramètres sont aussi appris en utilisant un algorithme EM ;
- Le modèle Netrate présenté dans [Gomez-Rodriguez et al., 2011] en utilisant la distribution exponentielle ;
- Le modèle UC, la variante que nous proposons du modèle IC qui tient compte des caractéristiques des utilisateurs, présenté dans le chapitre 2 ;
- Les modèles RUC et DRUC présentés dans le chapitre 2. Dans cette thèse, nous avons arbitrairement fixé le paramètre de délai δ à 0.9, ce qui correspond à une faible décroissance de l'influence des voisins au cours du temps.

3.1 Jeux de données

Dans cette première partie, nous expliquons les jeux de données que nous avons utilisés.

3.1.1 Sources

Nos jeux de données sont tirées de trois sources différentes :

- **ICWSM** ([Burton et al., 2009]), un jeu de données de billets de blogs utilisé lors du *data challenge* de la conférence ICWSM 2009. Les données ont été récupéré en utilisant l'outil spinn3r.
- **Memetracker** [Leskovec et al., 2009], qui est aussi un jeu de données de billets de blogs construit par l'université de Stanford.
- Des **jeux artificiels** générés par Ludovic Denoyer au laboratoire d'Informatique de Paris 6.

Pour les deux jeux de données de billets de blogs, nous avons considéré un blog comme étant un utilisateur du système. C'est un raccourci dans le sens où plusieurs personnes peuvent poster au sein d'un même blog. Dans les faits, mis à part quelques blogs particuliers, ils ne sont tenus et entretenus que par une seul personne réelle. Dans cette communauté, les bloggeurs donnent souvent les sources de leurs informations et de leurs motivations en donnant des liens vers ces sources. C'est grâce à ces liens que nous avons traqué la diffusion au sein de ces réseaux. Si un billet p_2 d'un blog b_2 cite un billet p_1 d'un autre blog b_1, nous considérons que b_2 a diffusé une information qui provenait du blog b_1. Nous inférons les liens entre utilisateurs en utilisant ces diffusions : si un blog b_2 a rediffusé au moins une information d'un blog b_1, on considère qu'il existe un lien de b_1 vers b_2 dans le réseau. Nous ne tenons pas compte des liens d'un blog vers lui même. Si un bloggeur se cite lui même, il n'y a pas eu de diffusion. Enfin, nous considérons la diffusion d'un contenu (ou cascade) comme un ensemble de billets tous reliés par des citations. Il s'agit

d'un sous-graphe du réseau complet des utilisateurs. Un billet qui n'est cité par personne et ne cite personne forme une cascade à lui tout seul. Il s'agit d'un contenu qui ne s'est pas diffusé. Nous transformons chaque billet de blog en un sac de mots et les traitons ensuite comme des vecteurs d'occurrences.

Pour l'ensemble de billets de blogs Memetracker, le blog auquel appartient chaque billet n'est pas mentionné. Nous avons donc inféré ce blog en utilisant l'url du billet : l'url du blog auquel appartient un billet est incluse dans l'url du billet. Pour ce faire, nous avons coupé l'url du billet au premier caractère "/" après "http ://". Il existe des plates-formes d'hébergement de blogs pour lesquels cette méthode ne donne pas l'adresse du blog mais de la plate-forme. Pour les blogs ainsi obtenus qui ont plus de 5000 billets en moins d'un mois, nous avons coupé l'url du billet au second caractère "/" afin d'éviter ces cas particuliers.

Une autre caractéristique du jeu de données Memetracker est que le contenu des blogs est représenté par des *memes*. Il s'agit d'expressions qui sont souvent retrouvées dans les billets de blogs. Cela a pour conséquence que le contenu utilisé pour le jeu de données Memetracker est moins complet que celui du jeu de données ICWSM.

D'un point de vu technique nous avons :

- un utilisateur correspond à un blog ;
- le profil d'un utilisateur est le vecteur moyenne des vecteurs de caractéristiques définissant tous les billets qu'il a posté. Il s'agit d'un vecteur de même taille que ceux représentants les billets ;
- un contenu est calculé comme le vecteur moyenne des vecteurs de caractéristiques définissant tous les billets qui le composent. Il s'agit encore d'un vecteur de même taille que les précédents.

Les pré-traitements suivants ont été effectués :

- nous n'avons gardé que les billets émis durant un mois ;
- nous n'avons gardé que les billets en anglais ;
- filtrage des mots vides en utilisant une liste de mots vides : il s'agit des mots qui n'apportent pas d'information au contenu tels que les déterminants, etc ;
- utilisation de la méthode de stemming de Porter pour lemmatiser tous les mots : l'idée est de ne garder qu'une forme basique de chaque mot afin de repérer plus facilement les mots identiques. Les pluriels, les temps de conjugaison, etc sont tous supprimés ;
- filtrage des mots apparaissant dans moins de cinq documents. Si un mot n'apparaît que très rarement il ne nous sera pas utile pour comparer les documents.

Nous avons extrait de chaque jeu de blogs trois types de jeux de données (décrits dans les sections 3.1.2 et 3.1.3) lors de nos expériences.

Jeux	# utilisateurs	# liens	# termes	# cascades	Taille moyenne
creux-meme-ident	29 177	8 501	69 355	104 978	1,008 (2,53)
creux-meme-aleat	39 427	10 816	70 602	104 973	1,01 (3,03)
creux-icwsm-ident	29 862	46 803	253 735	104 984	1,02 (2,52)
creux-icwsm-aleat	40 268	62 657	262 290	104 980	1,02 (2,61)

TABLE 3.1 – Statistiques sur les jeux de données creux. La dernière valeur est la taille moyenne des cascades ; nous donnons entre parenthèses la valeur en ne prenant en compte que les cascades de taille plus grande que 1.

3.1.2 Jeux aléatoires

Le but de ces jeux est de représenter les réseaux de blogs, en gardant les mêmes propriétés de diffusion. Pour ce faire, nous avons choisi aléatoirement 100000 cascades. Dans ce contexte, il y a beaucoup de contenus qui ne se diffusent pas (environ 2% des cascades contiennent au moins deux billets). Nous nous trouvons ainsi dans un cas où les modèles n'ont que peu d'information pour apprendre leurs paramètres. Nous avons séparé ces jeux de données en deux catégories :

- ceux que nous appelons **aleat**, pour lesquels les 100000 cascades sont tirés aléatoirement ;
- ceux que nous appelons **ident**, pour lesquels les utilisateurs apparaissant dans les cascades du jeu de test apparaissent obligatoirement dans au moins une cascade du jeu d'entraînement.

Nous nommons ces jeux de données en fonction de leurs propriétés :

- **creux-icwsm-aleat**
- **creux-icwsm-ident**
- **creux-meme-aleat**
- **creux-meme-ident**

Le tableau 3.1 montre les informations principales concernant ces jeux de données aléatoires.

3.1.3 Jeux pour la diffusion dense

Les jeux de données proposés précédemment font état d'une faible diffusion. Afin de tester les modèles dans un contexte où il y a plus de diffusion, nous proposons deux autres jeux de données tirés des mêmes ensembles de billets de blogs. Nous ne considérons ici que les 5000 utilisateurs les plus actifs du réseau (ceux qui ont diffusé le plus de contenus). Le jeu de données est ensuite formé de tous les contenus qui se diffusent entre ces utilisateurs les plus actifs.

De même que pour les jeux de données creux, nous nommons ces jeux de données en fonction de leurs propriétés :

- **dense-icwsm**
- **dense-meme**

Jeux	# utilisateurs	# liens	# termes	# cascades	Taille moyenne
dense-meme	5 000	4 373	24 482	2 977	1,1
dense-icwsm	5 000	17 746	173 014	23 738	1,075

TABLE 3.2 – Statistiques sur les jeux de données denses.

Jeux	# utilisateurs	# liens	# termes	# cascades	Taille moyenne
art-nosim	1 461	5 484	0	938	6,07
art-sim	1 461	5 484	1 000	999	86,11

TABLE 3.3 – Statistiques sur les jeux de données artificiels.

Le tableau 3.2 montre les informations principales concernant ces jeux de données denses.

3.1.4 Jeux artificiels

Les jeux dense contiennent plus de diffusion que les jeux creux mais ne permettent toujours pas de considérer la qualité des modèles dans un contexte de très forte diffusion. C'est pour cela que nous avons fait une étude sur des jeux de données générés.

Des cascades ont étés générées sur des graphes réels de réseaux sociaux en utilisant des modèles à cascades simples. Nous utilisons dans cette thèse deux jeux de données artificiels :

- **art-nosim**
- **art-sim**

Le premier jeu de données correspond à des cascades générés à partir de probabilités tirées aléatoirement. Il n'y a donc pas de contenu dans ce jeu. Pour le second jeu de données, les profils utilisateurs ainsi que les contenus des informations diffusées ont étés générés aléatoirement. Les probabilités du modèle à cascades qui a été utilisé sont liées à la similarité entre les profils des utilisateurs et les contenus. Ces deux jeux nous permettent, en plus de tester les différents modèles dans un cadre de très grande diffusion, de comparer les résultats avec et sans l'utilisation de contenu.

Le tableau 3.3 montre les informations principales concernant ces jeux de données artificiels.

3.1.5 Partage des données : entraînement et évaluation

Mis à part les jeux de données **creux** pour lesquels le partage a été fait lors de la création du jeu dû aux contraintes d'utilisateurs similaires, nous avons partagé les jeux de données en cinq parties bien distinctes. Les tests de prédiction de diffusion sont ensuite réalisés selon une validation croisée : quatre parties sont utilisées pour l'apprentissage des paramètres des différents modèles et la dernière sert au test de la prédiction des modèles.

Au cours des expériences chaque partie sert d'ensemble de test successivement et nous calculons la moyenne sur les cinq expériences.

3.2 Mesures d'évaluation

Afin de comparer la qualité des diffusions prédites par les différents modèles, nous utilisons trois mesures d'évaluation standards : des courbes de précision/rappel, la précision moyenne et l'erreur relative de volume.

3.2.1 Courbes Précision/Rappel

Pour chaque cascade, on classe les utilisateurs par leur probabilité d'être actif prédite par le modèle (l'utilisateur avec la plus forte probabilité est en première position). On calcule ensuite la précision à chaque point de rappel. Un point de rappel correspond à un utilisateur qui est réellement diffuseur du contenu. On fait enfin la moyenne des précisions à chaque point de rappel sur toutes les cascades.

Il est à noter que toutes les cascades ne font pas la même taille (ne contiennent pas le même nombre d'utilisateurs diffuseurs). La conséquence directe est que la variance de la précision calculée sur les derniers points de rappel est beaucoup plus importante que celle calculée sur les premiers.

Cette mesure est proposé dans [Manning et al., 2008]. L'idée de classer les utilisateurs est également présente dans [Song et al., 2007].

3.2.2 Précision moyenne

Cette mesure (MAP pour Mean Average Precision) fait ressortir la qualité globale d'un modèle sur tous les points de rappel. Elle est calculée de la façon suivante :

$$MAP = \frac{1}{K-\ell} \sum_{k=\ell+1}^{K} \frac{1}{|C^k(T^k)|} \sum_{n_i \in C^k(T^k)} Precision(R_{k,i})$$

où $R_{k,i}$ est l'ensemble des utilisateurs qui ont une probabilité d'être diffuseur supérieure ou égale à celle de l'utilisateur n_i (à noter que $n_i \in R_{k,i}$), qu'ils soient réellement diffuseurs ou juste prédits diffuseurs par le modèle.

3.2.3 Erreur relative de volume

Cette mesure (RVE pour Relative volume error) a été introduite dans [Yang and Leskovec, 2010]. Elle cherche à mesurer dans quelle mesure un modèle peut prédire le bon nombre d'utilisateurs touchés par la diffusion d'un contenu. En d'autres mots, est-ce qu'après un temps donné, le nombre de diffuseurs prédits par le modèle est proche du nombre réel de diffuseurs. C'est une mesure globale qui capture la propension d'un modèle à diffuser, que les

utilisateurs prédits soient les bons ou pas. Cette erreur est calculée comme la différence entre le nombre de diffuseurs prédit par le modèle et le nombre réel, moyenné sur toutes les cascades :

$$RVE = \frac{\sqrt{\sum_{k=\ell+1}^{K}(V_k(T^k) - \hat{V}_k(T^k))^2}}{\sqrt{\sum_{k=\ell+1}^{K} V_k(T^k)^2}}$$

où $V_k(t)$ représente le nombre réel d'utilisateurs diffuseurs au temps t et $\hat{V}_k(t)$ représente le volume d'utilisateurs diffuseurs prédits par le modèle au temps t. Pour le modèle RUC, on a :

$$\hat{V}_k(t) = \sum_{n_i \in \mathcal{N}} P(n_i, c^k, t)$$

3.3 Exécution des modèles à cascades dans une optique probabiliste

Les modèles à cascades nous donnent, lors de leur exécution une instance probable de la diffusion qu'ils engendrent. Afin d'obtenir les probabilités d'être actif de tous les utilisateurs après une diffusion (ce qui correspond à l'ensemble des instances de diffusions possibles), plusieurs méthodes sont possibles :

- faire une moyenne sur un grand nombre d'instances de diffusion
- utiliser la percolation de liens
- calculer les probabilités sur tous les chemins de diffusion possibles

La méthode qui consiste à calculer la moyenne sur un grand nombre d'instances de diffusion ne permet pas d'obtenir les probabilités de diffusion exactes. De plus elle est relativement coûteuse car elle nécessite d'exécuter un grand nombre de diffusions. La percolation de lien donne, elle le bon résultat mais nécessite de calculer la diffusion sur toutes les percolations de liens du graphe social. Si le graphe possède n liens, il faudra effectuer 2^n diffusions. Ce n'est pas envisageable pour des graphes sociaux non triviaux. Il reste la méthode qui consiste à calculer les probabilités de diffusion en suivant tous les chemins possibles.

L'algorithme 3 présente cette méthode pour l'algorithme de diffusion IC. Elle donne les valeurs exactes des probabilités tout en gardant une complexité abordable. Le principe est le suivant : on suit tous les chemins de diffusion possible en évitant les cycles (on ne veut pas qu'un même utilisateur apparaisse deux fois dans le même chemin de diffusion). En effet, un utilisateur appartenant à un chemin de diffusion est considéré comme déjà actif dans le chemin. Sinon il n'aurait pas pu activer ses successeurs. Il ne peut donc pas être activé à nouveau. Ensuite, la probabilité qu'un utilisateur ai été activé par un chemin donné est déterminée par les probabilités d'activation de tous ses prédécesseurs par le même chemin. Une subtilité subsiste : un utilisateur ne peut avoir été activé par un chemin que s'il n'a pas déjà été activé par un autre chemin. On calcule donc sa probabilité

Algorithme 3: Exécution des modèles à cascades. $prob(c)$ est la probabilité sur le chemin c, c'est à dire la multiplication de toutes les probabilités pour passer du premier utilisateur de c au dernier.

Données : Un graphe $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
Des probabilités de diffusion : $\forall (n_i, n_j) \in \mathcal{E}, p_{i,j}$
Un ensemble d'utilisateurs diffuseurs $Initiaux$
Résultat : Probabilité d'être actif pour l'ensemble des utilisateur du réseau
$Chemins \leftarrow \emptyset$;
$Etat \leftarrow$ Tableau($|\mathcal{V}|$);
pour tous les $u \in \mathcal{V}$ **faire**
  si $u \in Initiaux$ **alors**
    $Etat[u] = 0$;
    $Chemins \leftarrow Chemins \cup [u]$;
  sinon
    $Etat[u] = 1$;
  fin
fin
répéter
  $c \leftarrow Chemins.pop()$;
  $u \leftarrow c.dernier()$;
  pour tous les $v \in F(u)$ **faire**
    si $v \notin Initiaux$ *et* $v \notin c$ **alors**
      $Etat[v] \leftarrow Etat[v] * (1 - prob(c + v))$;
      $Chemins \leftarrow Chemins \cup c + v$;
    fin
  fin
jusqu'à $Chemins = \emptyset$;
pour tous les $u \in \mathcal{V}$ **faire**
  $Etat[u] = 1\text{-}Etat[u]$;
fin

d'être actif à la fin de la diffusion comme la probabilité qu'il n'ait pas été activé par aucun chemin.

3.4 Classement des utilisateurs : Précision/Rappel

3.4.1 Contexte réel : peu de diffusion

Dans un premier temps, nous observons les résultats de prédiction des modèles sur les jeux de données creux. Ces jeux sont ceux qui ont les propriétés les plus proches de l'ensemble de blogs d'origine. Il n'y a que très peu de diffusion dans ces jeux de données : seulement 2% des cascades impliquent au moins un utilisateur différent du diffuseur initial.

Cascades tirés aléatoirement

La figure 3.1 montre la précision aux différents points de rappel des six modèles sur le jeu de données **creux-meme-aleat**. Les modèles à cascades s'écroulent et ne parviennent pas à différencier correctement les utilisateurs diffuseurs des utilisateurs non diffuseurs. Les modèles centrés utilisateurs de leur coté obtiennent de meilleurs résultats. Seul le modèle UC obtient un résultat assez faible mais cependant un peu meilleur que les autres modèles à cascades.

La figure 3.2 montre la précision aux différents points de rappel des six modèles sur le jeu de données **creux-icwsm-aleat**. On retrouve le même schéma que pour le jeu de données de Memetracker. On voit cependant une nette amélioration des résultats pour le modèle UC, même s'il est toujours en dessous des modèles avec renforcement. Les modèles avec renforcement ont eux aussi une précision plus grande sur le jeu ICWSM par rapport au jeu Memetracker.

On observe sur ces jeux de données creux aléatoires des résultats bien supérieurs de la part des modèles centrés utilisateur. La raison principale de ces résultats vient du manque de diffusion. En effet, le peu de diffusion dans ces réseaux ne permet absolument pas aux modèles à cascades d'apprendre correctement tous leurs paramètres. De leur coté, les modèles centrés utilisateur n'ont que quatre paramètres à apprendre. De plus, certaines cascades des jeux d'évaluation impliquent des utilisateurs qui ne sont pas présents dans les jeux d'entraînement. N'ayant aucune information sur ces utilisateurs, les modèles à cascades standards ne peuvent pas estimer les paramètres correspondants.

Cascades de test avec seulement des utilisateurs du jeux d'apprentissage

Afin de tenter de résoudre le problème des utilisateurs que les modèles à cascades n'ont jamais vu, nous proposons une étude sur deux nouveaux jeux de données que nous avons biaisé en choisissant les cascades des jeux d'évaluation de sorte qu'elles n'impliquent que des utilisateurs déjà présents dans les jeux d'entraînement. Cela ne résout pas le problème du grand nombre de paramètres à estimer à partir de peu de données mais supprime le problème des nouveaux utilisateurs.

La figure 3.3 montre la précision aux différents points de rappel des six modèles sur le jeu de données **creux-meme-ident**. Sur ce jeu, seuls les deux modèles avec renforcement RUC et DRUC parviennent à différencier les utilisateurs diffuseurs des non diffuseurs.

La figure 3.4 montre la précision aux différents points de rappel des six modèles sur le jeu de données **creux-icwsm-ident**. On voit clairement que tous les modèles à cascades obtiennent de meilleurs résultats que sur les jeux de données aléatoires. Les modèles avec renforcement sont toujours au dessus, mais le modèle IC parvient néanmoins à obtenir des résultats similaires au modèle UC. Le fait de ne pas avoir de nouveaux utilisateurs permet au modèle IC d'estimer au moins aussi bien que le modèle UC les interactions

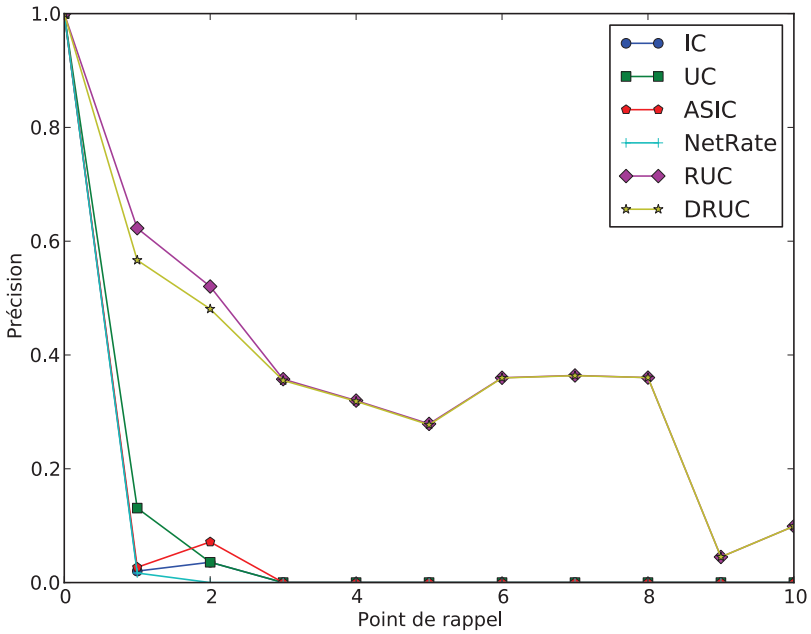


FIGURE 3.1 – Courbes de précision sur le jeu de données **creux-meme-aleat**.

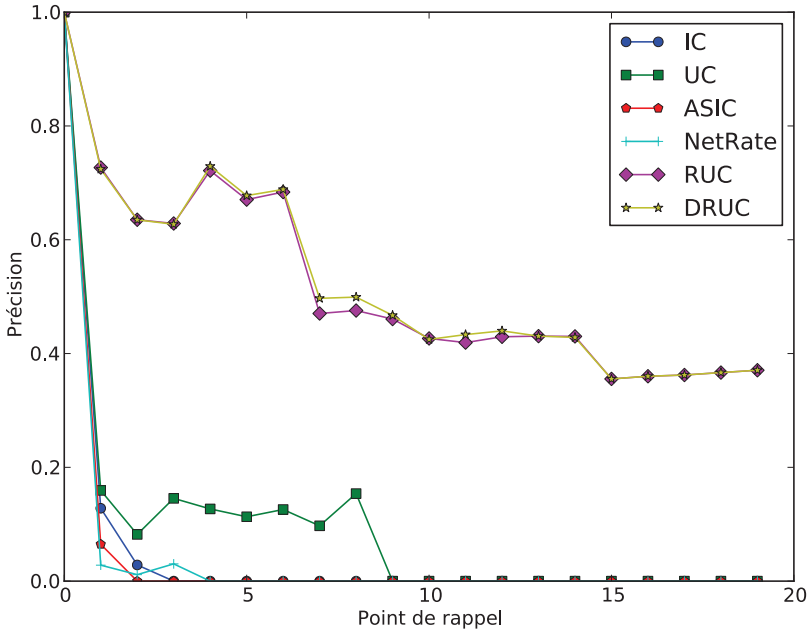


FIGURE 3.2 – Courbes de précision sur le jeu de données **creux-icwsm-aleat**.

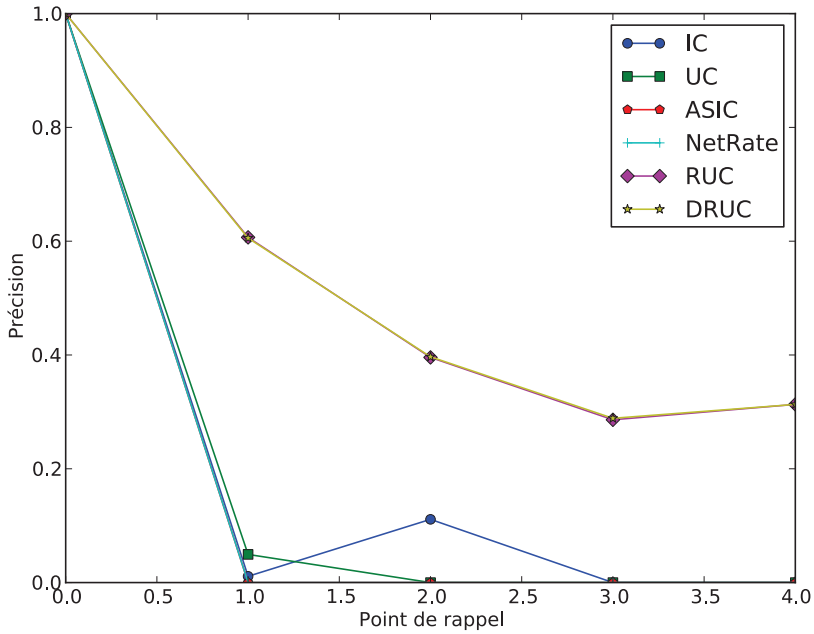


FIGURE 3.3 – Courbes de précision sur le jeu de données **creux-meme-ident**.

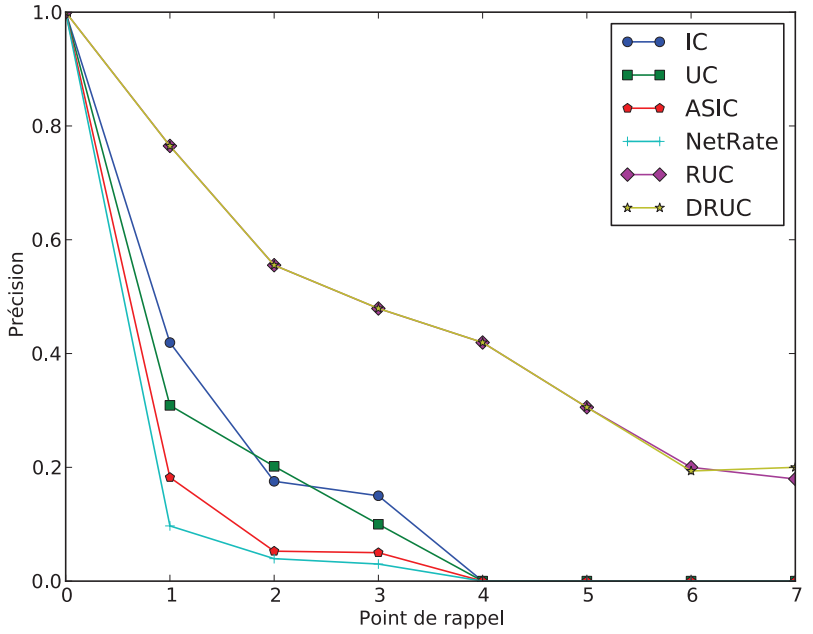


FIGURE 3.4 – Courbes de précision sur le jeu de données **creux-icwsm-ident**.

entre utilisateurs malgré le grand nombre de ses paramètres.

On remarque que dans tous ces jeux de données creux, les modèles asynchrones ASIC et Netrate obtiennent de très mauvais résultats. Le peu de diffusion dans les réseaux pousse les modèles vers la non diffusion. Les modèles asynchrones sont ainsi dirigés par le terme de décroissance qu'ils incorporent. Nous donnons plus d'explications sur ce phénomène dans la section 3.7.

3.4.2 Diffusion dense

Après avoir évalué les différents modèles dans un contexte de peu de diffusion, nous avons restreint ces jeux de données denses aux 5 000 utilisateurs les plus actifs des deux réseaux. Nous entendons par utilisateurs actifs les utilisateurs qui diffusent le plus de contenus. Le but de cette manipulation est d'obtenir un ensemble de cascades pour lequel on retrouve plus de diffusion que dans les jeux creux.

La figure 3.5 montre la précision aux différents points de rappel des six modèles sur le jeu de données **dense-meme**. Les modèles centrés utilisateurs obtiennent des résultats plus de deux fois supérieurs à ceux des modèles à cascades standards. Les modèles asynchrones parviennent cependant mieux à différencier les utilisateurs diffuseurs des non diffuseurs par rapport aux jeux de données creux. Le fait que l'on observe plus de diffusion, leur permet de moins se noyer dans la non diffusion. Plus de diffusion signifie aussi plus de contenus pour décrire les profils utilisateurs ce qui permet aussi aux modèles centrés utilisateurs d'obtenir de meilleurs résultats.

La figure 3.6 montre la précision aux différents points de rappel des six modèles sur le jeu de données **dense-icwsm**. On peut faire des observations similaires sur le jeu dense ICWSM que sur celui issu de Memetracker. Deux modèles ont cependant des résultats un peu différents. Quand le modèle IC obtient de meilleurs résultats, le modèle Netrate ne parvient pas du tout à distinguer les utilisateurs diffuseurs des autres.

Le jeu de données ICWSM contient plus de cascades que le jeu de données Memetracker. Ce phénomène permet aux modèles IC et ASIC de mieux estimer leurs paramètres grâce à un plus grand nombre de cascades d'entraînement. De plus, les contenus du jeu de données ICWSM sont plus complets car ils consistent en l'intégralité des billets de blogs plutôt qu'une liste de *memes*. Cela permet aux modèles centrés utilisateurs qui prennent en compte la similarité entre les profils des utilisateurs et le contenu d'obtenir des résultats meilleurs sur ICWSM plutôt que sur Memetracker.

3.4.3 Jeux Artificiels : diffusion très importante

Les jeux de données artificiels nous permettent d'évaluer deux choses. Tout d'abord, un des jeux de données ne contient pas de contenu et surtout a été généré sans utiliser de contenu, ce qui veut dire que les cascades générées n'obéissent qu'à une loi de cascades

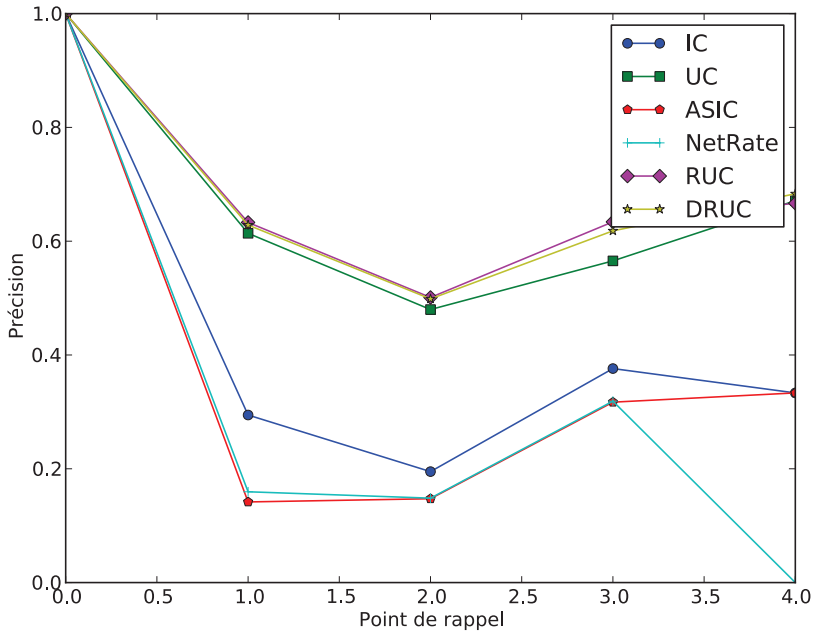


FIGURE 3.5 – Courbes de précision sur le jeu de données **dense-meme**.

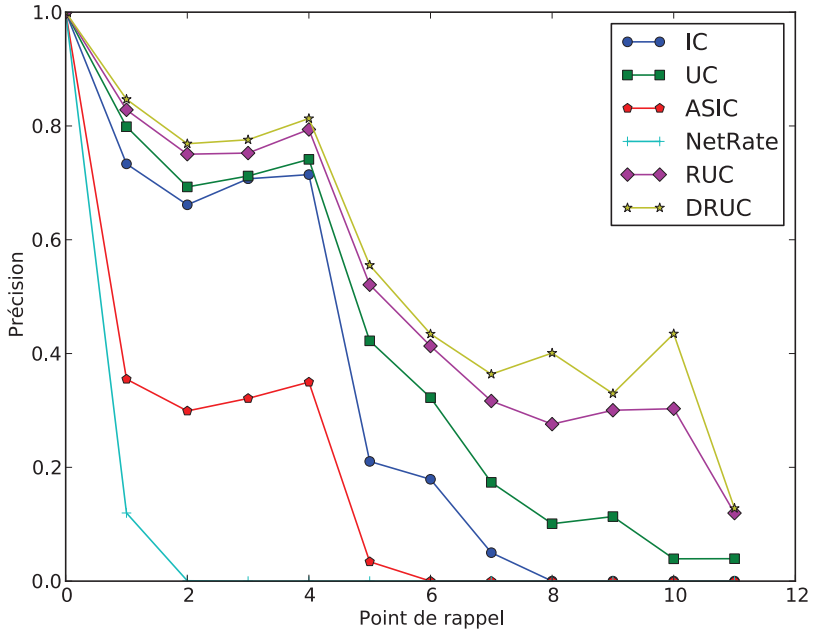


FIGURE 3.6 – Courbes de précision sur le jeu de données **dense-icwsm**.

standards. Le second jeu est par contre lui généré à partir des profils des utilisateurs, ce qui signifie que les modèles utilisant les contenus sont censés obtenir de meilleurs résultats. Ensuite, de façon artificielle, il est possible de créer des jeux de données dans lesquels la diffusion est beaucoup plus importante que dans les jeux de données réels.

La figure 3.7 montre la précision aux différents points de rappel des six modèles sur le jeu de données **art-nosim**. Tous les modèles ont clairement beaucoup de mal à différencier les utilisateurs diffuseurs des non diffuseurs. Les modèles centrés utilisateurs obtiennent néanmoins des précisions sensiblement supérieures à celles des modèles IC et ASIC. Le modèle Netrate obtient quant à lui de meilleurs résultats que tous les autres modèles.

La figure 3.8 montre la précision aux différents points de rappel des six modèles sur le jeu de données **art-sim**. La prise en compte des contenus par les modèles centrés utilisateur leur permet clairement d'obtenir des meilleurs résultats que les modèles à cascades standards IC et ASIC. Le modèle Netrate obtient ici aussi des résultats corrects.

Il y a deux points importants dans l'observation de ces résultats :

- Dans un jeu de données dont les cascades sont dirigées en partie par le contenu, les modèles centrés utilisateur obtiennent de très bons résultats comparés aux autres modèles. Dans le réseau pour lequel le contenu n'est pas pris en compte pour la génération des cascades, et dans un cadre où l'on observe beaucoup de diffusion, ils obtiennent des résultats similaires aux modèles standards.
- La présence de beaucoup de diffusion dans ces jeux de données entraîne de bons résultats pour le modèle Netrate. Il obtient même de très bons résultats sur le jeu de données **art-sim** pour lequel la diffusion est encore plus importante. En effet, comme nous l'avons dit précédemment ce modèle est dirigé par un terme de décroissance. Lorsqu'il y a peu de diffusion, il peine à différencier les diffuseurs des non diffuseurs, mais dans un cadre avec beaucoup de diffusion, l'estimation de ses paramètres est meilleure et il obtient donc de meilleurs résultats. Ce phénomène est ici aussi lié au terme de décroissance. Dans un contexte avec peu de diffusion, le modèle ne fait pas la différence entre les utilisateurs diffuseurs et les utilisateurs non diffuseurs.

3.4.4 Récapitulatif : Précision moyenne

Le tableau 3.4 donne la précision moyenne pour les six modèles sur l'ensemble des jeux de données. Elle correspond à un résumé des courbes que nous avons présenté précédemment. Sur les jeux de données creux, les modèles centrés utilisateur obtiennent des résultats clairement meilleurs que les modèles standards. Ceci est dû principalement au fait que les modèles standards ont beaucoup de paramètres à estimer et que dans ces jeux de données, le manque de diffusion ne permet pas de les estimer correctement. On voit cependant que ces modèles parviennent à obtenir une précision un peu plus élevée sur le jeu de données **creux-icwsm-ident**, dû au fait qu'il n'y a pas d'utilisateurs dans l'ensemble d'évaluation qui n'étaient impliqués dans aucune cascade de l'ensemble d'entraînement.

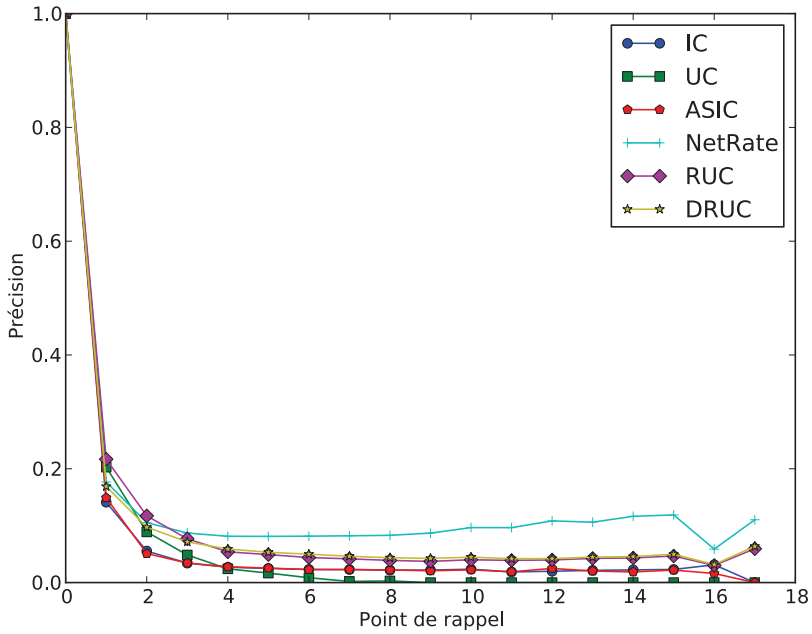


FIGURE 3.7 – Courbes de précision sur le jeu de données **art-nosim**.

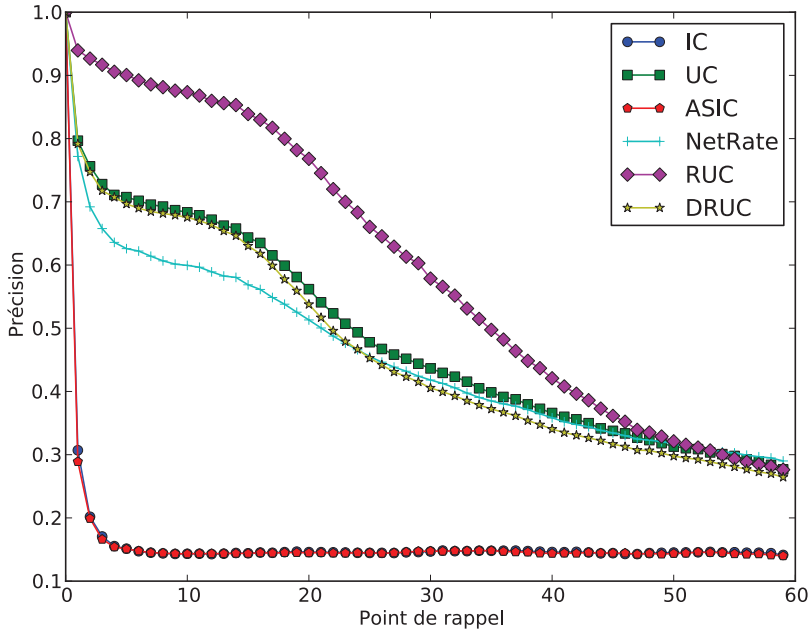


FIGURE 3.8 – Courbes de précision sur le jeu de données **art-sim**.

Jeu de données	IC	UC	ASIC	NetRate	RUC	DRUC
creux-meme-aleat	0.015	0.109	0.020	0.012	**0.606**	0.566
creux-icwsm-aleat	0.088	0.131	0.043	0.019	**0.710**	**0.707**
creux-meme-ident	0.010	0.043	0.000	0.000	**0.594**	**0.593**
creux-icwsm-ident	0.383	0.274	0.150	0.078	**0.743**	**0.743**
dense-meme	0.283	0.608	0.134	0.151	**0.627**	**0.622**
dense-icwsm	0.712	0.787	0.313	0.099	0.817	**0.836**
art-nosim	0.058	**0.100**	0.059	**0.101**	**0.099**	0.093
art-sim	0.167	0.446	0.166	0.363	**0.490**	0.397

TABLE 3.4 – Précision moyenne pour l'ensemble des modèles sur l'ensemble des jeux de données. Les valeurs en gras sont les meilleures.

Sur le jeu **dense-meme**, les modèles centrés utilisateurs sont clairement meilleurs que les autres. Sur le jeu tiré d'ICWSM en revanche les modèles à cascades standards obtiennent de meilleurs résultats. Ceci est dû au fait que le jeu ICWSM comprend plus de cascades que le jeu Memetracker. Il y a donc plus de données utilisées lors de l'estimation des paramètres. De plus, on peut voir une amélioration de la précision des modèles centrés utilisateur sur les jeux denses par rapport aux jeux creux. Dans les jeux denses, chaque utilisateur intervient dans plus de cascades, rendant ainsi leurs profils plus complets, et permettant aux modèles qui prennent en compte la similarité entre le contenu et les profils utilisateurs d'obtenir de meilleurs résultats.

Enfin, sur le jeu artificiel sans prise en compte des contenus, tous les modèles obtiennent des résultats similaires. Les modèles centrés utilisateurs restent toutefois légèrement meilleurs que les modèles IC et ASIC. Par contre, sur le jeu avec prise en compte des contenus, les modèles qui les utilisent pour faire leurs prédictions (UC, RUC et DRUC) obtiennent des résultats significativement meilleurs que les autres. On remarque néanmoins que le modèle Netrate obtient des bons résultats, comparé aux autres modèles, dû au grand nombre de diffusions. En effet, la structure du modèle l'empêche de faire la différence entre les utilisateurs diffuseurs et les utilisateurs non diffuseurs dans un contexte avec peu de diffusion. Ces jeux artificiels sont tout le contraire et ses paramètres sont ainsi mieux estimés.

3.5 Erreur de prédiction

Nous avons présenté la précision obtenue par les différents modèles sur les jeux de données. Les modèles centrés utilisateurs parviennent mieux à différencier les utilisateurs diffuseurs des utilisateurs non diffuseurs. Outre le fait de prédire les diffuseurs avec une probabilité plus forte que les non diffuseurs, un autre aspect de la diffusion est la capacité des modèles à prédire la bonne quantité de diffusion. En d'autres mots, être capable de savoir si un contenu va beaucoup se diffuser ou rester dans l'ombre.

Le tableau 3.5 montre l'erreur relative de volume pour les six modèles sur l'ensemble

Jeu de données	IC	UC	ASIC	NetRate	RUC	DRUC
creux-meme-aleat	**1.009**	**1.006**	**1.005**	1.041	1.036	1.040
creux-icwsm-aleat	0.996	**0.989**	1.000	1.012	**0.983**	0.994
creux-meme-ident	1.035	1.032	**1.013**	1.281	1.322	1.319
creux-icwsm-ident	**0.997**	1.765	1.009	1.079	1.585	1.444
dense-meme	0.916	1.391	0.954	**0.847**	1.412	1.203
dense-icwsm	0.756	6.259	0.893	0.959	0.901	**0.719**
art-nosim	3.513	**0.877**	3.418	3.878	1.121	0.982
art-sim	0.665	0.819	0.666	0.490	**0.246**	0.862

TABLE 3.5 – Erreur relative de volume pour l'ensemble des modèles sur l'ensemble des jeux de données. Un modèle simple qui ne diffuse pas aurait une erreur de 1.

des jeux de données. Un modèle qui diffuse parfaitement obtient une erreur relative de 0. Un modèle qui ne diffuse pas, et un modèle qui diffuse deux fois trop obtiennent une erreur relative de volume de 1. Sur les jeux de données creux, tous les modèles obtiennent une erreur proche de 1. Les modèles de cascades simples obtiennent cependant un résultat un peu meilleur par rapport aux modèles centrés utilisateurs. Le tableau 3.6 montre le volume de diffusion moyen prédit par les six modèles sur l'ensemble des jeux de données. Sur les jeux de données creux, les modèles centrés utilisateur ont tendance à diffuser plus que les modèles à cascades standards. En fait, le contexte de diffusion très faible fait que les modèles IC, ASIC et Netrate ne diffusent que très peu et obtiennent donc une erreur proche de 1. Au contraire, les modèles centrés utilisateurs ont tendance à diffuser trop.

Jeu de données	IC	UC	ASIC	NetRate	RUC	DRUC
creux-meme-aleat	0.002	0.003	0.001	0.004	0.003	0.004
creux-icwsm-aleat	0.002	0.005	0.001	0.002	0.010	0.009
creux-meme-ident	0.002	0.002	0.001	0.005	0.002	0.002
creux-icwsm-ident	0.002	0.010	0.001	0.003	0.012	0.010
dense-meme	0.09	0.85	0.05	0.30	0.57	0.72
dense-icwsm	0.43	1.81	0.21	0.12	0.59	0.81
art-nosim	23.99	0.70	23.45	26.25	5.48	0.10
art-sim	28.62	15.70	28.49	123.94	82.76	12.07

TABLE 3.6 – Volume de diffusion pour l'ensemble des modèles sur l'ensemble des jeux de données.

Pour ce qui est des jeux de données à diffusion dense, les modèles RUC et DRUC diffusent trop ce qui pénalise leur erreur sur le jeu Memetracker. Ils obtiennent tout de même de meilleurs résultats sur le jeu ICWSM. Le modèle UC quand à lui diffuse beaucoup trop entraînant une erreur assez importante.

En regardant les résultats sur les jeux artificiels à très forte diffusion, on voit que les modèles centrés utilisateur obtiennent de meilleurs résultats que les autres modèles. Si

l'on regarde le volume de diffusion prédit, on s'aperçoit que le modèle RUC est le seul à diffuser convenablement. L'erreur des modèles UC et DRUC avoisinant 1 est due à un manque de diffusion. Comme nous l'avons vu pour la précision, dans des contextes de diffusion importante, le modèle Netrate obtient de meilleurs résultats. C'est aussi le cas pour l'erreur de volume.

D'une manière générale, ces six modèles ont du mal à prédire le volume de diffusion quand celui-ci est trop faible. Les modèles IC, ASIC et Netrate tendent à ne pas diffuser du tout alors que les modèles UC, RUC et DRUC prédisent trop de diffusion. On voit cependant sur les jeux artificiels pour lesquels le volume de diffusion est très important que les modèles Netrate et RUC parviennent à obtenir de meilleurs résultats, voir même de très bons résultats en ce qui concerne le modèle centré utilisateur.

3.6 Valeurs des paramètres pour les modèles centrés utilisateur

Après avoir étudié les résultats des différents modèles dans une tâche de prédiction, on s'intéresse à l'importance des caractéristiques utilisateur (intérêt thématique, activité et pression sociale) prises en compte par les modèles centrés utilisateur. Nous ne présentons ici qu'une partie des valeurs. L'intégralité des valeurs de paramètres pour l'ensemble des jeux de données est disponible en annexe C. Certains paramètres sont estimés à la même valeur, ceci est dû à la méthode dérivée de la montée de gradient que nous utilisons (cf. section 2.6).

Modèle	UC	RUC	DRUC
λ_0 (biais)	-1.77	-8.23	-8.23
λ_1 (intérêt thématique)	6.8	9.52	10.52
λ_2 (activité)	9.1	3.69	2.36
λ_3 (pression sociale)	0.73	1.78	1.78

TABLE 3.7 – Valeurs des paramètres après apprentissage sur le jeu de données **creux-meme-ident**.

Le tableau 3.7 montre les valeurs estimées des paramètres des trois modèles centrés utilisateur pour le jeu de données **creux-meme-ident**. On retrouve le même schéma pour les autres jeux de données creux. Tout d'abord, on remarque que le paramètre de biais est toujours négatif, ce qui implique qu'il pousse les modèles vers la non diffusion. C'est d'autant plus vrai pour les modèles avec renforcement qui ont la propriété d'augmenter la probabilité de diffusion des utilisateurs à chaque étape de temps. Ce phénomène n'est pas étonnant dans le cadre des jeux de données creux pour lesquels le volume de diffusion est faible.

Les trois caractéristiques utilisateurs sont du même ordre de grandeur. Dans la théorie,

l'intérêt thématique et l'activité sont compris entre -1 et 1 alors que la pression sociale est strictement positive et correspond au nombre de voisins diffuseurs. Dans la pratique, la pression sociale est peu souvent supérieure à 1 car les utilisateurs n'ont que très rarement une probabilité de diffuser prédite par le modèle proche de 1. On peut donc comparer les valeurs estimées des trois paramètres.

Le paramètre de l'intérêt thématique est estimé plus important que les deux autres, montrant que la prise en compte du contenu et des profils utilisateurs est très important pour ces modèles. Le modèle UC utilise beaucoup plus l'activité que la pression sociale pour faire ses prédictions alors que les modèles RUC et DRUC prennent en compte la pression sociale à un niveau plus important.

Modèle	UC	RUC	DRUC
λ_0 (biais)	-2.36	-5.47	-3.33
λ_1 (intérêt thématique)	2.36	7.01	9.49
λ_2 (activité)	6.27	5.92	3.99
λ_3 (pression sociale)	1.19	2.78	0.95

TABLE 3.8 – Valeurs des paramètres après apprentissage sur le jeu de données **dense-meme**.

Le tableau 3.8 montre les valeurs estimées des paramètres des trois modèles centrés utilisateur pour le jeu de données **dense-meme**. On retrouve le même schéma pour le jeu de données **dense-icwsm**. Mis à part la valeur du paramètre du modèle UC correspondant à l'intérêt thématique qui est plus faible, l'importance des trois caractéristiques utilisateur reste similaire à celle observée pour les jeux de données creux. Cependant, la valeur absolue du biais est plus faible pour les modèles avec renforcement, surtout pour le modèle DRUC. Ces jeux de données contiennent plus de diffusion que les jeux creux, ce qui explique ce changement. De plus, le modèle DRUC possède un paramètre de diminution de l'influence des voisins au cours du temps impliquant une baisse de la probabilité de diffusion prédite par le modèle plus le temps passe, ce qui explique qu'il ait une valeur plus faible que le modèle RUC.

Modèle	UC	RUC	DRUC
λ_0 (biais)	-7.12	-6.23	-5.12
λ_1 (intérêt thématique)	-	-	-
λ_2 (activité)	7.12	6.23	6.23
λ_3 (pression sociale)	1.77	1.77	0.77

TABLE 3.9 – Valeurs des paramètres après apprentissage sur le jeu de données **art-nosim**.

Les tableaux 3.9 et 3.10 montrent les valeurs estimées des paramètres des trois modèles centrés utilisateur pour les jeux de données artificiels. Nous n'avons pas affiché de valeur pour les paramètre λ_1 sur le jeu artificiel sans contenu. En effet, la dérivée partielle de la vraisemblance par rapport à λ_1 est toujours égale à 0 lorsqu'il n'y a pas de contenu,

Modèle	UC	RUC	DRUC
λ_0 (biais)	-7.61	-4.89	-2.86
λ_1 (intérêt thématique)	7.61	4.89	5.11
λ_2 (activité)	0.36	3.78	2.86
λ_3 (pression sociale)	0.31	1.77	0.12

TABLE 3.10 – Valeurs des paramètres après apprentissage sur le jeu de données **art-sim**.

et la valeur de ce paramètre n'influence pas le modèle. On remarque encore une fois que l'intérêt thématique est fortement pris en compte quand il est disponible. Dans le jeu de données sans contenu, le paramètre correspondant à l'activité est plus important que pour le jeu de données avec contenu. La valeur du paramètre associé à la pression sociale est ici très faible pour les modèles UC et DRUC qui peinent à estimer correctement le volume de diffusion.

Les valeurs obtenues lors de l'estimation des paramètres des modèles centrés utilisateur nous montre l'importance de la prise en compte du contenu dans la modélisation de la diffusion par les modèles centrés utilisateur. D'autre part, lorsque le contenu n'est pas disponible, ces modèles utilisent principalement l'activité des utilisateurs afin de prédire les diffusions. La pression sociale est globalement peu utilisée pour modéliser la diffusion dans ces jeux de données de blogs. On ne voit une réelle influence de cette pression sociale que sur le jeu de données **art-sim** sur lequel RUC prédit assez bien le volume de diffusion.

3.7 Etude de la fonction de décroissance des modèles à cascades

Nous avons vu précédemment que sur plusieurs exemples les modèles à cascades basés sur le temps (ASIC et Netrate) obtiennent de moins bons résultats que le modèle simple IC. Nous en expliquons ici la raison en comparant les modèles IC et ASIC. Dans le modèle IC, lorsqu'un utilisateur n_i tente d'activer un de ses voisins n_j à une étape de temps t, il a une probabilité $p_{i,j}$ d'y parvenir et celui-ci sera actif à l'étape de temps qui suit $\tau = t + 1$. Dans le modèle ASIC, cette même probabilité existe mais l'étape à laquelle n_j va diffuser le contenu est déterminé par une distribution exponentielle de paramètre $r_{i,j}$. En clair il peut diffuser le contenu à n'importe quelle étape $\tau > t$, mais la probabilité de le faire décroît plus τ s'éloigne de t. Des détails supplémentaires sur ces deux méthodes sont présentes dans le chapitre 1.

Lors de l'estimation des paramètres de ces deux modèles, les valeurs de $p_{i,j}$ sont les mêmes. C'est donc l'ajout de la décroissance exponentielle qui influe sur la différence dans la diffusion.

La figure 3.9 montre deux distributions exponentielles de paramètres différents. L'étape de temps 0 correspond au moment ou l'utilisateur n_i tente d'activer l'utilisateur n_j, c'est

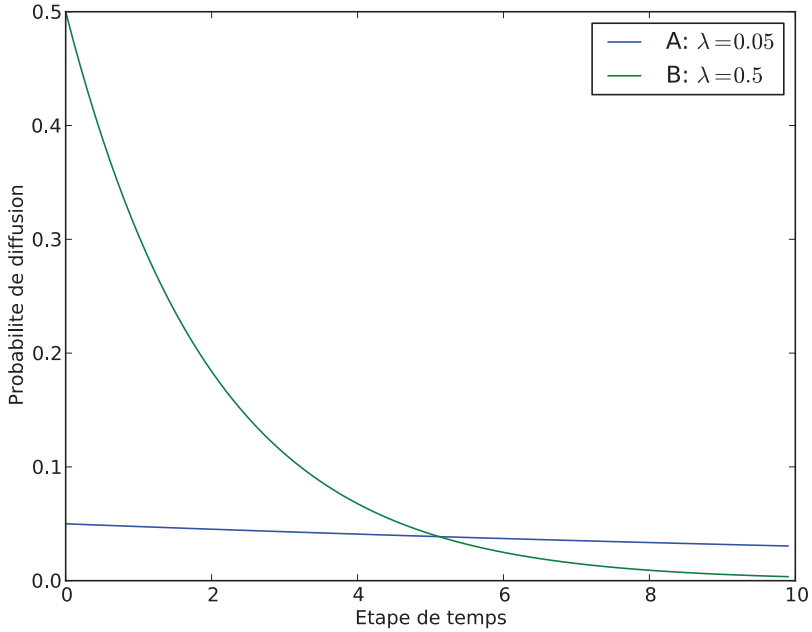


FIGURE 3.9 – Comparaison des fonctions de décroissance exponentielle pour le modèle ASIC. $f(x) = \lambda e^{-\lambda x}$.

à dire t. En supposant que la probabilité d'activer ait réussi (en utilisant $p_{i,j}$), il reste à déterminer quand l'utilisateur n_j sera actif en utilisant ces courbes. La probabilité qu'il soit actif à l'étape $t + \kappa$ est $\int_0^\kappa \lambda e^{-\lambda x} dx$. Dans toutes nos expériences, nous nous sommes limités à une fenêtre de temps. Prenons ici l'étape $t + 8$ comme la fin de notre fenêtre. En suivant ce raisonnement on s'aperçoit que pour la courbe B, la probabilité que n_j ait diffusé le contenu sera presque 1 alors que pour la courbe A, elle sera seulement de 0.3.

En d'autres mots, les modèles basé sur une décroissance exponentielle permettent de prédire une diffusion en dehors de la fenêtre de temps considéré. Cela a pour effet qu'une partie des utilisateurs sont déclarés comme non actifs à la fin de la diffusion mais auraient pu être déclarés actifs si notre fenêtre de temps avait été plus longue.

Un autre phénomène important est en rapport avec la quantité de diffusion observée. Dans un contexte de diffusion importante, le modèle ASIC va estimer un paramètre $r_{i,j}$ grand pour les liens par lesquels la diffusion est importante et faible pour les liens par lesquels peu d'information se propage. C'est encore plus vrai pour le modèle Netrate qui ne possède que ce paramètre. Dans un cadre où il y a peu de diffusions, tous les paramètres $r_{i,j}$ seront estimés faibles et les probabilités de diffusions seront toutes proches les unes des autres. C'est pourquoi ces modèles peinent à différencier les utilisateurs diffuseurs quand la quantité de diffusion observée est faible.

3.8 Modèles de regression "centrés utilisateur"

Lors d'un travail [Lagnier et al., 2013a] avec Ludovic Denoyer et Patrick Gallinari du Laboratoire d'Informatique de Paris 6, nous avons étendu l'utilisation des caractéristiques utilisateur. Ils ont en effet proposé des modèles de regression pour la prédiction de diffusion de contenus ayant pour paramètres ceux définis pour les modèles centrés utilisateur. Nous présentons ici les résultats du modèle DUC (Discriminative User Centric model).

Jeu de données	RUC	DRUC	DUC
creux-meme-aleat	0.606	0.566	**0.612**
creux-icwsm-aleat	**0.710**	**0.707**	0.701
dense-meme	**0.627**	**0.622**	0.581
dense-icwsm	0.817	**0.836**	0.820

TABLE 3.11 – Précision moyenne pour les modèles de regression. Les valeurs en gras sont les meilleures.

Le tableau 3.11 montre la précision moyenne du modèle DUC en comparaison de celle des modèles avec renforcement. Les résultats sont similaires : mis a part pour le jeu de données, **creux-meme-aleat**, les modèles avec renforcement obtiennent des résultats légèrement meilleurs.

Les modèles de regression attribuent pour chaque cascade un score de contamination à tous les utilisateurs. Ce score n'est en rien une probabilité d'être actif et il n'y a donc pas de transition direct vers le volume de contamination.

Jeu de données	dense-meme	dense-icwsm
λ_0 (biais)	-0.42	0.025
λ_1 (intérêt thématique)	2.47	7.9
λ_2 (activité)	-0.21	0.012
λ_3 (pression sociale)	-0.14	-0.66

TABLE 3.12 – Valeurs des paramètres après apprentissage pour les moddèles de regression.

Le tableau 3.12 montre les valeurs des paramètres associés aux caractéristiques utilisateur après apprentissage du modèles de regression DUC. Le paramètre associé à l'intérêt thématique est, comme pour les modèles probabilistes, bien plus grand que les autres, montrant l'importance de la prise en compte du contenu.

Pour finir, ces modèles de regressions ont deux grandes différences avec les modèles probabilistes proposés dans cette thèse :

- ils ne permettent pas de prédire la diffusion étape par étape à moins d'effectuer une regression par étape de temps. Ils ne modélisent pas le comportement des utilisateurs mais celui du système et prédisent ainsi directement un score final de diffusion ;
- le temps d'apprentissage est nettement moins important que celui des modèles avec renforcement car ils n'ont pas besoin d'apprendre le comportement des utilisateurs

étape par étape.

3.9 Conclusion

Malgré le fait que les six modèles (IC, UC, ASIC, Netrate, RUC et DRUC) n'arrivent globalement pas à prédire correctement le volume de diffusion, les modèles centrés utilisateurs ont de meilleurs capacités, quel que soit le contexte, pour classer les utilisateurs : ils prédisent une probabilité de diffusion plus importante pour les utilisateurs étant réellement diffuseurs. Ceci est dû à trois propriétés de ces modèles :

- ils reposent sur peu de paramètres, ce qui leur permet de mieux apprendre dans un contexte où il n'y a pas beaucoup de données disponibles :
- la prise en compte du contenu diffusé leur permet de mieux modéliser la propagation d'information dans un réseau social pour lequel les contenus sont importants ;
- le renforcement permet une modélisation plus fine de la diffusion en répartissant le choix d'un utilisateur de diffuser un contenu sur plusieurs étapes plutôt que sur une simple probabilité.

De plus, les paramètres des modèles centrés utilisateurs étant globaux et pas propres à chaque utilisateur, il est possible d'ajouter de nouveaux utilisateurs dans le réseau sans que cela pose de problème à ces modèles pour les incorporer. La connaissance acquise par l'expérience d'un utilisateur est retransmise aux autres.

Chapitre 4

Le problème de maximisation de l'influence

Sommaire

Nous avons commencé ce travail dans le cadre du stage de Master de François Kawala [Kawala, 2011].

4.1 Définition du problème

Le problème de maximisation de l'influence (IM) est définit sur un graphe social $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ pour un modèle de diffusion $\mathcal{M}$ et un nombre $\kappa \leq |\mathcal{V}|$. En partant du principe que κ utilisateurs du réseau diffusent une information, le but est de savoir lesquels entraîneront une diffusion maximale. Selon le modèle pour lequel on cherche à maximiser l'influence, nous aurons besoin de paramètres supplémentaires comme les profils des utilisateurs $\mathcal{P}$,

le contenu diffusé c^k, etc. On définit la fonction σ que l'on cherche à optimiser comme l'influence du contenu diffusé sur le réseau social. En d'autres mots il s'agit du nombre d'utilisateurs actifs à la fin d'une diffusion. Dans la pratique, un certain nombre de modèles étant stochastiques, elle correspond à l'espérance du nombre d'utilisateurs ayant diffusé le contenu :

$$\sigma_{\mathcal{M}}(e, \mathcal{G}) = E[|C^k(T^k)|] \tag{4.1}$$

où e est un ensemble de κ diffuseurs initiaux. Le problème de maximisation de l'influence est définit de la façon suivante :

$$IM_{\mathcal{M}}(\mathcal{G}, \kappa) = argmax_{e \subseteq \mathcal{V}, |e|=\kappa} \ \ \sigma_{\mathcal{M}}(e, \mathcal{G}) \tag{4.2}$$

Il s'agit de l'ensemble de κ utilisateurs qui entraînera une diffusion maximale au sein du réseau.

4.2 Un problème NP-difficile

Le problème de maximisation de l'influence, en utilisant les modèles IC et RUC, est un problème NP-difficile. Afin de prouver cette complexité, nous réduisons dans cette section le problème de couverture d'ensemble (SC), qui fait parti des 21 problèmes NP-complets de Karp ([Karp, 1972]), au problème de maximisation de l'influence.
Nous définissons, pour cette réduction, deux fonctions :

- Γ, qui transforme un problème de couverture d'ensemble x en un problème de maximisation de l'influence $\Gamma(x)$
- ζ, qui transforme une solution d'un problème de maximisation de l'influence en une solution d'un problème de couverture d'ensemble

La réduction est valide si ces fonctions ont une complexité polynomiale et si, pour un problème quelconque de couverture d'ensemble x et sa solution y, $\zeta(IM_{\mathcal{M}}(\Gamma(x)) = y$. On pourra alors en déduire que le problème de maximisation de l'influence en utilisant le modèle $\mathcal{M}$ est NP-difficile.

4.2.1 Définition du problème de couverture d'ensemble

Le problème de couverture d'ensemble consiste à trouver, à partir d'un ensemble d'ensembles, une couverture de l'univers en ne choisissant qu'un certain nombre de ces ensembles. De manière plus formelle, on possède un ensemble $\mathcal{C}$ de sous-ensembles de l'univers $\mathcal{U}$ et un nombre κ tel que $\kappa \leq |\mathcal{C}|$. Le problème de couverture d'ensemble cherche à trouver un famille $\mathcal{F}$ d'éléments de $\mathcal{C}$ tel que :

- $|\mathcal{F}| \leq \kappa$
- $\cup_{f \in \mathcal{F}}(f) = \mathcal{U}$

En d'autre termes, la famille $\mathcal{F}$ doit contenir au maximum κ éléments et couvrir l'univers. La figure 4.1 montre une illustration du problème de couverture d'ensemble.

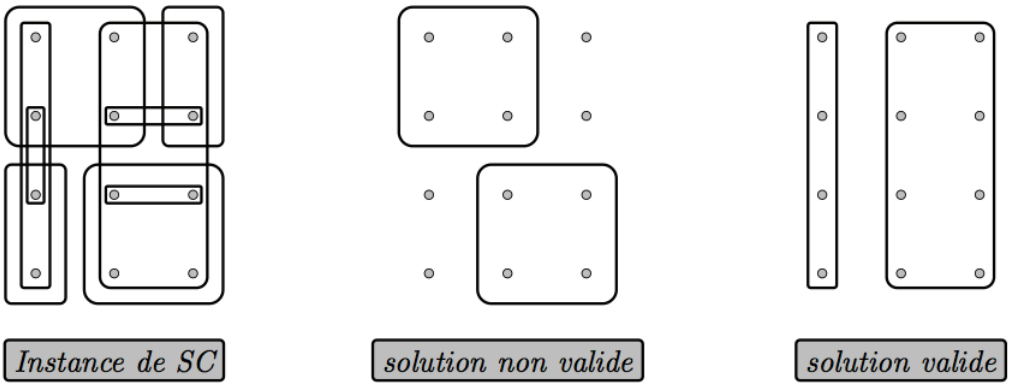


FIGURE 4.1 – Exemple d'instance de l'algorithme de couverture d'ensemble (SC). Dans cet exemple, $k = 2$.

4.2.2 Complexité du problème pour le modèle IC

Nous expliquons ici la réduction du problème de couverture d'ensemble vers le problème de maximisation de l'influence en utilisant le modèle IC. Cette réduction a été proposé par [Kempe et al., 2003]. L'idée est de trouver une instance particulière du problème de maximisation de l'influence qui donne un résultat transposable vers le problème de couverture d'ensemble.

Pour une instance du problème de couverture d'ensemble, on crée une instance du problème de maximisation de l'influence comme suit :

- Un graphe $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ avec deux types de nœuds :
  - Un nœud pour chaque élément de l'univers : $\forall e \in \mathcal{U}, n_e \in \mathcal{V}$
  - Un nœud pour chaque sous-ensemble de l'univers : $\forall c \in \mathcal{C}, n_c \in \mathcal{V}$
- Un lien qui part de chaque ensemble vers chacun des éléments qu'il contient : $\forall c \in \mathcal{C}, \forall e \in c, (n_c, n_e) \in \mathcal{E}$
- On place sur chaque lien une probabilité de diffusion de 1
- On cherche un ensemble de κ nœuds qui donnent une diffusion maximale

La fonction de transformation du problème est donc :

$$\Gamma(< \mathcal{U}, \mathcal{C}, \kappa >) =< \mathcal{G}, \kappa > \tag{4.3}$$

En créant cette instance, tous les nœuds sont soit atteignables en une seule étape, soit inatteignables. Une autre propriété intéressante est que comme $\kappa \leq |\mathcal{C}|$, les nœuds correspondants aux sous-ensembles de l'univers seront choisis comme diffuseurs initiaux lors de la maximisation de l'influence car ils sont les seuls à entraîner une réelle diffusion.

En résolvant le problème de maximisation de l'influence, on obtient le nombre de nœuds atteints par la diffusion. On peut ensuite répondre au problème de la couverture d'ensemble de la façon suivante : il est possible de couvrir l'univers avec seulement κ sous-ensembles si et seulement si l'ensemble de nœuds atteints par la diffusion regroupe tous les nœuds correspondants aux éléments de l'univers plus κ nœuds correspondants aux sous-ensembles de l'univers. Autrement dit, la fonction de transformation de la solution est la suivante :

$$\zeta(IM_{IC}(\Gamma(\mathcal{G},\kappa)) = IM_{IC}(\Gamma(\mathcal{G},\kappa) \geq |\mathcal{U}| + \kappa \tag{4.4}$$

Comme le problème de couverture d'ensemble est NP-complet, cette réduction prouve que le problème de maximisation de l'influence en utilisant le modèle IC est NP-difficile. La figure 4.2 montre un exemple de la réduction que nous venons de présenter. En choisissant les ensembles A et D (les utilisateurs n_A et n_D) on obtient une solution du problème qui n'est pas valide car l'élément 2 n'est pas capturé par l'union des deux ensembles (et donc l'utilisateur n_2 n'est pas atteint par la diffusion). Le choix des ensembles B et D permet d'englober tout l'univers dans l'union des deux ensembles (et par fortiori tous les utilisateurs correspondants à des éléments de l'univers sont atteints par la diffusion).

4.2.3 Complexité du problème pour le modèle RUC

Dans cette partie, nous prouvons que le problème de maximisation de l'influence en utilisant le modèle RUC est aussi NP-difficile. Contrairement au modèle IC pour lequel le flux de diffusion se fait par rapport aux probabilités associées aux liens du réseau, dans le modèle RUC chaque utilisateur à une probabilité qui lui est propre de s'activer à chaque étape de temps. Dans la démonstration précédente nous avons compté les utilisateurs atteignables par la diffusion afin de déterminer si oui ou non l'ensemble des utilisateurs correspondant aux éléments de l'univers avait été atteint. Afin de pouvoir reproduire un tel comportement, il faut définir une probabilité d'activation pour les utilisateurs qui ne dépend pas d'eux. C'est un peu contradictoire avec le but même de notre modèle mais cette caractéristique est capturée en choisissant tous les paramètres nuls : $\lambda_0 = \lambda_1 = \lambda_2 = \lambda_3 = 0$. Pour un utilisateur n_i quelconque, sa probabilité d'activation à une étape de temps t pour la diffusion d'un contenu c^k (que nous n'avons pas besoin de définir au vu du contexte) est :

$$P(n_i, c^k, t) = \begin{cases} \frac{1}{1+e^0} = \frac{1}{2} \text{ si } SP(n_i, \mathcal{G}, M^k, t) > 0 \\ \\ 0 \text{ sinon} \end{cases} \tag{4.5}$$

La fonction de transformation du problème Γ est ainsi définie de la même manière que pour le modèle IC, et les probabilités de diffusion sont définies par l'équation 4.5.

En résolvant le problème de maximisation de l'influence, on obtient une probabilité

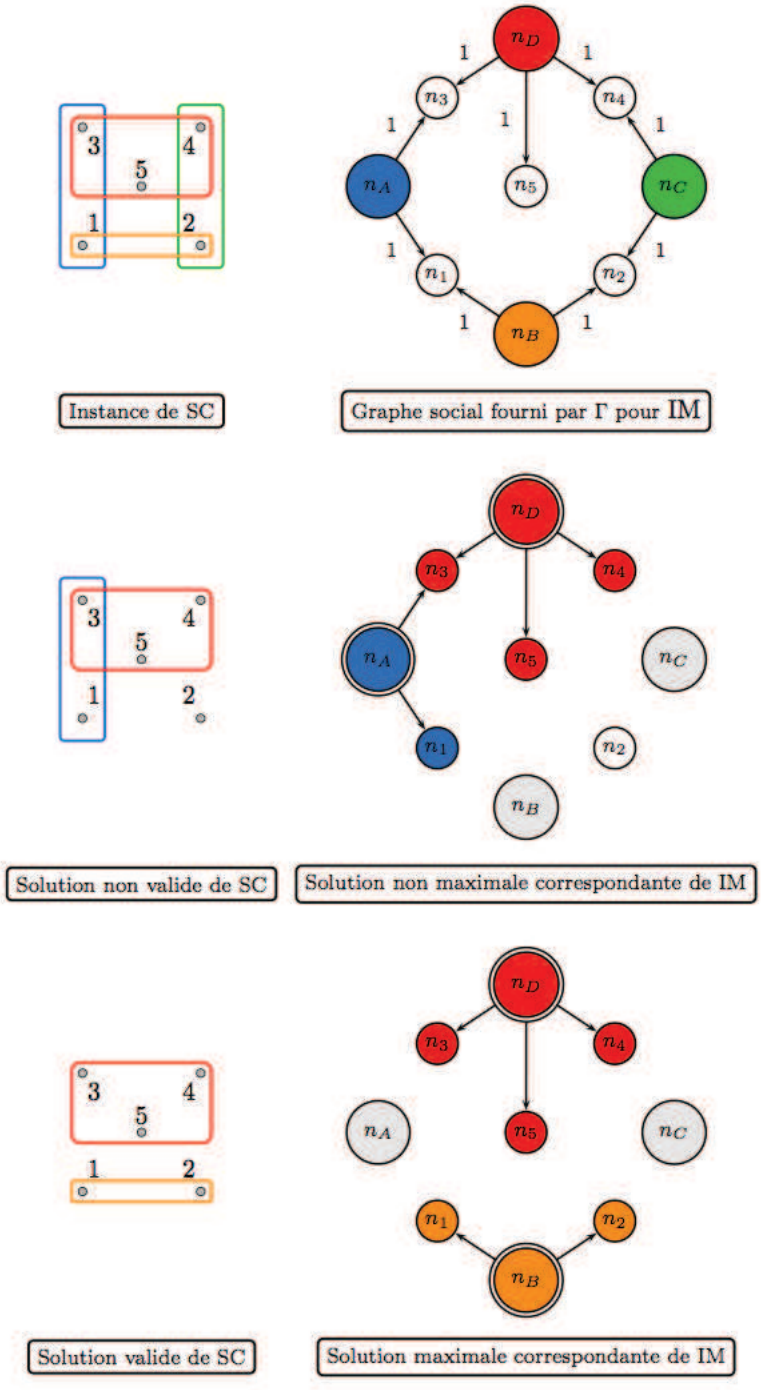


FIGURE 4.2 – Exemple de réduction du problème de couverture d'ensemble vers le problème de maximisation de l'influence.

d'être actif pour chacun des utilisateurs du réseau. Nous allons retrouver trois classes d'utilisateurs : ceux initiateurs de la diffusion, ceux ayant été atteints par la diffusion, et ceux n'ayant pas été atteints par la diffusion. Une propriété de l'instance du problème telle que nous l'avons défini est que pour tous les utilisateurs d'une même classe leur probabilité d'être actif a chaque étape de temps est la même :

- si n_i est initiateur : $P(n_i, c^k, \leq t) = 1$
- si n_i est atteignable : $P(n_i, c^k, \leq t) = 1 - \left(\frac{1}{2}\right)^t$
- si n_i est non atteignable : $P(n_i, c^k, \leq t) = 0$

Nous rappelons que le graphe ne contient que des chemins de longueur 1 et que donc un utilisateur est soit atteignable dès la première étape de temps, soit ne le sera jamais. Ce résultat est trivial pour les utilisateurs initiateurs et non atteignables mais demande un peu plus d'explications pour les utilisateurs atteignables. Dans le modèle RUC nous avons défini la probabilité d'être actif d'un utilisateur en fonction de sa probabilité d'être actif à l'étape précédente (équation 2.3). On peux ainsi montrer simplement la propriété suivante :

Propriété 1. *Lors de la diffusion d'un contenu c^k en utilisant le modèle RUC pour lequel tous les paramètres sont nuls, la probabilité d'être actif pour un utilisateur n_i directement relié à un diffuseur initial est la suivante :*

$$P(n_i, c^k, \leq t) = 1 - \left(\frac{1}{2}\right)^t$$

Démonstration. On réécrit l'équation 2.3 :

$$\begin{aligned} P(n_i, c^k, \leq t+1) \quad &= P(n_i, c^k, \leq t) + (1 - P(n_i, c^k, \leq t))P(n_i, c^k, t) \\ &= P(n_i, c^k, \leq t)(1 - P(n_i, c^k, t)) + P(n_i, c^k, t) \end{aligned}$$

Les probabilités d'un utilisateur d'être actif au cours du temps forment une suite arithmético-géométrique dont le premier terme est 0. Pour une telle suite définie de la forme
$\forall n \geq 0, u_{n+1} = au_n + b$, le calcul d'un terme général se fait de la façon suivante :
soit $r = \frac{b}{1-a}$
alors $\forall n \geq 0, u_n = a^n(-r) + r$
Dans notre cas, $a = b = \frac{1}{2}$ et $r = 1$. □

A partir de ces résultats, nous pouvons définir la fonction de transformation de la solution :

$$\zeta(IM_{RUC}(\Gamma(\mathcal{G}, \kappa)) = IM_{RUC}(\Gamma(\mathcal{G}, \kappa) \geq |\mathcal{U}| \times (1 - \left(\frac{1}{2}\right)^t) + \kappa \tag{4.6}$$

En effet, chaque utilisateur ayant été atteint aura une probabilité de $1 - \left(\frac{1}{2}\right)^t$ d'être actif

et les κ utilisateurs initiateurs auront une probabilité de 1 d'être actifs. Le problème de maximisation de l'influence en utilisant le modèle RUC est donc lui aussi NP-difficile.

4.3 Approximation du problème

Le problème de maximisation de l'influence étant NP-difficile, il n'est pas possible de le résoudre pour des instances non triviales. C'est pourquoi on va chercher à trouver des algorithmes d'approximation du résultat optimal. Nous présentons ici l'algorithme glouton "Greedy Hill Climbing" qui garantit une bonne approximation pour la plupart des modèles.

4.3.1 Algorithme glouton "Greedy Hill Climbing"

Algorithme 4: Algorithme "Greedy Hill Climbing" pour la maximisation de l'influence

Données : Un graphe $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
Un nombre d'initiateurs κ
Résultat : Ensemble des κ diffuseurs initiaux entraînant une diffusion maximale
$A \leftarrow \emptyset$
pour tous les $i = 1$ *à* κ **faire**
  pour tous les $n_j \in \mathcal{N} \setminus A$ **faire**
    $d = \sigma_{\mathcal{M}}(A \cup \{n_j\}, \mathcal{G}) - \sigma_{\mathcal{M}}(A, \mathcal{G})$
    si d *est maximal* **alors**
      $n_{max} = n_j$
    fin
  fin
  $A \leftarrow A \cup \{n_{max}\}$
fin

Le schéma d'exécution se trouve dans l'algorithme 4. Le principe est le suivant : si on ne peut pas trouver l'ensemble de κ initiateurs apportant une diffusion maximale alors on les choisit un par un. Ainsi, l'algorithme commence par choisir le meilleur initiateur. Ensuite, il choisit le second qui offre le meilleur gain marginal par rapport au premier utilisateur déjà choisi. L'algorithme continue ensuite jusqu'à avoir sélectionné un ensemble de κ utilisateurs. Cet algorithme offre une bonne approximation quand la fonction que l'on veut maximiser respecte certaines propriétés décrites dans le théorème suivante :

Théorème 4. *Pour une fonction non négative, monotone et sous-modulaire f, soit A un ensemble de κ utilisateurs obtenus par l'algorithme "Greedy Hill Climbing" maximisant la fonction f. Soit A^* l'ensemble qui maximise la valeur de f pour κ éléments. Alors $f(A) \geq (1 - 1/e)f(A^*)$, en d'autres mots, A est une $(1 - 1/e)$-approximation.*

Ce théorème à été utilisé dans le domaine de l'optimisations et la preuve se trouve dans [Nemhauser and Wolsey, 1988].

4.3.2 Maximisation en utilisant le modèle IC

La fonction d'influence σ_{IC} est non négative et monotone. Non négative car le nombre d'utilisateurs touchés par une diffusion est obligatoirement positif ou nul et monotone car strictement croissante par construction. Un utilisateur ayant diffusé une information ne peut pas revenir en arrière, le nombre de personnes touchées par la diffusion d'un contenu ne peut donc pas diminuer au cours du temps. Afin de pouvoir utiliser l'algorithme glouton présenté précédemment avec une garantie de résultat, il reste à montrer que la fonction d'influence est sous-modulaire.

Propriété 2. *Quelque soient les deux ensembles S et T tels que $S \subseteq T$ et un élément v tel que $v \notin T$, une fonction f est sous-modulaire si et seulement si elle satisfait l'équation suivante :*

$$f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)$$

En d'autres mots, la fonction f est sous-modulaire si le gain marginal pour l'ajout d'un nouvel élément à l'ensemble S est au moins aussi important que le gain marginal pour l'ajout du même élément à un sur-ensemble T.

Nous avons montré dans le chapitre 1 que le modèle IC était équivalent à une percolation de lien. Rappelons les détails d'une percolation de lien :

- Chaque lien du graphe d'origine peut être soit ouvert, soit fermé. Il existe donc une instance de graphe pour chaque ensemble de configurations possible des liens. On appelle $\mathcal{X}$ l'ensemble de ces configurations. Les probabilités de diffusion sur les liens ouverts sont de 1, elles sont de 0 sur les liens fermés.
- A chaque configuration du graphe $X \in \mathcal{X}$ correspond une probabilité d'apparition $Prob[X]$ telles que $\sum_{X \in \mathcal{X}} Prob[X] = 1$.

On appelle $\sigma_X(A)$ le nombre total d'utilisateurs touchés par une diffusion si l'ensemble initial d'utilisateurs actifs est A et que l'on se place dans la configuration X. Dans ce contexte, le calcul de $\sigma_X(A)$ est déterministe car les probabilités de diffusion sont soit 1 soit 0. Il est ainsi facile d'exprimer sa valeur. Soit $R(n_i, X)$ l'ensemble des utilisateurs atteignables depuis l'utilisateur v dans la configuration de graphe X. On peut exprimer la fonction d'influence comme suit :

$$\sigma_X(A) = |\cup_{n_i \in A} R(n_i, X)| \tag{4.7}$$

Si l'on considère maintenant les deux ensembles S et T tels que $S \subseteq T$ ainsi qu'un utilisateur n_j tel que $n_j \notin T$, la quantité $\sigma_X(S \cup \{n_j\}) - \sigma_X(S)$ correspond au nombre

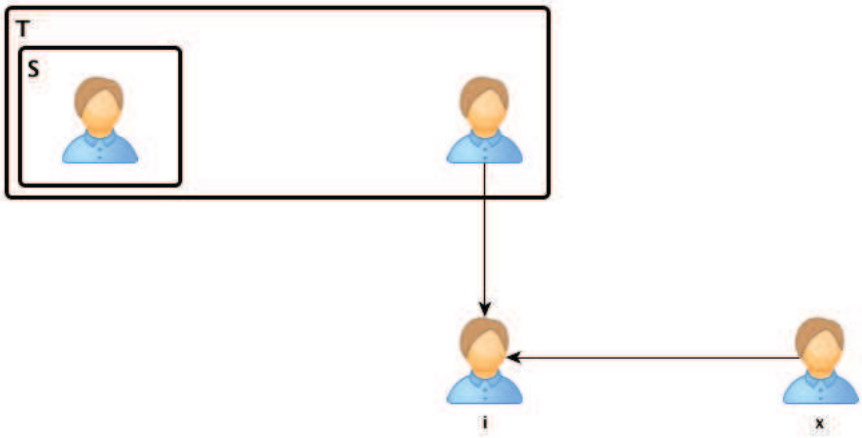


FIGURE 4.3 – Graphe social avec 4 utilisateurs et 2 ensembles d'utilisateurs S et T tels que $S \subseteq T$

d'éléments de $R(n_j, X)$ qui ne sont pas dans l'union $\cup_{n_i \in S} R(n_i, X)$. Cette valeur est au moins aussi grande que le nombre d'éléments de $R(n_j, X)$ qui ne sont pas dans la plus grande union $\cup_{n_i \in T} R(n_i, X)$. On a donc montré que la fonction σ_X est sous-modulaire :

$$\sigma_X(S \cup \{n_j\}) - \sigma_X(S) \geq \sigma_X(T \cup \{n_j\}) - f(T) \tag{4.8}$$

Enfin on a par l'équivalence entre le modèle IC et la percolation de lien :

$$\sigma_{IC}(A) = \sum_{X \in \mathcal{X}} Prob[X] \;\; \sigma_X(A) \tag{4.9}$$

Sachant qu'une combinaison linéaire de fonctions positives sous-modulaires est elle même une fonction sous-modulaire, nous venons de montrer que la fonction σ_{IC} est sous-modulaire. Il est ainsi possible d'obtenir une bonne approximation au problème de maximisation de l'influence pour le modèle IC en utilisant l'algorithme glouton "greedy hill climbing". Une autre approche a été proposé dans [Nguyen and Zheng, 2012] où les auteurs décomposent le problème en plusieurs graphes acycliques.

4.3.3 Maximisation en utilisant le modèle RUC

Malheureusement pour nous, la borne de l'algorithme de "Greedy Hill Climbing" n'est pas valide pour résoudre la maximisation de l'influence en utilisant le modèle RUC.

Propriété 3. *La fonction σ n'est pas sous-modulaire quand elle est appliquée au modèle RUC.*

Démonstration. Nous exhibons un contre-exemple en utilisant le graphe social présenté dans la figure 4.3.

Si σ est sous-modulaire, alors

$$\begin{aligned}
&\sigma_{RUC}(S \cup \{x\}) - \sigma_{RUC}(S) \geq \sigma_{RUC}(T \cup \{x\}) - \sigma_{RUC}(T) \\
\Leftrightarrow \quad &\frac{1}{1+e^{-\lambda_0-\lambda_1\times s-\lambda_2\times w-\lambda_3}} \geq \frac{1}{1+e^{-\lambda_0-\lambda_1\times s-\lambda_2\times w-2\lambda_3}} - \frac{1}{1+e^{-\lambda_0-\lambda_1\times s-\lambda_2\times w-\lambda_3}} \\
\Leftrightarrow \quad &2\times\frac{1}{1+e^{-\alpha-\lambda_3}} \geq \frac{1}{1+e^{-\alpha-2\lambda_3}} \text{ avec } \alpha = \lambda_0 - \lambda_1 \times s - \lambda_2 \times w \\
\Leftrightarrow \quad &\frac{1+e^{-\alpha-2\lambda_3}}{1+e^{-\alpha-\lambda_3}} \geq \frac{1}{2}
\end{aligned}$$

Si $\lambda_3 = 2$ et $\alpha = -4$ alors $\frac{2}{1+e^2} < \frac{1}{2}$. Ce qui amène une contradiction, donc σ appliquée au modèle RUC n'est pas sous-modulaire. □

L'algorithme glouton "Greedy Hill Climbing" reste toutefois une approximation du résultat optimal pour la maximisation de l'influence en utilisant le modèle RUC. Nous développons dans la section suivante un certain nombre d'exemples pour lesquels, même si nous n'avons pas de garantie quant à sa qualité, l'approximation donne de bons résultats.

4.4 Exemples d'approximation pour le modèle RUC

Nous n'avons pas pu définir de borne quant à la qualité de l'approximation de la méthode gloutonne pour le modèle RUC. Dans cette section, nous comparons la méthode gloutonne d'une part à l'optimal quand c'est possible, et d'autre part à des méthodes naïves.

4.4.1 Méthodes naïves

Dans le but d'obtenir plus d'informations sur la qualité de la méthode gloutonne, et pas seulement le seuil d'approximation quand on le connaît, 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 κ.
- 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 κ.

4.4.2 Généralisation Greedy-n

Il existe une généralisation de l'algorithme glouton "greedy hill climbing" qui a été présentée dans [Du et al., 2008]. Le principe est le suivant : au lieu de choisir les utilisateurs apportant le meilleur gain marginal un par un, on les choisit deux par deux, trois par trois, et ainsi de suite. Il va de soit que si l'on choisit les utilisateurs κ par κ, on obtient le résultat optimal. L'idée en choisissant plusieurs utilisateurs à la fois est d'essayer d'améliorer l'approximation. Le désavantage se trouve dans la complexité.

Pour un réseau de N utilisateurs, l'algorithme standard a une complexité de κN fois le temps de calcul de l'influence. A chaque choix d'utilisateur, on doit calculer l'influence de tous les utilisateurs du réseau pour choisir le meilleur. Si l'on prend maintenant la variante qui choisit les utilisateurs deux par deux, la complexité est de κN^2 fois le temps de calcul de l'influence. On augmente donc radicalement la complexité du calcul.

Nous montrons dans les sections suivantes quelques exemples de calcul de l'influence et nous allons voir que la généralisation n'apporte pas beaucoup d'améliorations pour un temps de calcul beaucoup plus important.

4.4.3 Jeux de données "jouets"

#initiateurs	1	2	3	4	5	6	7
Optimal	7.85	12.70	15.91	17.89	19.15	19.86	20.27
Greedy-1	7.85	12.60	15.65	17.54	18.79	19.61	20.10
Greedy-2	7.85	12.70	15.72	17.69	18.89	19.67	20.14
Degré	7.85	11.99	14.43	16.05	17.20	17.92	18.52
Centralité	6.48	10.58	13.20	14.95	16.14	17.10	17.79

TABLE 4.1 – Valeur de l'influence maximale pour le modèle RUC sur des petits jeux de données générés aléatoirement

Dans un premier temps, afin de pouvoir tester la qualité de la méthode gloutonne par rapport à l'optimal, nous avons fait une étude sur des petits réseaux générés aléatoirement. Ils possèdent tous entre 10 et 30 utilisateurs et entre 30 et 200 liens. Le tableau 4.1 montre le résultat de la maximisation de l'influence sur ces jeux de données "jouets". Les résultats sont moyennés sur 100 réseaux générés aléatoirement : aussi bien les liens entre utilisateurs que les paramètres de la diffusion ont été générés, rien n'a été fixé.

On voit tout d'abord que les méthodes naïves obtiennent de résultats moins bons que les algorithmes gloutons. Ensuite, la différence entre les deux méthodes gloutonnes n'est pas significative : elles obtiennent des résultats très similaires. En creusant un peu les résultats, on s'aperçoit que dans la plupart des cas, elles donnent exactement le même résultat et que la méthode greedy-2 est meilleure seulement sur quelques réseaux spécifiques. Enfin sur ces jeux de données "jouets", les méthodes gloutonnes sont très proches de l'optimal, malgré le fait que nous n'ayons pas prouvé de borne minimale.

Algorithme	Optimal	Greedy-1	Greedy-2	Degré	Centralité
Temps (millisecondes)	11347.15	3.13	14.15	0.18	0.41

TABLE 4.2 – Temps d'exécution des algorithmes de maximisation de l'influence sur des petits jeux de données générés aléatoirement pour 7 diffuseurs initiaux

Le tableau 4.2 montre la moyenne du temps de calcul de chaque algorithme sur l'ensemble des 100 réseaux dans le contexte de sept diffuseurs initiaux. Les méthodes naïves sont extrêmement rapides, mais la méthode gloutonne standard ne prend que 10 fois plus de temps. Il faut par contre de nouveau ajouter un facteur 5 entre les méthode greedy-1 et greedy-2, ce qui vient appuyer la première idée comme quoi la généralisation de la méthode gloutonne n'est viable que pour les versions les plus petites (on choisit peu d'utilisateurs à la fois). Enfin, le calcul de l'optimal prend 11 secondes pour des réseaux très petits et n'est pas possible dans le cas de jeux de données réelles.

4.4.4 Jeux de données réels

Dans cette section, nous avons utilisé un réseau réel et des paramètres fixés aléatoirement (le contenu diffusé, le profil des utilisateurs, la volonté de diffuser des utilisateurs ainsi que les paramètres du modèle RUC ont tous été générés). Le réseau que nous utilisons ici pour apprendre les paramètres et tester la qualité des modèles est **dense-meme**[1]. Ce réseau contient 5000 utilisateurs et 4373 liens entre eux.

#initiateurs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Greedy-1	80	124	149	168	181	193	202	211	218	225	232	238	244	249	254
Greedy-2	80	124	149	168	181	193	202	211	218	225	232	238	244	249	254
Degré	80	124	149	159	174	186	195	201	206	211	214	216	219	224	225
Centralité	80	124	135	145	154	159	162	163	174	177	183	195	210	211	212

TABLE 4.3 – Valeur de l'influence maximale pour le modèle RUC sur le réseau Memetracker dense

Le tableau 4.3 montre la valeur de l'influence maximale trouvée, pour un nombre donné de diffuseurs initiaux, par les différents algorithmes sur ce jeu de données. La valeur optimale n'est pas présente car trop longue à calculer. Le classement des algorithmes est le même que pour les jeux précédents : les méthodes gloutonnes donnent de meilleurs résultats que les heuristiques. On remarque cependant que les deux variantes de la méthode gloutonne donnent exactement le même résultat. Comme vu précédemment, la variante où l'on choisit les utilisateurs deux à deux ne donne de meilleurs résultats que dans certains cas particuliers et ce réseau social (le réseau et les paramètres) n'en fait pas parti.

Pour venir appuyer ces résultats, le tableau 4.4 montre les temps d'exécution des différents algorithmes pour 15 diffuseurs initiaux. Il est clair que les heuristiques sont très

1. nous avons présenté ce réseau en détail dans le chapitre 3

Algorithme	Optimal	Greedy-1	Greedy-2	Degré	Centralité
Temps	NP-difficile	36.24 sec	11 h	0.32 sec	0.61 sec

TABLE 4.4 – Temps d'exécution des algorithmes de maximisation de l'influence sur le réseau Memetracker dense pour 15 utilisateurs initiaux

rapides à coté de l'algorithme glouton. Leurs résultats n'étant pas simplement mauvais, il est à noter qu'elles peuvent être utilisées sur des réseaux pour lesquels il n'est pas possible d'utiliser l'algorithme glouton. On voit aussi que le temps d'exécution de la variante gloutonne greedy-2 prend déjà beaucoup de temps de calcul pour un réseau de seulement 5000 utilisateurs. Sachant qu'elle n'apporte que rarement une faible augmentation de la maximisation de l'influence, il n'est pas intéressant de l'utiliser dans le cas de réseaux de taille importante.

Conclusion

La modélisation des processus de diffusion dans les réseaux sociaux touche plusieurs domaines. Il peut s'agir de la diffusion de virus dans une population, de l'adoption de nouveaux produits par des clients potentiels, ou de la diffusion de contenus entre des utilisateurs. La plupart des modèles proposés jusqu'à présent décrivent le processus de diffusion en fonction de la topologie du réseau : une propagation ne peut avoir lieu entre deux individus que s'ils sont liés. Dans le contexte de la diffusion d'information, ces modèles mettent cependant de coté un certain nombre de paramètres importants tels que :

- le contenu de l'information diffusé ;
- les goûts des utilisateurs, leurs thématiques d'intérêt.

Afin de pallier ce problème, nous avons proposé dans cette thèse une nouvelle famille de modèles probabilistes centrés utilisateur, dans lesquels nous définissons la probabilité de diffusion d'un utilisateur en fonction de sont intérêt pour le contenu diffusé, de sa tendance à diffuser et de la pression sociale qu'il subit. Afin de faire le lien avec les modèles standards, le premier modèle centré utilisateur que nous avons présenté est une variante du modèle à cascades indépendantes IC pour lequel les probabilités de diffusion dépendent de ces caractéristiques utilisateur. Nous avons proposé deux autres modèles dans lesquels nous avons intégré un phénomène de renforcement étape par étape pour modéliser de manière plus fine le processus de diffusion. Le modèle DRUC intègre quant à lui un paramètre de décroissance qui n'est pas présent dans le modèle RUC : un utilisateur tient plus compte d'une information récente que d'une information usagée. Les modèles standards peuvent être écrits comme des processus de percolation de liens. Nous avons montré que ce n'est pas le cas pour les modèles centrés utilisateur avec renforcement. De plus, nous avons illustré les processus de diffusion induits par ces modèles sur des jeux de données artificiels ainsi que sur deux graphes de réseaux sociaux réels.

Afin de tester leur qualité de modélisation, nous avons comparé cette nouvelle famille de modèles avec les modèles standards IC, ASIC et Netrate pour une tâche de prédiction de la diffusion sur des réseaux artificiels ainsi que sur deux jeux de données de blogs ICWSM et Memetracker. Un étude de la précision nous a montré que les modèles contrés utilisateur parviennent mieux à différentier les utilisateurs diffuseurs des utilisateurs non diffuseurs. De plus, une étude de l'erreur de volume de la diffusion nous a montré qu'au-

cun de ces modèles, qui modélisent la diffusion étape par étape, ne parvient à estimer correctement le volume de diffusion. Enfin une étude des paramètres des modèles centrés utilisateur nous permet de montrer l'importance de l'intérêt thématique des utilisateurs dans la modélisation de la diffusion par ces modèles.

Dans un second temps, nous avons étudié le problème de maximisation de l'influence dans les réseaux sociaux. Comme pour les modèles standards, nous avons montré que ce problème est NP-difficile en utilisant les modèles centrés utilisateur. Ce modèle est habituellement approché en utilisant l'algorithme glouton *greedy hill climbing* car la fonction à optimiser est sous-modulaire. Ce n'est malheureusement pas le cas pour les modèles RUC et DRUC. Nous avons néanmoins montré que cet algorithme obtient de bons résultats en maximisant l'influence sur de petits jeux artificiels afin de le comparer à l'optimal et sur des jeux de données réels en le comparant à des heuristiques simples.

Perspectives

Nous avons présenté dans cette thèse une nouvelle famille de modèles qui prennent en compte des caractéristiques utilisateur afin de modéliser la diffusion de contenus dans les réseaux sociaux. Nous les avons comparés à des modèles standards IC, ASIC et Netrate sur des jeux de données de blogs dans lesquels le contenu a un impact important sur la diffusion. Afin de les comparer dans plusieurs contextes de diffusion plus ou moins importante nous avons artificiellement générés des jeux dans lesquels on peut observer une activité faible, moyenne et forte. Il reste cependant un certain nombre d'analyses intéressantes à effectuer sur ces modèles de diffusion.

Analyse des modèles de diffusion

Les modèles Netrate et ASIC qui incluent un paramètre de délai ont tendance, comme nous l'avons vu dans la section 3.7, à prédire une diffusion loin dans le futur lorsque l'on se place dans un contexte de diffusion faible. Les jeux de données que nous avons sélectionnés sont définis sur une durée assez courte d'un mois. Il serait intéressant de faire une étude avec des jeux de données s'écoulant sur plusieurs mois voir plusieurs années afin de savoir si la précision de ces modèles s'améliore plus la durée d'étude est grande. Il faut néanmoins noter que le modèle Netrate définit tous les utilisateurs comme diffuseurs. En fait, les utilisateurs considérés comme non diffuseurs sont ceux qui sont prédits diffuseurs suffisamment loin dans le futur pour qu'on ne les prennent pas en compte.

Dans un ordre d'idée similaire, le modèle RUC prédit lui aussi une probabilité de diffuser proche de 1 pour tous les utilisateurs au bout d'une durée suffisamment grande. C'est d'ailleurs pour cela que nous avons proposé le modèle avec décroissance de la probabilité au cours du temps DRUC. Les expériences que nous avons faites dans cette thèse n'ont pas permis de montrer un apport important de ce paramètre de décroissance. Une étude sur des jeux de données d'une longue durée pourrait montrer des résultats plutôt moyens pour le modèle RUC et une nette amélioration de ceux du modèle DRUC en comparaison.

Enfin, dans cette thèse nous n'avons comparé les modèles centrés utilisateurs qu'à des modèles qui ne prennent en compte que la pression sociale lors de nos expériences. Depuis, un certain nombre de modèles prenant en compte d'autres paramètres ont été proposé

comme ceux que nous avons présenté dans la section 1.5. Nous avons montré l'importance de certaines caractéristiques comme l'intérêt des utilisateurs pour le contenu diffusé. Il serait donc intéressant de comparer les modèles centrés utilisateur à ces modélisations plus standards qui prennent en compte des nouveaux paramètres.

Améliorations des modèles centrés utilisateur

Le modèle DRUC inclue un paramètre de décroissance de l'influence que les utilisateurs ont sur leur voisin plus ils ont diffusé un contenu loin dans le passé. Dans cette thèse nous l'avons fixé arbitrairement à 0.9 car nous voulions simplement qu'il modélise une faible décroissance avec le temps. Il est possible d'estimer ce paramètre au même titre que les quatre autres. A noter tout de même que le temps d'apprentissage des modèles avec renforcement est assez important et qu'ajouter un nouveau paramètre à estimer ne fait que renforcer la complexité des calculs.

Les modèles avec renforcement RUC et DRUC assignent à chaque utilisateur une probabilité qu'il ai diffusé le contenu à chaque étape de la diffusion. Nous avons vu de par les résultats que le classement qu'ils font des utilisateurs est bien meilleur que celui des modèles standards : ils assignent une plus forte probabilité aux utilisateurs diffuseurs. Il y a donc normalement un seuil qui permet de séparer les utilisateurs diffuseurs des utilisateurs non diffuseurs en fonction de leur probabilité de diffusion. Nous avons essayé de fixer un seuil identique pour toutes les cascades sans obtenir de résultat satisfaisant. En effet, les valeurs des probabilités assignées aux utilisateurs sont très différentes d'une cascade à l'autre. Par exemple, un seuil de 0.2 donnera de très bons résultats sur une cascade car il sépare correctement les utilisateur mais sera nettement inférieur à toutes les probabilités de diffusion sur une autre cascade. Il faudrait donc trouver une méthode pour estimer ce seuil en fonction des contenus diffusés.

Dans nos expériences, les modèles avec renforcement obtiennent de meilleurs résultats que le modèle centré utilisateur simple. Cela montre un apport de la modélisation avec renforcement pour ce type de modèles. Nous pensons que cet apport est principalement dû aux variations des probabilités dans le temps en fonction des utilisateurs diffuseurs, ce qui permet de modéliser de manière plus fine la diffusion. Nous avons l'idée d'adapter le modèle IC pour y inclure ce phénomène de renforcement pour voir si l'apport est aussi important sur un modèle plus simple ou si la prise en compte des caractéristiques utilisateur est nécessaire.

Prise en compte de différents types de réseaux

Nous n'avons travaillé dans cette thèse que sur des réseaux sociaux dans lesquels la diffusion est ouverte. Lorsqu'un utilisateur diffuse un contenu, il le rend disponible à tous

ceux qui veulent le lire. En particulier ses voisins auront l'information de la présence de ce nouveau contenu. Dans certains réseaux sociaux, comme les e-mails, le principe régissant la diffusion est quelque peu différent. Un utilisateur qui cherche à diffuser un contenu choisit une liste d'utilisateurs à qui il veut envoyer ce contenu. Se pose alors un problème de choix des destinataires lors de la diffusion. Les modèles à cascades peuvent plus ou moins traiter ce problème car les probabilités de diffusion se trouvent sur les liens du graphe. En adaptant le modèle centré utilisateur RUC avec des probabilités de diffusion sur les liens, nous pourrions prédire les destinataires d'un contenu en prenant en compte des caractéristiques autant de l'utilisateur source que de l'utilisateur destinataire.

La plupart des réseau sociaux n'ont pas une topologie prédéfinie, des liens entre les utilisateurs peuvent être créés et détruits. La plupart des modèles de diffusion et en particulier les modèles centrés utilisateur sont définis pour un réseau social statique. Une approche similaire à celle du modèle Netrate qui consiste à estimer les influences entre les utilisateurs de par les diffusions d'un jeu de données d'entraînement pourrait permettre d'adapter les modèles centrés utilisateur pour qu'ils ne prennent plus en compte les liens explicitement définis dans le réseau social étudié.

Problèmes liés à la diffusion de contenu

Un certain nombre de problèmes sont liés au problème de diffusion dans les réseaux sociaux. Nous avons par exemple travaillé sur le problème de maximisation de l'influence dans cette thèse. D'autres problèmes comme la détection de lien ou de communautés sont aussi reliés à la diffusion. En effet, les utilisateurs qui se lient entre eux le font toujours pour une raison. Ils peuvent habiter au même endroit, fréquenter les même lieux ou partager des centres d'intérêt. Dans tous les cas, l'étude des diffusions entre les utilisateurs peut permettre d'inférer des liens entre utilisateurs qui ne sont pas déjà reliés.

Les études que nous avons faites sur le problème de maximisation de l'influence ont montrés que sur des réseaux artificiels très simples, la méthode gloutonne "greedy hill climbing" obtient très souvent le résultat optimal. Une étude plus poussée pourrait permettre de savoir dans quels cas de figure cette méthode ne permet pas d'obtenir le résultat optimal. Dans ce cas là, il serait possible de reporter les résultats sur des graphes réels et de connaître plus précisément la qualité de l'approximation voir de découper le problème en sous-problèmes pour lesquels la méthode est optimale.

Annexe A

Preuve de la probabilité de diffusion globale

$\forall t \geq 0$, notre hypothèse H_1(t) est :

$$P(n_i, c^k, \leq t) = \sum_{t'=0}^{t-1} P(n_i, c^k, t') \prod_{\tau=0}^{t'-1} (1 - P(n_i, c^k, \tau))$$

$H_1(0)$ est vrai.

Si H_1(t) est vrai, alors

$$\begin{aligned}
P(n_i, c^k, \leq t+1) &= P(n_i, c^k, \leq t) + (1 - P(n_i, c^k, \leq t))P(n_i, c^k, t) \\
&= \sum_{t'=0}^{t-1} P(n_i, c^k, t') \prod_{\tau=0}^{t'-1} (1 - P(n_i, c^k, \tau)) \\
&\quad + P(n_i, c^k, t) \left(1 - \sum_{t'=0}^{t-1} P(n_i, c^k, t') \prod_{\tau=0}^{t'-1} (1 - P(n_i, c^k, \tau)) \right) \\
&= \sum_{t'=0}^{t-1} P(n_i, c^k, t') \prod_{\tau=0}^{t'-1} (1 - P(n_i, c^k, \tau)) \\
&\quad + P(n_i, c^k, t) \prod_{t'=0}^{t-1} (1 - P(n_i, c^k, t')) \\
&= \sum_{t'=0}^{t} P(n_i, c^k, t') \prod_{\tau=0}^{t'-1} (1 - P(n_i, c^k, \tau))
\end{aligned}$$

Alors H_1(t+1) est vrai.

Il reste quand même à prouver que :

$$\left(1 - \sum_{t'=0}^{t-1} P(n_i, c^k, t') \prod_{\tau=0}^{t'-1} (1 - P(n_i, c^k, \tau)) \right) = \prod_{t'=0}^{t-1} (1 - P(n_i, c^k, t'))$$

Nous avons :

$$\left(1 - \sum_{t'=0}^{t-1} P(n_i, c^k, t') \prod_{\tau=0}^{t'-1} (1 - P(n_i, c^k, \tau))\right)$$

$$= [1 - P(n_i, c^k, 0)] \left(1 - \sum_{t'=1}^{t-1} P(n_i, c^k, t') \prod_{\tau=1}^{t'-1} (1 - P(n_i, c^k, \tau))\right)$$

en inférant cette étape, on prouve l'équation précédente.

Annexe B

Preuve d'équivalence pour l'espérance du nombre de voisins diffuseurs

On veut prouver

$$a_1(i) = \sum_{j \in B(i)} p(j)$$
$$=$$
$$a_2(i) = \sum_{k=0}^{n} k \times \left(\sum_{S \subseteq B(i), |S|=k} \prod_{j \in S} p(j) \prod_{j \notin S} (1 - p(j)) \right)$$
$$avec |B(i)| = n$$

Démonstration par récurrence

Le cas pour $n = 1$ ($B(i) = \{u\}$) est trivial :

$$a_1(i) = \sum_{j \in B(i)} p(j) = p(u)$$
$$a_2(i) = \sum_{k=0}^{n} k \times \left(\sum_{S \subseteq B(i), |S|=k} \prod_{j \in S} p(j) \prod_{j \notin S} (1 - p(j)) \right) = p(u)$$

Supposons maintenant que la propriété soit vraie pour un utilisateur i tel que $B(i) = n-1$.

Pour un utilisateur i tel que $B(i) = n$, on a :

$$a_2(i) = \sum_{k=0}^{n} k \times \left(\sum_{S \subseteq B(i), |S|=k} \prod_{j \in S} p(j) \prod_{j \notin S} (1 - p(j)) \right)$$

$$a_2(i) = n\prod_{k=1}^{n} p(k) + \sum_{k=0}^{n-1} k\left(\sum_{S\subseteq B(i),|S|=k),n\notin S}\prod_{j\in S} p(j)\prod_{j\notin S}(1-p(j))\right)$$
$$+\sum_{k=0}^{n-1} k\left(\sum_{S\subseteq B(i),|S|=k,n\in S}\prod_{j\in S} p(j)\prod_{j\notin S}(1-p(j))\right)$$
$$a_2(i) = n\prod_{k=1}^{n} p(k) + T_1 + T_2$$

$$T_1 = \sum_{k=0}^{n-1} k\left(\sum_{S\subseteq B(i),|S|=k,n\notin S}\prod_{j\in S} p(j)\prod_{j\notin S}(1-p(j))\right)$$
on factorise par $1-p(n)$
$$T_1 = (1-p(n))\sum_{k=0}^{n-1} k\left(\sum_{S\subseteq B(i),|S|=k),n\notin S}\prod_{j\in S} p(j)\prod_{j\notin S\cup\{n\}}(1-p(j))\right)$$
hypothèse de récurence
$$T_1 = (1-p(n))\sum_{k=1}^{n-1} p(k)$$

$$T_2 = \sum_{k=0}^{n-1} k\left(\sum_{S\subseteq B(i),|S|=k,n\in S}\prod_{j\in S} p(j)\prod_{j\notin S}(1-p(j))\right)$$
on sépare $k\in[2,(n-1)]$ et $k=1$ ($k=0$ implique 0)
$$T_2 = \sum_{k=2}^{n-1} k\left(\sum_{S\subseteq B(i),|S|=k,n\in S}\prod_{j\in S} p(j)\prod_{j\notin S}(1-p(j))\right) + \sum_{S\subseteq B(i),|S|=1,n\in S}\prod_{j\in S} p(j)\prod_{j\notin S}(1-p(j))$$
$\frac{k-1}{k-1} = 1$ et sortir p(n)
$$T_2 = p(n)\sum_{k=2}^{n-1}\frac{k}{k-1}\left((k-1)\sum_{S\subseteq B(i),|S|=k-1,n\notin S}\prod_{j\in S} p(j)\prod_{j\notin S}(1-p(j))\right)$$
$$+\sum_{S\subseteq B(i),|S|=1,n\in S}\prod_{j\in S} p(j)\prod_{j\notin S}(1-p(j))$$
$$T_2 = T_{2,1} + \sum_{S\subseteq B(i),|S|=1,n\in S}\prod_{j\in S} p(j)\prod_{j\notin S}(1-p(j))$$

$$T_{2,1} = p(n)\sum_{k=2}^{n-1}\frac{k}{k-1}\left((k-1)\sum_{S\subseteq B(i),|S|=k-1,n\notin S}\prod_{j\in S} p(j)\prod_{j\notin S}(1-p(j))\right)$$
séparer $k=(k-1)+1$

$$T_{2,1} = p(n) \sum_{k=2}^{n-1} (k-1) \left(\sum_{S \subseteq B(i), |S|=k-1, n \notin S} \prod_{j \in S} p(j) \prod_{j \notin S} (1-p(j)) \right)$$
$$+ p(n) \sum_{k=2}^{n-1} \left(\sum_{S \subseteq B(i), |S|=k-1, n \notin S} \prod_{j \in S} p(j) \prod_{j \notin S} (1-p(j)) \right)$$

$k = 0$ implique 0 et changement de variable $k = k - 1$

$$T_{2,1} = p(n) \sum_{k=0}^{n-2} k \left(\sum_{S \subseteq B(i), |S|=k, n \notin S} \prod_{j \in S} p(j) \prod_{j \notin S} (1-p(j)) \right)$$
$$+ p(n) \sum_{k=2}^{n-1} \left(\sum_{S \subseteq B(i), |S|=k-1, n \notin S} \prod_{j \in S} p(j) \prod_{j \notin S} (1-p(j)) \right)$$

on rentre p(n)

$$T_{2,1} = p(n) \sum_{k=0}^{n-2} k \left(\sum_{S \subseteq B(i), |S|=k, n \notin S} \prod_{j \in S} p(j) \prod_{j \notin S} (1-p(j)) \right)$$
$$+ \sum_{k=2}^{n-1} \left(\sum_{S \subseteq B(i), |S|=k-1, nin S} \prod_{j \in S} p(j) \prod_{j \notin S} (1-p(j)) \right)$$
$$T_{2,1} = T_{2,1,1} + \sum_{k=2}^{n-1} \left(\sum_{S \subseteq B(i), |S|=k-1, n \in S} \prod_{j \in S} p(j) \prod_{j \notin S} (1-p(j)) \right)$$

$$T_{2,1,1} = p(n) \sum_{k=0}^{n-2} k \left(\sum_{S \subseteq B(i), |S|=k, n \notin S} \prod_{j \in S} p(j) \prod_{j \notin S} (1-p(j)) \right)$$
$$T_{2,1,1} = p(n) \sum_{k=0}^{n-1} k \left(\sum_{S \subseteq B(i), |S|=k, n \notin S} \prod_{j \in S} p(j) \prod_{j \notin S} (1-p(j)) \right)$$
$$- p(n) (n-1) \left(\sum_{S(|n-1|), n \notin S} \prod_{j \in S} p(j) \prod_{j \notin S} (1-p(j)) \right)$$

hypothèse et expansion

$$T_{2,1,1} = p(n) \sum_{k=1}^{n-1} p(k) - p(n) \times (n-1) \prod_{i=1}^{n-1} p(i)$$

$$T_2 = \sum_{k=0}^{n-1} k \left(\sum_{S \subseteq B(i), |S|=k, n \in S} \prod_{j \in S} p(j) \prod_{j \notin S} (1-p(j)) \right)$$

réécriture

$$T_2 = p(n) \sum_{k=1}^{n-1} p(k) - p(n) \times (n-1) \prod_{i=1}^{n-1} p(i) + \sum_{k=2}^{n-1} \left(\sum_{S \subseteq B(i), |S|=k, n \in S} \prod_{j \in S} p(j) \prod_{j \notin S} (1-p(j)) \right)$$

$$+ \sum_{S\subseteq B(i),|S|=1,n\in S} \prod_{j\in S} p(j) \prod_{j\notin S} (1-p(j))$$

contracter

$$T_2 = p(n)\sum_{k=1}^{n-1} p(k) - p(n)\times(n-1)\prod_{i=1}^{n-1} p(i) + \sum_{k=1}^{n-1}\left(\sum_{S\subseteq B(i),|S|=k,n\in S} \prod_{j\in S} p(j) \prod_{j\notin S} (1-p(j))\right)$$

sortir p(n) et changement de variable $k = k-1$

$$T_2 = p(n)\sum_{k=1}^{n-1} p(k) - p(n)\times(n-1)\prod_{i=1}^{n-1} p(i) + p(n)\sum_{k=0}^{n-2}\left(\sum_{S\subseteq B(i),|S|=k,n\notin S} \prod_{j\in S} p(j) \prod_{j\notin S} (1-p(j))\right)$$

$$a_2(i) = n\prod_{k=1}^{n} p(k) + (1-p(n))\sum_{k=1}^{n-1} p(k) + p(n)\sum_{k=1}^{n-1} p(k) - p(n)\times(n-1)\prod_{i=1}^{n-1} p(i)$$
$$+ p(n)\sum_{k=0}^{n-2}\left(\sum_{S\subseteq B(i),|S|=k,n\notin S} \prod_{j\in S} p(j) \prod_{j\notin S} (1-p(j))\right)$$

réécriture

$$a_2(i) = \left((1-p(n))\sum_{k=1}^{n-1} p(k) + p(n)\sum_{k=1}^{n-1} p(k)\right) + \left(n\prod_{k=1}^{n} p(k) - (n-1)\prod_{i=1}^{n} p(i)\right)$$
$$+ p(n)\sum_{k=0}^{n-2}\left(\sum_{S\subseteq B(i),|S|=k,n\notin S} \prod_{j\in S} p(j) \prod_{j\notin S} (1-p(j))\right)$$

$$a_2(i) = \sum_{k=1}^{n-1} p(k) + \prod_{k=1}^{n} p(k) + p(n)\sum_{k=0}^{n-2}\left(\sum_{S\subseteq B(i),|S|=k,n\notin S} \prod_{j\in S} p(j) \prod_{j\notin S} (1-p(j))\right)$$

contracter

$$a_2(i) = \sum_{k=1}^{n-1} p(k) + p(n)\sum_{k=0}^{n-1}\left(\sum_{S\subseteq B(i),|S|=k,n\notin S} \prod_{j\in S} p(j) \prod_{j\notin S} (1-p(j))\right)$$

$$a_2(i) = \sum_{k=1}^{n-1} p(k) + p(n)$$

$$a_2(i) = a_1(i)$$

Car : $\sum_{k=0}^{n-1}\left(\sum_{S\subseteq B(i),|S|=k,n\notin S} \prod_{j\in S} p(j) \prod_{j\notin S} (1-p(j))\right) = 1$ (voir page suivante)

Alors la propriété est vrai pour un utilisateur i tel que $B(i) = n$.

On montre la propriété suivante par récurrence :
$\forall n, \sum_{S \in P([1,n])} \prod_{j \in S} p(j) \prod_{j \notin S} (1 - p(j)) = 1$
où $P([1, n])$ est l'ensemble des parties de l'ensemble $[1, n]$

Si $n = 1$, on a : $\sum_{S \in P(1)} \prod_{j \in S} p(j) \prod_{j \notin S} (1 - p(j)) = p(1) + 1 - p(1) = 1$

Supposons la propriété vraie pour $n - 1$, on a :
$$\sum_{S \in P([1,n])} \prod_{j \in S} p(j) \prod_{j \notin S} (1 - p(j))$$
$$= p(n) \left(\sum_{S \in P([1,n-1])} \prod_{j \in S} p(j) \prod_{j \notin S} (1 - p(j)) \right) + (1 - p(n)) \left(\sum_{S \in P([1,n-1])} \prod_{j \in S} p(j) \prod_{j \notin S} (1 - p(j)) \right)$$
$$= \sum_{S \in P([1,n-1])} \prod_{j \in S} p(j) \prod_{j \notin S} (1 - p(j))$$
$$= 1$$
Alors la propriété est vraie pour n.

Annexe C

Valeurs des paramètres pour les modèles centrés utilisateurs

Modèle	UC	RUC	DRUC
λ_0 (biais)	-0.28	-6.22	-5.69
λ_1 (intérêt thématique)	6.46	9.1	9.52
λ_2 (activité)	9.28	6.22	2.36
λ_3 (pression sociale)	0	1.78	0.52

TABLE C.1 – Valeurs des paramètres après apprentissage sur le jeu de données **creux-meme-aleat**

Modèle	UC	RUC	DRUC
λ_0 (biais)	-1.77	-7.39	-6.77
λ_1 (intérêt thématique)	6.8	14.1	11.23
λ_2 (activité)	11.77	7.39	6.77
λ_3 (pression sociale)	0.73	2.6	3.23

TABLE C.2 – Valeurs des paramètres après apprentissage sur le jeu de données **creux-icwsm-aleat**

Modèle	UC	RUC	DRUC
λ_0 (biais)	-1.77	-8.23	-8.23
λ_1 (intérêt thématique)	6.8	9.52	10.52
λ_2 (activité)	9.1	3.69	2.36
λ_3 (pression sociale)	0.73	1.78	1.78

TABLE C.3 – Valeurs des paramètres après apprentissage sur le jeu de données **creux-meme-ident**

Modèle	UC	RUC	DRUC
λ_0 (biais)	-1.39	-7.28	-6.77
λ_1 (intérêt thématique)	6.69	10.77	9.89
λ_2 (activité)	9.28	7.28	6.77
λ_3 (pression sociale)	0	2.72	3.23

TABLE C.4 – Valeurs des paramètres après apprentissage sur le jeu de données **creux-icwsm-ident**

Modèle	UC	RUC	DRUC
λ_0 (biais)	-2.36	-5.47	-3.33
λ_1 (intérêt thématique)	2.36	7.01	9.49
λ_2 (activité)	6.27	5.92	3.99
λ_3 (pression sociale)	1.19	2.78	0.95

TABLE C.5 – Valeurs des paramètres après apprentissage sur le jeu de données **dense-meme**

Modèle	UC	RUC	DRUC
λ_0 (biais)	-3.13	-6.77	-2.61
λ_1 (intérêt thématique)	7.06	6.77	10.75
λ_2 (activité)	3.13	5.52	4.27
λ_3 (pression sociale)	1.94	3.23	1.55

TABLE C.6 – Valeurs des paramètres après apprentissage sur le jeu de données **dense-icwsm**

Modèle	UC	RUC	DRUC
λ_0 (biais)	-7.12	-6.23	-5.12
λ_1 (intérêt thématique)	-	-	-
λ_2 (activité)	7.12	6.23	6.23
λ_3 (pression sociale)	1.77	1.77	0.77

TABLE C.7 – Valeurs des paramètres après apprentissage sur le jeu de données **art-nosim**

Modèle	UC	RUC	DRUC
λ_0 (biais)	-7.61	-4.89	-2.86
λ_1 (intérêt thématique)	7.61	4.89	5.11
λ_2 (activité)	0.36	3.78	2.86
λ_3 (pression sociale)	0.31	1.77	0.12

TABLE C.8 – Valeurs des paramètres après apprentissage sur le jeu de données **art-sim**

Publications de l'auteur

Journaux internationaux

[Lagnier et al., 2013b] (Soumis) Lagnier, S. C., Gaussier, E., and cois Kawala, F. (2013b). User-centered probabilistic models for content diffusion in social networks. *Social Networks*.

Conférences internationales

[Lagnier et al., 2013a] Lagnier, C., Denoyer, L., Gaussier, E., and Gallinari, P. (2013a). Predicting information diffusion in social networks using content and users profiles. In *ECIR*.

Journaux nationaux

[Lagnier and Gaussier, 2012] Lagnier, C. and Gaussier, E. (2012). Modéliser l'utilisateur pour la diffusion de l'information dans les réseaux sociaux. *ISI*.

[Lagnier and Gaussier, 2011] Lagnier, C. and Gaussier, E. (2011). Un modèle de diffusion de l'information dans les réseaux sociaux. *RNTI-AAFD*.

Conférences nationales

[Lagnier et al., 2011] Lagnier, C., Gaussier, E., and cois Kawala, F. (2011). Modéliser l'utilisateur pour la diffusion de l'information dans les réseaux sociaux. In *MARAMI*.

Bibliographie

[Adar and Adamic, 2005] Adar, E. and Adamic, L. A. (2005). Tracking information epidemics in blogspace. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '05, pages 207–214. IEEE Computer Society.

[Anagnostopoulos et al., 2011] Anagnostopoulos, A., Brova, G., and Terzi, E. (2011). Peer and authority pressure in information-propagation models. In *Proceedings of the 2011 European conference on Machine learning and knowledge discovery in databases - Volume Part I*, ECML PKDD'11, pages 76–91. Springer-Verlag.

[Anderson et al., 2012] Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. (2012). Effects of user similarity in social media. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12. ACM.

[Anderson, 1984] Anderson, R. M., editor (1984). *The Population Dynamics of Infectious Diseases : Theory and Applications.* Chapman & Hall, London.

[Apolloni et al., 2009] Apolloni, A., Channakeshava, K., Durbeck, L., Khan, M., Kuhlman, C., Lewis, B., and Swarup, S. (2009). A study of information diffusion over a realistic social network model. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, pages 675–682.

[Asur and Huberman, 2010] Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10. IEEE Computer Society.

[Bailey et al., 2004] Bailey, D. J., Kleczkowski, A., and Gilligan, C. A. (2004). Epidemiological dynamics and the efficiency of biological control of soil-borne disease during consecutive epidemics in a controlled environment. *New Phytologist*, 161(2) :569–575.

[Bakshy et al., 2012] Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12. ACM.

[Beutel et al., 2012] Beutel, A., Prakash, B. A., Rosenfeld, R., and Faloutsos, C. (2012). Interacting viruses in networks : can both survive ? In *Proceedings of the 18th ACM*

SIGKDD international conference on Knowledge discovery and data mining, KDD '12. ACM.

[Borodin et al., 2010] Borodin, A., Filmus, Y., and Oren, J. (2010). Threshold models for competitive influence in social networks. In *WINE*, pages 539–550. Springer.

[Boyd et al., 2010] Boyd, D., Golder, S., and Lotan, G. (2010). Tweet, tweet, retweet : Conversational aspects of retweeting on twitter. In *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, HICSS '10. IEEE Computer Society.

[Budak et al., 2011] Budak, C., Agrawal, D., and El Abbadi, A. (2011). Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 665–674. ACM.

[Burton et al., 2009] Burton, K., Java, A., and Soboroff, I. (2009). The ICWSM 2009 Spinn3r Dataset. In *The Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*.

[Cha et al., 2009] Cha, M., Antonio, J., Prez, N., and Haddadi, H. (2009). Flash floods and ripples : The spread of media content through the blogosphere. In *Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media*, ICWSM '09.

[Chierichetti et al., 2011] Chierichetti, F., Kleinberg, J. M., and Liben-Nowell, D. (2011). Reconstructing patterns of information diffusion from incomplete observations. In *NIPS*, pages 792–800.

[Chierichetti et al., 2010] Chierichetti, F., Lattanzi, S., and Panconesi, A. (2010). Rumour spreading and graph conductance. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10. Society for Industrial and Applied Mathematics.

[Dodds and Watts, 2004] Dodds, P. and Watts, D. (2004). Universal Behavior in a Generalized Model of Contagion. *Physical Review Letters*, 92(21).

[Du et al., 2008] Du, D.-Z., Graham, R. L., Pardalos, P. M., Wan, P.-J., Wu, W., and Zhao, W. (2008). Analysis of greedy approximations with nonsubmodular potential functions. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '08, pages 167–175, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

[Gomez-Rodriguez et al., 2011] Gomez-Rodriguez, M., Balduzzi, D., and Schölkopf, B. (2011). Uncovering the temporal dynamics of diffusion networks. In Getoor, L. and Scheffer, T., editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 561–568. ACM.

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

[Goyal et al., 2010] Goyal, A., Bonchi, F., and Lakshmanan, L. V. (2010). Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 241–250. ACM.

[Granovetter, 1978] Granovetter, M. (1978). Threshold Models of Collective Behavior. *American Journal of Sociology*, 83(6) :1420–1443.

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

[Grassberger, 1983] Grassberger, P. (1983). On the critical behavior of the general epidemic process and dynamical percolation. *Mathematical Biosciences*, 63 :157–172.

[Grimal, 2012] Grimal, C. (2012). *Apprentissage de co-similairtés pour la classification automatique de données monovues et multivues*. PhD thesis, Université de Grenoble.

[Gruhl and Liben-nowell, 2004] Gruhl, D. and Liben-nowell, D. (2004). Information diffusion through blogspace. In *In WWW 04*, pages 491–501. ACM Press.

[Guille and Hacid, 2012] Guille, A. and Hacid, H. (2012). A predictive model for the temporal dynamics of information diffusion in online social networks. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 1145–1152. ACM.

[Gupte et al., 2009] Gupte, M., Hajiaghayi, M., Han, L., Iftode, L., Shankar, P., and Ursu, R. M. (2009). News posting by strategic users in a social network. In *Proceedings of the 5th International Workshop on Internet and Network Economics*, WINE '09, pages 632–639. Springer-Verlag.

[Hanneman and Riddle, 2005] Hanneman, R. A. and Riddle, M. (2005). *Introduction to social network methods*. University of California.

[Heaukulani and Ghahramani, 2013] Heaukulani, C. and Ghahramani, Z. (2013). Dynamic probabilistic models for latent feature propagation in social networks. In Dasgupta, S. and Mcallester, D., editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 275–283. JMLR Workshop and Conference Proceedings.

[Hong et al., 2011] Hong, L., Dan, O., and Davison, B. D. (2011). Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11. ACM.

[Jaccard, 1901] Jaccard, P. (1901). Etude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37 :547–579.

[Jackson and Yariv, 2005] Jackson, M. O. and Yariv, L. (2005). Diffusion on social networks.

[Jansen et al., 2009] Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter power : Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11) :2169–2188.

[Java et al., 2007] Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter : understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07. ACM.

[Karp, 1972] Karp, R. M. (1972). Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pages 85–103.

[Kawala, 2011] Kawala, F. (2011). Study on the user centric diffusion model. Master's thesis, Université Joseph Fourier.

[Kempe et al., 2005] Kempe, D., Kleinberg, J., and Éva Tardos (2005). Influential nodes in a diffusion model for social networks. In *IN ICALP*, pages 1127–1138. Springer Verlag.

[Kempe et al., 2003] Kempe, D., Kleinberg, J., and 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*, pages 137–146. ACM Press.

[Kimura et al., 2007] Kimura, M., Saito, K., and Nakano, R. (2007). Extracting influential nodes for information diffusion on a social network. *Proceedings Of The National Conference On Artificial Intelligence*, 22(2) :1371.

[Kleinberg, 2007] Kleinberg, J. (2007). *Cascading Behavior in Networks : Algorithmic and Economic Issues*. Cambridge University Press.

[Klimt and Yang, 2004] Klimt, B. and Yang, Y. (2004). Introducing the enron corpus. In *CEAS*.

[Koide et al., 2011] Koide, A., Saito, K., Ohara, K., Kimura, M., and Motoda, H. (2011). Estimating diffusion probability changes for asic-sis model. *Journal of Machine Learning Research - Proceedings Track*, 20 :297–313.

[Leskovec et al., 2009] Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, KDD '09, pages 497–506. ACM.

[Leskovec and Horvitz, 2006] Leskovec, J. and Horvitz, E. (2006). Worldwide buzz : Planetary-scale views on an instant-messaging network. Technical report.

[Leskovec et al., 2007] Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N. S., and Hurst, M. (2007). Patterns of cascading behavior in large blog graphs. In *SDM*.

[Li et al., 2012] Li, C.-T., Kuo, T.-T., Ho, C.-T., Hong, S.-C., Lin, W.-S., and Lin, S.-D. (2012). Modeling and evaluating information propagation in a microblogging social network. *Social Network Analysis and Mining*, pages 1–17.

[Liben-Nowell and Kleinberg, 2008] Liben-Nowell, D. and Kleinberg, J. (2008). Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12) :4633–4638.

[Lieberman et al., 2005] Lieberman, E., Hauert, C., and Nowak, M. A. (2005). Evolutionary dynamics on graphs. *Nature*, 433 :312–316.

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

[Ma et al., 2008] Ma, H., Yang, H., Lyu, M. R., and King, I. (2008). Mining social networks using heat diffusion processes for marketing candidates selection. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 233–242. ACM.

[Macy, 1991] Macy, M. W. (1991). Chains of Cooperation : Threshold Effects in Collective Action. *American Sociological Review*, 56(6) :730–747.

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

[Mercklé, 2004] Mercklé, P. (2004). *Les réseaux sociaux, les origines de lanalyse des réseaux sociaux*. CNED / ens-lsh.

[Myers and Leskovec, 2010] Myers, S. A. and Leskovec, J. (2010). On the convexity of latent social network inference. *CoRR*, abs/1010.5504.

[Myers and Leskovec, 2012] Myers, S. A. and Leskovec, J. (2012). Clash of the contagions : Cooperation and competition in information diffusion. In *ICDM*, pages 539–548. IEEE Computer Society.

[Najar et al., 2012] Najar, A., Denoyer, L., and Gallinari, P. (2012). Predicting information diffusion on social networks with partial knowledge. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 1197–1204.

[Nekovee et al., 2007] Nekovee, M., Moreno, Y., Bianconi, G., and Marsili, M. (2007). Theory of rumour spreading in complex social networks. *Physica A : Statistical Mechanics and its Applications*, 374 :457–470.

[Nemhauser and Wolsey, 1988] Nemhauser, G. L. and Wolsey, L. A. (1988). *Integer and combinatorial optimization*. Wiley-Interscience, New York, NY, USA.

[Newell and Simon, 1972] Newell, A. and Simon, H. A. (1972). *Human problem solving*. Prentice-Hall.

[Newman, 2003] Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review*, 45(2) :167–256.

[Nguyen and Zheng, 2012] Nguyen, H. and Zheng, R. (2012). Influence spread in large-scale social networks — a belief propagation approach. In *Proceedings of the 2012 European conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*, ECML PKDD'12, pages 515–530. Springer-Verlag.

[Pathak et al., 2010] Pathak, N., Banerjee, A., and Srivastava, J. (2010). A generalized linear threshold model for multiple cascades. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pages 965–970. IEEE Computer Society.

[Petrovic et al., 2011] Petrovic, S., Osborne, M., and Lavrenko, V. (2011). Rt to win ! predicting message propagation in twitter. In *ICWSM*. The AAAI Press.

[Prakash et al., 2011] Prakash, B. A., Chakrabarti, D., Faloutsos, M., Valler, N., and Faloutsos, C. (2011). Threshold conditions for arbitrary cascade models on arbitrary networks. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, ICDM '11, pages 537–546. IEEE Computer Society.

[Prakash et al., 2010] Prakash, B. A., Tong, H., Valler, N., Faloutsos, M., and Faloutsos, C. (2010). Virus propagation on time-varying networks : theory and immunization algorithms. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases : Part III*, ECML PKDD'10, pages 99–114. Springer-Verlag.

[Romero et al., 2011a] Romero, D. M., Galuba, W., Asur, S., and Huberman, B. A. (2011a). Influence and passivity in social media. In *Proceedings of the ECML/PKDD 2011*.

[Romero et al., 2011b] Romero, D. M., Meeder, B., and Kleinberg, J. (2011b). Differences in the mechanics of information diffusion across topics : idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 695–704. ACM.

[Saito et al., 2009] Saito, K., Kimura, M., Ohara, K., and Motoda, H. (2009). Learning continuous-time information diffusion model for social behavioral data analysis. *Learning*, 5828 :322–337.

[Saito et al., 2010a] Saito, K., Kimura, M., Ohara, K., and Motoda, H. (2010a). Generative Models of Information Diffusion with Asynchronous Timedelay. *Journal of Machine Learning Research*, 13 :193–208.

[Saito et al., 2010b] Saito, K., Kimura, M., Ohara, K., and Motoda, H. (2010b). Selecting information diffusion models over social networks for behavioral analysis. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases : Part III*, ECML PKDD'10, pages 180–195. Springer-Verlag.

[Saito et al., 2008] Saito, K., Nakano, R., and 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*, KES '08, pages 67–75. Springer-Verlag.

[Saito et al., 2011] Saito, K., Ohara, K., Yamagishi, Y., Kimura, M., and Motoda, H. (2011). Learning diffusion probability based on node attributes in social networks. In Kryszkiewicz, M., Rybinski, H., Skowron, A., and Ras, Z. W., editors, *ISMIS*, volume 6804 of *Lecture Notes in Computer Science*, pages 153–162. Springer.

[Schaffer and Bronnikovab, 2007] Schaffer, W. M. and Bronnikovab, T. V. (2007). Parametric dependence in model epidemics. i : Contact-related parameters. *Journal of Biological Dynamics*, 1 :183–195.

[Song et al., 2007] Song, X., Chi, Y., Hino, K., and Tseng, B. L. (2007). Information flow modeling based on diffusion rate for prediction and ranking. In *Proceedings of the 16th international c1onference on World Wide Web*, WWW '07, pages 191–200. ACM.

[Spearman, 1904] Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15 :88–103.

[Stauffer and Aharony, 1992] Stauffer, D. and Aharony, A. (1992). *Introduction to Percolation Theory, 2nd ed.* London : Taylor & Francis.

[Stavrianou et al., 2009] Stavrianou, A., Velcin, J., and Chauchat, J.-H. (2009). Definition and measures of an opinion model for mining forums. In *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining*, ASONAM '09. IEEE Computer Society.

[Suh et al., 2010] Suh, B., Hong, L., Pirolli, P., and Chi, E. H. (2010). Want to be retweeted ? large scale analytics on factors impacting retweet in twitter network. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, SOCIALCOM '10. IEEE Computer Society.

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

[Vázquez et al., 2006] Vázquez, A., ao Gama Oliveira, J., Dezsö, Z., Goh, K. I., Kondor, I., and Barabási, A. L. (2006). Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73.

[Wang et al., 2012a] Wang, F., Wang, H., and Xu, K. (2012a). Diffusive logistic model towards predicting information diffusion in online social networks. In *Proceedings of the 2012 32nd International Conference on Distributed Computing Systems Workshops*, ICDCSW '12, pages 133–139. IEEE Computer Society.

[Wang et al., 2012b] Wang, L., Ermon, S., and Hopcroft, J. E. (2012b). Feature-enhanced probabilistic models for diffusion network inference. In *Proceedings of the 2012 European conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*, ECML PKDD'12, pages 499–514. Springer-Verlag.

[Wei et al., 2012] Wei, X., Valler, N., Prakash, B. A., Neamtiu, I., Faloutsos, M., and Faloutsos, C. (2012). Competing memes propagation on networks : a case study of composite networks. *SIGCOMM Comput. Commun. Rev.*, 42(5) :5–12.

[Wortman, 2008] Wortman, J. (2008). Viral marketing and the diffusion of trends on social networks. Technical report, University of Pennsylvania.

[Wu and Huberman, 2007] Wu, F. and Huberman, B. A. (2007). Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45) :17599–17601.

[Yang and Leskovec, 2010] Yang, J. and Leskovec, J. (2010). Modeling Information Diffusion in Implicit Networks. *Data Mining, IEEE International Conference on*, 0 :599–608.

[Zaman et al., 2010] Zaman, T. R., Herbrich, R., Van Gael, J., and Stern, D. (2010). Predicting Information Spreading in Twitter. In *Computational Social Science and the Wisdom of Crowds Workshop*.