Augmenter les retweets sur Twitter : comment tirer parti des mentions ?

Soumajit Pramanik, Qinna Wang, Maximilien Danisch, Mohit Sharma, Sumanth Bandi, Jean-Loup Guillaume, Stéphane Raux, Bivas Mitra

Department of Computer Science and Engineering, IIT Kharagpur, India
Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu 75005 Paris
L3I,University of La Rochelle, France
Linkfluence, 5 rue Choron, 75009 Paris, France

RÉSUMÉ. *Alors que Twitter est devenu incontournable, la propagation des tweets et hashtags est toujours largement incomprise. Le propagation d'information sur Twitter est principalement due aux retweets et aux mentions mais, alors que les retweets ne permettent d'atteindre que les abonnés d'un individu, les mentions permettent d'atteindre n'importe qui directement. De nombreuses études ont montré que les mentions sont largement utilisées sur Twitter, mais surtout qu'elles sont fondamentales pour augmenter la popularité des tweets et hashtags. Des méthodes automatiques pour choisir les bons utilisateurs à mentionner pourraient donc permettre d'augmenter la visibilité des tweets.*
Dans cet article nous proposons un système de recommandation de mentions en temps réel pour augmenter la popularité d'un tweet. Ce système est basé sur un modèle de propagation de tweet dans un graphe multiplexe construit à partir d'une étude de données réelles. Il permet de clairement faire la différence entre les propagations dues aux mentions et celles dues aux abonnements. Les simulations du modèle donnent des résultats similaires aux observations empiriques et sont également fondées sur des résultats analytiques. En utilisant ces différents résultats nous proposons une stratégie de recommandation effective et une application Twitter associée.

ABSTRACT. *While Twitter has become one of the most influential micro-blogging systems, the propagation of tweets or hashtags is still widely misunderstood. Information propagation in Twitter is mainly due to 'retweets' and 'mentions' but, while retweets only reach the circle of a source, mentions allows to spread an information far beyond its neighborhood in just one step. Studies show that mentions are not only widely used by Twitter users, but are fundamental in the popularity of tweets or hashtags. Solutions to help in mentioning the proper users could therefore give maximal exposure to a given tweet.*
In this paper, we propose a real time mention recommendation system to enhance the popularity of tweets by a strategic use of the mention utility. This system is based on a model of tweet propagation in a multiplex network whose basic bricks are supported by the study of a real dataset and that allows to make a clear difference between retweets due to mentions and other factors. Simulations of the model show a nice agreement with the empirical tweet popularity observed in the dataset and are further supported by analytical results. Using all these results, we propose an effective mention recommendation strategy and implement it in a Twitter Application.

MOTS-CLÉS : *Twitter, diffusion, cascades, mention, modèle épidémiologique.*

KEYWORDS: *Twitter, diffusion, cascades, mention, epidemic model.*

1. Introduction

Twitter est devenu l'une des plus importantes plateformes de micro-blogging pour diffuser et partager des nouvelles, des informations personnelles ou des idées (González-Bailón *et al*.,2011). Cependant, la popularité des tweets et des hashtags est très variable, et ce dans tous les jeux de données étudiés : seule une très faible proportion de tweets et de hashtags sont fortement populaires (voir encart Fig. 1). La propagation dans Twitter se produit par deux activités principales : les retweets et les mentions (Kato *et al*.,2012). Dans le cas des retweets, l'information est simplement relayée aux abonnés de l'utilisateur qui retweete. La mention permet au contraire d'atteindre directement n'importe qui et donc de rendre l'information plus visible en visant les utilisateurs les plus appropriés. De plus, les mentions sont listées dans un onglet spécifique sur Twitter ce qui les rend plus visibles que les messages classiques. La Fig. 1 montre que les utilisateurs utilisent les mentions de manière très importante et il est admis que cette utilisation massive joue un rôle majeur dans les cascades de tweets et de hashtags. Par exemple, dans le jeu de données présenté section 2, nous observons que la probabilité de retweet d'un utilisateur mentionné est en moyenne 32% plus élevée que celle d'un follower normal.

De nombreuses études ont été menées pour comprendre la popularité des tweets. Dans (Suh *et al*.,2010) et (Malhotra *et al*.,2012), les auteurs étudient le rôle du contenu et des propriétés contextuelles des tweets pour identifier les facteurs significatifs dans le taux de retweet et la popularité. Dans (Petrovic *et al*.,2011) Petrovic et al. développent un système automatique pour prédire la popularité d'un tweet. Dans (Uysal, Croft,2011), Uysal et Croft proposent des méthodes pour recommander des tweets que les utilisateurs devraient trouver intéressants et donc retweeter. Plusieurs modèles d'influence ont également été étudiés. Cependant, dans (Cha *et al*.,2010), Cha et al. montrent que le nombre d'abonnés n'est pas nécessairement une bonne mesure d'influence sur Twitter. Plusieurs modèles ont alors été proposés (Borge-Holthoefer *et al*.,2012 ;Chen *et al*.,2009 ;Kempe *et al*.,2003 ;Zhan *et al*.,2015) pour identifier ces influenceurs. Malgré tout, cibler une personne influente n'assure pas que cette personne va retweeter. Ceci dépend de plusieurs facteurs tels que le contenu du tweet, l'utilisateur lui-même, etc. qui sont ignorés dans le calcul de l'influence. Cela a motivé l'introduction de systèmes pour identifier les bons utilisateurs à mentionner. Ainsi, Wang et al. (Wang *et al*.,2013) proposent l'heuristique 'Whom-to-Mention' qui utilisent des paramètres (tels que l'intérêt partagé de l'utilisateur, le contenu et l'influence) et utilisent des techniques d'apprentissage pour identifier les bons utilisateurs à mentionner. D'autres systèmes similaires sont disponibles dans (Lee *et al*.,2014) et (Tang *et al*.,2014). Ce bref état de l'art souffre cependant de plusieurs limites. Tout d'abord, la plupart des méthodes reposent sur de nombreux paramètres qui ne sont pas calculables en temps réel à l'exécution, ils ne sont donc pas utilisables dans une application. De plus ces systèmes sont des boites noires qui ne donnent pas d'indications sur les liens entre les différents facteurs qui jouent dans la popularité d'un tweet. Par exemple, dans quelle mesure est-ce que l'intérêt de l'utilisateur mentionné joue un rôle ; comment est-ce que l'activité de l'utilisateur mentionné (taux de retweet) doit jouer dans la stratégie de mention ? Pour comprendre cela, un modèle de propagation est nécessaire. Un tel modèle peut permettre de comprendre le rôle des facteurs individuels et donc d'aider à la conception d'un système de recommandation. Cet article est un pas important dans cette direction.

Dans cet article nous développons un système de recommandation de mentions en temps-réel pour améliorer la popularité d'un tweet basé sur un modèle analytique rigoureux. Nous

commençons avec une étude de cas pour montrer l'importance des mentions pour la propagation des tweets. Cette étude nous permet d'identifier les propriétés importantes des utilisateurs mentionnés : popularité, activité et intérêt. Nous modélisons la propagation des tweets comme un réseau multiplexe (Granell *et al*.,2013) et proposons un modèle analytique pour simuler le flux de tweets. Les simulations du modèle s'accordent bien avec la popularité réellement observée. Nous montrons et prouvons également un seuil critique sur le taux de retweet au-delà duquel les mentions sont inutiles. Finalement, en nous basant sur ces différentes observations, nous proposons une stratégie de mention qui dépasse la méthode "Whom to Mention" de l'état de l'art (Wang *et al*.,2013). Pour montrer l'utilité de cette stratégie, nous développons une application Twitter disponible à http://bit.y/Easy-Mention.

Cet article est organisé comme suit : dans la section 2 nous décrivons les jeux de données et les étudions pour montrer l'importance de la mention et identifier les caractéristiques principales des utilisateurs mentionnés. Dans la section 3 nous développons le modèle SIR_{MF} pour étudier la propagation des tweets et comprendre les liens entre les différents paramètres du modèle. Le modèle est complété dans la section 4 par l'identification analytique d'un seuil critique sur la formation de cascades de retweets. Enfin, nous développons le système de recommandation de mentions dans la section 5 et évaluons ses performances avant de conclure.

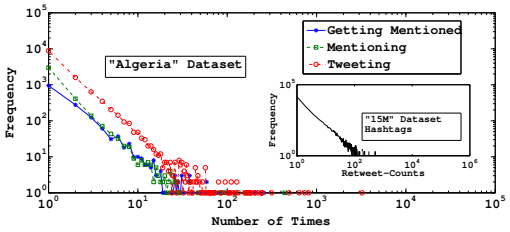


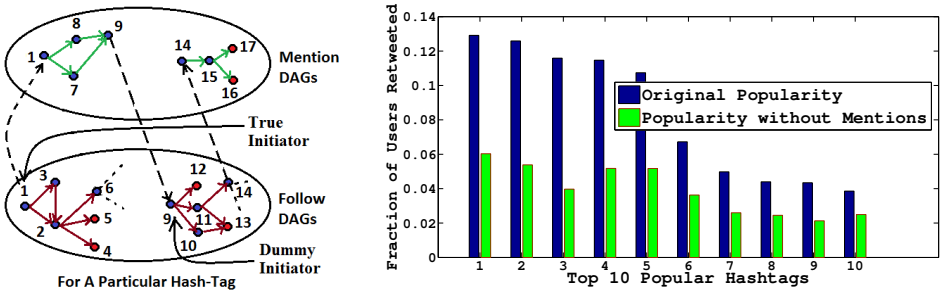
Figure 1. Utilisation de la mention sur Twitter. ***Encart*** *distribution de retweets de hashtag.*

2. Jeux de données et analyses

Nous avons récupéré des tweets durant deux événements (a) le mouvement du printemps Arabe en 2011 et (b) la Coupe du Monde de football en 2014. Dans chacun de ces événements, Twitter a été utilisé de manière intensive pour propager des nouvelles et des opinions. Cependant, les lieux et temporalités des deux événements sont très différents, nous considérerons donc que des résultats similaires peuvent potentiellement correspondre à des schémas généraux sur Twitter.

(a) Mouvement du printemps Arabe : ceci correspond à deux jeux de données disponibles publiquement (i) Algeria Dataset est un ensemble d'environ 60K tweets et 20K utilisateurs qui les ont postés durant les événements en Algérie ("http://dfreelon.org/2012/02/11/arab-spring-twitter-data-now-available-sort-of/",s. d.). Nous avons crawlé le contenu des tweets, des informations sur les utilisateurs et le réseau des abonnements. (ii) 15M dataset est un ensemble de 86K utilisateurs et 0.5M tweets (uniquement les hashtags) postés en Espagne en mai 2011. Le réseau des abonnements est aussi disponible (González-Bailón *et al*.,2011).

(b) World-Cup : contient 2.8M tweets postés durant la coupe du monde de football 2014 et contenant les hashtas officiels des équipes (#BRA, #CRO, etc.) et de matchs (#BRACRO, #MEXCMR, etc.) ("http://linkfluence.com/en/",s. d.).



(a) (b)

FIGURE 2. *(gauche) Exemple de graphe multiplexe. (droite) Popularité des 10 hashtags les plus populaires de "Algeria" avec et sans mentions.*

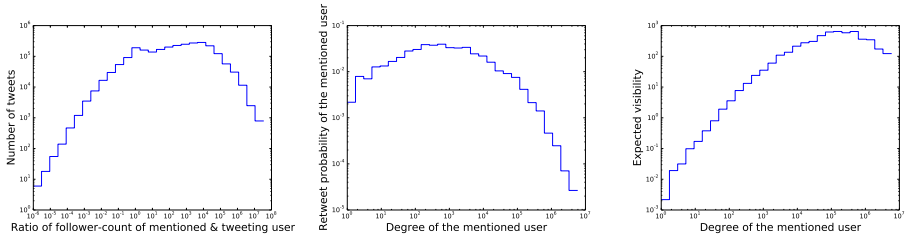
2.1. Représentation par graphe multiplexe

Pour un hashtag donné '#h', la représentation par graphe multiplexe contient deux couches : celle du bas correspond à la transmission aux abonnés, celle du haut correspond à la transmission utilisant la mention (Fig. 2(a)). Plus précisément, tous les utilisateurs ayant tweeté '#h' apparaissent dans la couche du bas et un arc connecte 'A' à 'B' si 'A' a (re)tweeté '#h' avant que 'B' ne retweete et 'B' est abonné à 'A'. Dans la couche du haut, un arc connecte 'C' à 'D' si 'C' tweete '#h' avant que 'D' ne retweete et 'C' *mentionne* 'D' dans son message ('D' peut être un follower de 'C' ou pas). Plus précisément, chaque couche est un ensemble de graphes dirigés acycliques (DAG). Nous appelons la racine de chaque DAG un initiateur et, afin de distinguer les différentes façons dont un initiateur a obtenu l'information, nous identifions deux classes d'initiateurs : les vrais initiateurs et initiateurs fictifs. Un *vrai initiateur* de '#h' est un utilisateur qui est une racine d'un DAG dans une des deux couches mais n'apparaît jamais comme non-racine d'un autre DAG. Ces utilisateurs ont donc débuté la diffusion de '#h' en réponse à une influence externe ou de leur propre chef. Un *initiateur fictif* est une racine d'un follow-DAG mais n'est pas la racine d'un mention-DAG. Un tel initiateur a donc reçu l'information d'un autre utilisateur et ré-initie une diffusion vers ses followers.

2.2. Importance des mentions et propriétés des utilisateurs mentionnés

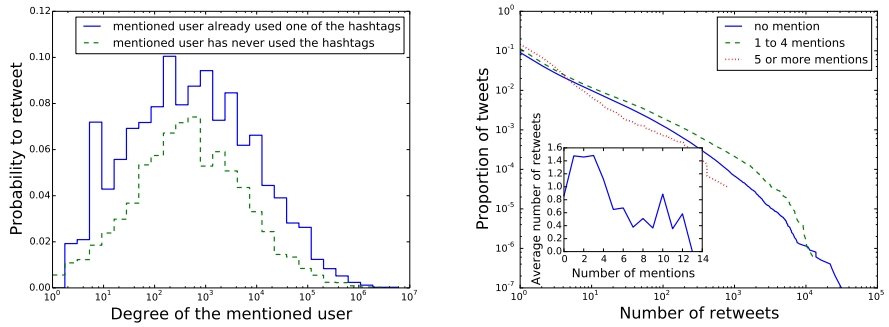
Avec cette représentation, nous pouvons mesurer l'impact des utilisateurs mentionnés dans la popularité, en définissant la popularité d'un hashtag par son nombre de retweets. Il est possible de sélectionner des hashtags très populaires et d'estimer la baisse de popularité en supprimant toutes les mentions. Pour cela, nous identifions tous les initiateurs fictifs (ensemble D) pour un hastag donné '#h' et tous les utilisateurs (ensemble S) qui n'appartiennent qu'aux DAG enracinés par un initiateur de D. L'activité de retweet de D est donc dépendante des mentions et nous mesurons cette dépendance à la mention par $(|S \cup D|)/n$. On peut observer que les hashtags les plus populaires sont très dépendants à la mention (Fig. 2(b)). Nous étudions maintenant les utilisateurs mentionnés via quelques propriétés simples :

Impact de la popularité et du taux de retweet : Pour confirmer une tendance à mentionner des utilisateurs populaires, nous traçons la distribution du ratio du nombre de followers des utilisateurs tweetant et étant mentionnés (Fig. 3(a)). Cette figure montre une tendance claire



(a) (b) (c)

FIGURE *3. (gauche) Probabilité de mentionner un utilisateur populaire, (milieu) Probabilité de retweeter des utilisateurs mentionnés, (droite) Visibilité attendue*



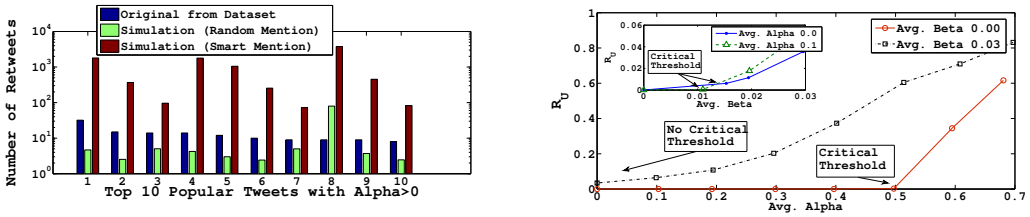
(a) (b)

FIGURE *4. (gauche) Probabilité de retweet pour un utilisateur mentionné dans "World-Cup", (droite) Nombre de mentions par tweet et nombre de retweets*

à mentionner des utilisateurs plus populaires qu'eux. En revanche, la Fig. 3(b) montre que la probabilité d'être retweeté diminue fortement si l'utilisateur mentionné a plus de 1000 abonnés où le degré représente le nombre des followers. Deux forces opposées agissent donc : les utilisateurs populaires retweetent moins mais donnent une forte visibilité aux tweets quand ils le font. Pour mesurer cela nous introduisons la mesure de "visibilité" qui est le produit du nombre de followers et de la probabilité de retweet (Fig. 3(c)).

Impact du contenu : La similarité entre le profil de l'utilisateur mentionné et le tweet reçu est un autre facteur qui détermine la probabilité de retweet. Dans "World-Cup" (voir Fig. 4(a)), nous avons calculé la probabilité qu'un utilisateur mentionné retweete (i) un tweet contenant au moins un hashtag qu'il a déjà utilisé (probabilité de 0.029) et (ii) un tweet ne contenant aucun hashtag qu'il a déjà utilisé (probabilité de 0.017). Ainsi, si l'utilisateur a déjà tweeté le hashtag en question, sa probabilité de retweet est presque doublée. La Fig. 4(a) montre également que cela dépend de la popularité des utilisateurs.

Nombre de mentions : Mentionner le bon nombre d'utilisateurs est également important pour obtenir de nombreux retweets. Dans "15M" nous observons que 23% des tweets contiennent des mentions. Parmi eux, 76% contiennent une mention, 17% contiennent deux, 4% contiennent trois et les 3% restant en contiennent plus de trois. La Fig. 4(b) montre que mentionner peu (2 ou 3) d'utilisateurs est toujours préférable à en mentionner beaucoup. En effet de nombreuses mentions indiquent un contenu du tweet plus court et donc certainement moins intéressant.



(a) (b)

FIGURE 5. *(gauche) Popularité réelle et simulée pour "Algeria" avec le même α (droite) Impact de α et β (Inset) sur la fraction d'utilisateurs infectés*

3. Modèle de simulation

Nous proposons un modèle basé sur SIR, SIR_{MF}, pour décrire et simuler la propagation de tweets dans le graphe multiplexe (Fig. 2(a)). Initiallement tous les individus sont susceptibles. Un individu v est infecté par un tweet T s'il le retweete à l'instant suivant. La transition $I \rightarrow R$ représente qu'un utilisateur infecté est supprimé directement à l'instant suivant. Nous supposons qu'il y a une seule information qui se propage et que chaque utilisateur ne peut le tweeter qu'une fois. La simulation s'arrête quand plus aucun individu ne peut être infecté. Nous supposons que l'infection est dirigée par trois facteurs (i) v doit être exposé à T, (ii) v doit être intéressé par T et (iii) v doit décider s'il retweete ou pas. Plus précisément, (i) un individu v peut être exposé à T par u si u tweete T et que v est un abonné de u ou si v n'est pas un abonné de u mais que u mentionne v dans T. Cela dépend entièrement de la structure du graphe multiplexe. (ii) L'intérêt de v pour T dépend du fait qu'il ait été exposé par une mention ou un lien direct. Comme les mentions sont plus visibles, nous utilisons deux distributions de Poisson de moyenne μ_1 et μ_2 normalisées entre 0 et 1 pour les liens de mention et d'abonnement respectivement. (iii) Le taux de retweet de v, κ_v, est modélisé par un loi puissance d'exposant κ normalisée entre 0 et 1 (Lerman, Ghosh,2010). Pour un individu v, nous notons α_v (resp. β_v) le taux d'infection par mention (resp. par abonnement). Un individu est donc infecté par mention avec probabilité $\alpha_v = f(\kappa_v, \mu_{1_v})$ et par abonnement avec probabilité $\beta_v = f(\kappa_v, \mu_{2_v})$. f peut simplement être le produit des deux probabilités. Enfin nous considérons qu'un utilisateur mentionne en moyenne λ individus par tweet. Le choix des (λ) individus à mentionner peut se faire suivant deux stratégies simples : **Mention aléatoire :** Les utilisateurs sont choisis au hasard parmi les susceptibles. **Mention optimale :** L'utilisateur u à mentionner est celui qui maximise $f_u \times \alpha_u$ parmi tous les susceptibles, où f_u est le nombre de followers de u. Cette stratégie correspond à maximiser le nombre attendu d'individu exposés.

3.1. Simulation

Nous simulons le modèle SIR_{MF} en utilisant le réseau des abonnements de "Algeria". Ce réseau est composé par N utilisateurs. Pour toutes les simulations nous avons utilisé $\lambda = 2$ et $\kappa = -2.5$ pour supposer que ces deux paramètres ne changent pas. Nous fixons μ_1 et μ_2 pour réguler les probabilités α et β et nous permettre d'observer l'impact de ces valeurs sur les cascades. Chaque résultat présenté dans la suite est une moyenne de 500 simulations et nous avons à chaque fois utilisé les stratégies aléatoire ou optimale en les comparant quand

cela avait du sens. Nous utilisons quatre métriques pour mesurer la propagation et évaluer les performances des algorithmes de recommandation de mentions : (a) le nombre de retweet des tweets avec mention R_U (dans les simulations il y a un seul tweet et il contient λ mentions), R_U est donc le nombre d'utilisateurs infectés, (b) le nombre de retweet des tweets sans mention NR_U, (c) la fraction de retweet dues aux mentions F_M et, (d) la fraction des utilisateurs mentionnés ayant retweeté FR_M. NR_U est inutile pour les simulations car nous n'utilisons que des tweets avec mentions. De même, FR_M n'est pas observable dans les simulations mais correspond au paramètre α. Nous utiliserons cependant ces métriques pour comparer les heuristiques dans la section 5.

3.2. Résultats

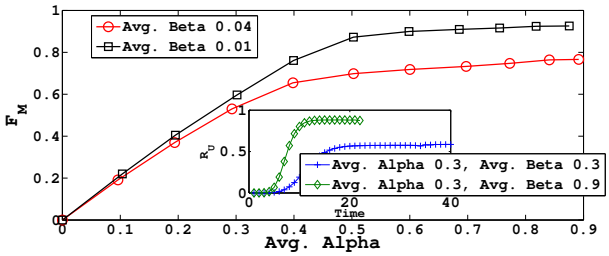


Figure 6. Impact de α sur la fraction d'utilisateurs retweetant à cause des mentions.
***Encart :** Augmentation du nombre d'utilisateurs retweetants au cours de la simulation.*

Tout d'abord nous étudions le taux de retweet R_U des 10 tweets les plus populaires ayant $\alpha > 0$ dans "Algeria". Nous simulons le modèle pour chacun de ces 10 tweets sur le réseau d'abonnement de "Algeria" en estimant μ_1 et μ_2 de sorte que les probabilités d'infection α et β soient proches des observations. De plus, nous initions les diffusions des mêmes individus avec le même nombre de mentions qu'observé. Nous observons Fig. 5(a) que pour presque tous les tweets les choix réels sont meilleurs que la stratégie aléatoire mais moins bons que le choix optimal. Cela corrobore notre intuition selon laquelle il est possible d'augmenter le taux de retweet en choisissant les mentions de manière adéquate. Nous étudions maintenant l'impact des différents paramètres du modèle SIR_{MF} sur le taux de retweet :

Impact des taux d'infection α et β : L'encart de la Fig. 5(b) montre qu'en dessous d'une valeur critique de β, les tweets sont très peu retweetés. Au-dela de cette valeur le taux de retweet augmente par contre presque linéairement. Cette valeur critique de β décroît quand α croît. Un effet similaire est observé si β est fixé et que α varie (Fig. 5(b)). Ces seuils peuvent être calculés analytiquement (voir section 4). La Fig. 6 montre que le nombre d'individus retweetants augmente avec le temps mais que, pour un taux d'infection donné, on observe une transition de phase qui aboutit à une épidémie. Si l'on observe l'impact de α sur les utilisateurs retweetant grâce aux mentions, F_M, on voit (Fig. 6) que pour α fixé, F_M décroît quand β croît. C'est naturel car si β est élevé, plus d'individus retweetent par les liens d'abonnements ce qui diminue F_M.

Impact de λ : Augmenter λ améliore la popularité R_U. Ceci dépend cependant de β : si β est faible, de nombreuses mentions sont utiles. Cependant quand β augmente, l'impact de λ sur R_U diminue du fait que les liens d'abonnement deviennent le facteur principal d'infection.

Impact des utilisateurs mentionnés : Nous étudions cela suivant deux angles. Tout d'abord l'encart de la Fig. 7 montre que la stratégie optimale est très pertinente si β est faible. L'aug-

mentation de β réuit l'écart. Ensuite, bien que la stratégie optimale semble le bon choix, il n'est pas concevable de l'utiliser en pratique. Nous étudions donc des stratégies plus réalistes comme par exemple de ne considérer que les voisins à distance 1 dans le réseau d'abonnement (abonnés non réciproques, abonnement non réciproques ou réciproques). La Fig. 7 montre que le plus pertinent est de mentionner les abonnements non réciproques. Ceci peut se justifier du fait que les abonnés sont en général autant ou moins populaires alors que les abonnements sont au contraire plus populaires. Un retweet d'un abonnement non réciproque rend donc le tweet plus largement accessible.

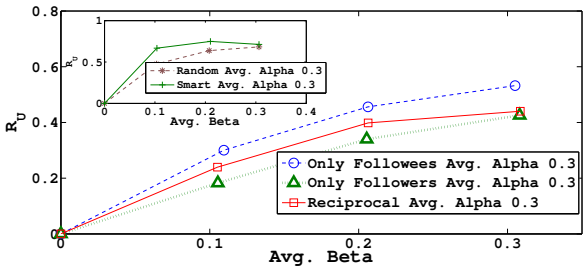


Figure 7. Impact de la relation entre les utilisateurs mentionnant et mentionné. ***Encart :*** *Impact de la stratégie.*

4. Analyse du modèle SIR_{MF}

Nous dérivons maintenant le seuil visible sur la Fig. 5(b) pour le modèle SIR_{MF} en utilisant une approche MMCA (Microscopic Markov Chain Approach). Les équations sont obtenues à partir du schéma de la Fig. 8(a) qui décrit les probabilités de transition entre les états. À l'instant t, un utilisateur susceptible v mentionné par un utilisateur infecté devient susceptible mentionné (MS), les autres sont susceptibles non mentionnés (US). À l'instant suivant il peuvent retweeter et devenir infectés ou retourner dans l'état susceptible. Les équations qui régissent les transitions sont les suivantes (nous ne détaillons pas les calculs) :

$$p_v^S(t+1) = p_v^{US}(t)q_v^{US}(t) + p_v^{MS}(t)q_v^{MS}(t) \tag{1}$$

$$p_v^I(t+1) = p_v^{US}(t)(1-q_v^{US}(t)) + p_v^{MS}(t)(1-q_v^{MS}(t)) - p_v^I(t) \tag{2}$$

$$p_v^R(t+1) = p_v^R(t) + p_v^I(t) \tag{3}$$

Nous validons ces résultats en les comparant avec des simulations Monte-Carlo (MC). Dans ces simulations nous débutons avec un seul individu infecté, choisissons les paramètres adéquats puis déroulons la diffusion. La Fig. 8(b) montre l'accord entre les deux approches ce qui valide les équations précédentes. Il faut aussi noter que MMCA donne des valeurs plus élevées que MC ce qui vient du fait que MMCA suppose les événements indépendants. Il est également possible de déterminer le seuil épidémique à partir des équations MMCA (à nouveau nous ne détaillons pas les calculs en nous nous basons sur l'approche (Youssef, Scoglio,2011)) : $\beta_c^{MC} = \frac{1-\alpha\lambda}{0.7\Lambda_{max}(A)}$, où A est la matrice d'adjacence du réseau d'abonnement et $\Lambda_{max}(A)$ sa plus grande valeur propre. Pour $\alpha > \frac{1}{\lambda}$, β_c n'existe pas. L'équation donne des résultats très similaires à ceux de la Fig. 5(b). For example, when $\alpha = 0.03$, there does not exist β_c. And when $\alpha = 0$, $\beta_c = 0.5$.

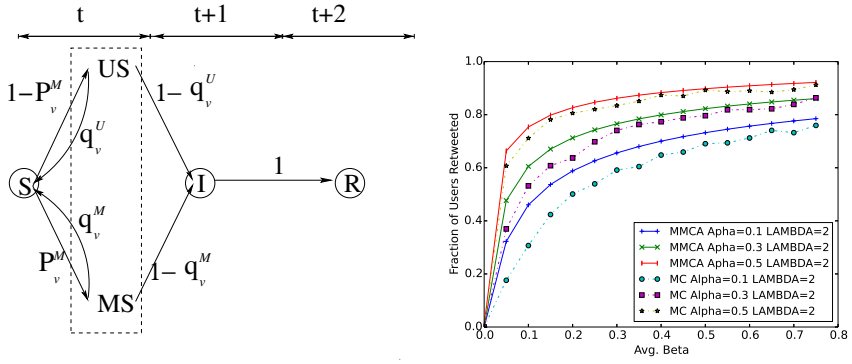


FIGURE 8. *(gauche) Probabilités de transition du modèle SIR_{MF}. Chaque utilisateur susceptible est mentionné (MS) ou pas (US). (droite) Comparaison des fractions d'utilisateurs ayant retweeté R_U en utilisant le modèle MMCA et l'approche MC.*

5. 'Easy-Mention': heuristique de Recommandation

Le but du système 'Easy-Mention' est de recommander à un utilisateur postant un tweet le meilleur ensemble d'utilisateurs à mentionner afin de maximiser le nombre de retweets. Par conséquent, l'entrée du système est le tweet soumis et la sortie est une liste d'utilisateurs susceptibles d'être mentionné. Nous devons d'abord déterminer l'espace de recherche. Pour rendre l'heuristique rapide il est nécessaire de pouvoir récupérer les informations pertinentes à l'exécution en tenant compte des restrictions de l'API Twitter. Nous sélectionnons uniquement les abonnements non réciproques pour les raisons suivantes : (i) ils ne peuvent obtenir le tweet que par une mention (ii) dans la réalité, les gens préfèrent mentionner leurs abonnements non-réciproques à leurs abonnés(iii) en général, les abonnements non réciproques ont plus d'abonnés ce qui augmente la visibilité attendue. Une fois les utilisateurs potentiels sélectionnés, 'Easy-Mention' attribue un score de qualité à chacun. Ce score évalue essentiellement le gain de popularité attendue du tweet T, si u est mentionné dans T. Les études précédentes montrent que les facteurs suivants peuvent réguler le score de qualité : (i) le nombre d'abonnés (ii) le taux de retweet (iii) le similarité de contenu entre T et le profil de l'utilisateur mentionné u. On obtient alors le score suivant : $S(u,T) = f_P(u)^x \times f_R(u)^y \times f_I(u,m)^z$ où $f_P(u)$ est le nombre d'abonnés normalisé de u, $f_R(u)$ est son taux de retweet normalisé et $f_I(u,T)$ capture la similarité [1]. Les différents exposants permettent de contrôler l'importance relative des trois paramètres et devraient être choisis en fonction de l'utilisateur. Par exemple les utilisateurs peu actifs sont beaucoup plus sensibles à la similarité de contenu que les utilisateurs moins actifs.

5.1. Protocole expérimental

Nous allons maintenant évaluer la performance de 'Easy-Mention' et la comparer avec l'état de l'art. Pour cela, (a) nous mettons en $\frac{1}{2}$uvre un modèle de retweet standard qui simule

1. Calculé comme la similarité cosinus entre le vecteur des termes contenus dans le tweet et les termes dans les tweets récents de u.

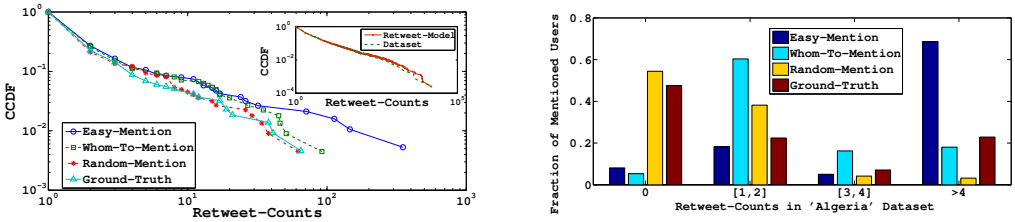


FIGURE 9. *(gauche) CCDF du nombre de retweet des tweets après l'utilisation de stratégies de mention différents.* ***Encart :*** *Comparaison de la distribution de la popularité des tweet à partir de "Algeria Dataset" et du modèle. (droite) Comparaison des taux de retweets sur "Algérie" des utilisateurs mentionnés.*

la propagation des tweets via l'activité de retweet puis, (b) au dessus de ce modèle nous déployons les algorithmes de recommandation de mention. Nous choisissons le modèle de retweet bien accepté de Vespignani et al. (Weng *et al.*,2012) qui est un modèle à base d'agent parcimonieux et que nous adaptons pour prendre en compte les mentions.

D'abord nous sélectionnons un ensemble de tweets D_T à partir de l'ensemble de données de telle sorte que 50% des tweets contient mentionne et 50% d'entre eux ne la contient pas. À tout moment, un utilisateur est choisi de préférence basée sur son retweet-taux pour poster un tweet ou retweet. Si il choisit de poster un tweet nouveau, un tweet est choisi au hasard parmi D_T et il le tweete avec le même nombre de mentions (y compris zéro) comme dans le tweet d'origine. Les utilisateurs spécifiques pour être mentionnés dans ce tweet est réglementé par l'algorithme de "recommandation de mention" (mention-recommendation). L'autre option est qu'il choisit de retweeter un poste déjà reçu. Pour chaque utilisateur u, nous maintenons une "fenêtre de l'écran" et une "fenêtre de mention" où les tweets reçus via des liens de follow et les tweets reçus via les liens de mention sont stockés respectivement. Si l'utilisateur sélectionné choisit de retweeter, l'une de ces deux fenêtres est choisie en fonction du profil de similarité de u (calculés comme la similarité cosinus des termes), ensuite le poste le plus similaire (en référence au profil de u) dans la fenêtre est retweeté. Cependant, il y a une possibilité équitable de ne retweeter aucun poste, si la similarité est inférieure à un seuil.

Afin de valider, nous simulons ce modèle retweet sur "Algérie" avec des tweets ne contenant pas de mentions. Il est réconfortant pour nous d'observer que le résultat explique l'hétérogénéité dans la distribution de popularité de tweets avec une précision raisonnable (voir l'encadré de la Fig. 9(a)). Maintenant, nous sommes prêts à utiliser ce modèle pour évaluer les heuristiques de recommandation de mention. En plus de ce modèle de retweet, nous appliquons l'heuristique 'Easy-Mention' et comparons ses performances avec les algorithmes "Whom-to-mention" (Wang *et al.*,2013), "Random mention" qui consiste à mentionner des utilisateurs au hasard et "Ground Truth" qui consiste à mentionner les utilisateurs du jeu de données.

5.2. Résultats

Nous effectuons les expériences sur "Algérie" (tweets et réseau des abonnements). Les mesures d'évaluation sont déjà décrites dans la section 3.1. Dans cette expérience, en affi-

chant un tweet T, nous enlevons les mentions originales du tweet T et remplaçons par celles choisies par l'algorithme étudié. Pour assurer l'équité, nous conservons le nombre de mentions du tweet original. Une fois les utilisateurs choisis, nous simulons le modèle de retweet. La Fig. 9(a) illustre clairement le fait que "Easy-Mention" surpasse les autres algorithmes en obtenant plus de retweets. Dans le Tableau 1, nous énumérons les mesures d'évaluation observées pour les différents algorithmes. "Easy-mention" arrive à mentionner des utilisateurs qui retweetent souvent ce tweet (FR_M élevé), et sont assez populaires pour augmenter la visibilité (le nombre moyen des followers des utilisateurs recommandés par "Easy-Mention" est 1317, contre 932 pour "Whom-To-Mention"). Cela permet d'obtenir plus de retweets pour les tweets avec mentions (R_U) que ceux sans mention (NR_U). La Fig. 9(b) souligne le fait que les utilisateurs mentionnés par "Easy-Mention" retweetent plus, ce qui contribue directement à la propagation. En résumé, "Easy-Mention" permet de populariser efficacement les tweets en créant plus de cascades plus grandes.

Tableau 1. Résultats pour les différentes stratégies de mention.

$Algorithmes$	R_U	F_M	$R_U - NR_U$	FR_M
$Easy - Mention$	4.13	0.048	3.87	0.101
$Whom - To - Mention$	3.92	0.012	3.50	0.024
$Random - Mention$	3.37	0.011	2.48	0.022

6. Conclusion

Dans cet article, nous présentons la première étude approfondie expliquant le rôle de la mention dans la propagation de l'information sur Twitter. Pour cela, nous avons utilisé une modélisation par réseau multiplexe qui capture efficacement la dynamique de propagation utilisant les liens directs et de mention. De plus, nous avons proposé la notion de dépendance à la mention pour mesurer l'amélioration de popularité due aux mentions. Nous avons identifié qu'une fraction significative (allant jusqu'à $50\% - 60\%$) des retweets pourrait disparaître si les gens arrêtaient d'utiliser les mentions. Nous avons également adapté le modèle SIR classique en SIR_{MF} pour les réseaux multiplexes en distinguant les liens directs et les liens de mention. En utilisant le MMCA basé sur un modèle analytique, nous avons pu mesurer les taux de retweets critiques (α et β) pour les informations à propager. Ces expériences nous ont permis d'identifier trois paramètres contrôlant l'efficacité des mentions : nombre de followers, taux de retweet et similarité de contenu que nous avons utilisé dans un outil de recommandation de mentions en ligne, "Easy-Mention". "Easy-Mention" dépasse la méthode de l'état de l'art "Whom-To-Mention" et, contrairement à la plupart des approches similaires, il peut facilement donner des réponses à l'exécution. En évaluant l'algorithme proposé, nous avons constaté que l'importance relative de ces paramètres varie pour différents types d'utilisateurs. Par exemple, les utilisateurs très actifs donnent moins d'importance à la similarité des tweets reçus. Il conviendra donc d'intégrer une composante apprentissage dans notre algorithme afin de choisir au mieux la meilleure combinaison de paramètres pour chaque utilisateur.

Remerciements

Ce travail a été partiellement soutenu par le SAP Labs India Doctoral Fellowship program, le projet DST - CNRS Indo - Français 'Evolving Communities and Information Spreading' et le projet ANR CODDDE ANR-13-CORD-0017-01.

Bibliographie

Borge-Holthoefer J., Rivero A., Moreno Y. (2012). Locating privileged spreaders on an online social network. *Physical review E*, vol. 85, n° 6, p. 066123.

Cha M., Haddadi H., Benevenuto F., Gummadi P. K. (2010). Measuring user influence in twitter: The million follower fallacy. *ICWSM*, vol. 10, n° 10-17, p. 30.

Chen W., Wang Y., Yang S. (2009). Efficient influence maximization in social networks. In *Acm sigkdd international conference*, p. 199–208.

González-Bailón S., Borge-Holthoefer J., Rivero A., Moreno Y. (2011). The dynamics of protest recruitment through an online network. *Scientific reports*, vol. 1.

Granell C., Gómez S., Arenas A. (2013). Dynamical interplay between awareness and epidemic spreading in multiplex networks. *Physical review letters*, vol. 111, n° 12, p. 128701.

http://dfreelon.org/2012/02/11/arab-spring-twitter-data-now-available-sort-of/. (s. d.).

http://linkfluence.com/en/. (s. d.).

Kato S., Koide A., Fushimi T., Saito K., Motoda H. (2012). Network analysis of three twitter functions: Favorite, follow and mention. In *Knowledge management and acquisition for intelligent systems*, p. 298–312. Springer.

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

Lee K., Mahmud J., Chen J., Zhou M., Nichols J. (2014). Who will retweet thisfi. In *Proceedings of the 19th international conference on intelligent user interfaces*, p. 247–256.

Lerman K., Ghosh R. (2010). Information contagion: An empirical study of the spread of news on digg and twitter social networks. *ICWSM*, vol. 10, p. 90–97.

Malhotra A., Malhotra C. K., See A. (2012). How to get your messages retweeted. *MIT Sloan Management Review*, vol. 53, n° 2, p. 61–66.

Petrovic S., Osborne M., Lavrenko V. (2011). Rt to win! predicting message propagation in twitter. In *Icwsm.*

Suh B., Hong L., Pirolli P., Chi E. H. (2010). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social computing (socialcom)*, p. 177–184.

Tang L., Ni Z., Xiong H., Zhu H. (2014). Locating targets through mention in twitter. *World Wide Web*, p. 1–31.

Uysal I., Croft W. B. (2011). User oriented tweet ranking: a filtering approach to microblogs. In *Proceedings of the 20th acm international conference on information and knowledge management*, p. 2261–2264.

Wang B., Wang C., Bu J., Chen C., Zhang W. V., Cai D. *et al*. (2013). Whom to mention: Expand the diffusion of tweets by @ recommendation on micro-blogging systems. In *Proceedings of the 22nd international conference on world wide web*, p. 1331–1340.

Weng L., Flammini A., Vespignani A., Menczer F. (2012). Competition among memes in a world with limited attention. *Sci. Rep.*, vol. 2, n° 335. http://dx.doi.org/10.1038/srep00335

Youssef M., Scoglio C. (2011). An individual-based approach to sir epidemics in contact networks. *Journal of theoretical biology*, vol. 283, n° 1, p. 136–144.

Zhan Q., Zhang J., Wang S., Yu P. S., Xie J. (2015). Influence maximization across partially aligned heterogenous social networks. In *Pakdd*, p. 58–69.