Couverture par Ensembles Flous et Maximum d'Influence en Marketing Viral

Stefan Janaqi

LGI2P – EMA, 69, rue G.Besse, 30035 Nimes
stefan.janaqi@mines-ales.fr

RESUME. Le Marketing Viral repose sur le choix, dans un réseau social, d'un ensemble initial d'individus I et la diffusion de son influence. La maximisation de f(I) – l'espérance du nombre de personnes influencées par I, est NP-Difficile pour plusieurs modèles de diffusion. Pour des modèles de diffusion standards, f(I) est sous modulaire et un simple algorithme glouton fournit de bons ensembles initiaux. Pour d'autres modèles de diffusion comme RUC, f(I) n'est pas sous modulaire. Néanmoins, pour tous ces modèles, la complexité de calcul de f(I) reste élevée. Nous présentons une borne inférieure exacte g(I) ≤ f(I) basée sur les opérateurs d'ensembles flous. Nous observons que g(I) est sous modulaire indépendamment du modèle de diffusion. La maximisation de g(I) est NP-Difficile mais, la complexité de calcul d'une bonne approximation de g(I) est de deux ordres inférieure que celle de f(I) permettant le passage à l'échelle. Nous illustrons la validité de notre approche sur des graphes jouets ainsi que sur un graphe de 83K sommets extrait de co-publications dans le domaine biomédical.

ABSTRACT. Viral Marketing is based on the choice, in a social network, of an initial set of individuals I and the diffusion of its influence. Maximizing f(I) – the expected number of persons influenced by I, is NP-Hard for several diffusion models. For standard diffusion models f(I) is submodular and a simple greedy algorithm finds good initial sets. For other diffusion models like RUC, f(I) isn't submodular. Nevertheless, for all models, the complexity of computing f(I) remains high. We present a tight lower bound g(I) ≤ f(I) based on the fuzzy set operators. We prove that g(I) is submodular independently from diffusion model. Maximizing g(I) is NP-Hard but, the complexity of a good approximation for g(I) is two orders less than that of f(I) allowing to scale to big networks. We illustrate the validity of our approach on toy graphs as well as on a real graph of 83K nodes extracted from co-author publications in biomedical domain.

MOTS-CLES : Marketing Viral, Opérateurs Flous, Modèles de Diffusion, Sous modularité, Complexité, Algorithme Glouton.

KEYWORDS: Viral Marketing, Fuzzy Operators, Diffusion Models, Submodularity, Complexity, Greedy Algorithm.

DOI:XXXXXXXXX

1. Introduction

L'influence dans les réseaux sociaux a suscité un intérêt croissant récemment. La fouille et l'estimation de cette influence sont motivées entre autres par des applications en marketing [BRU 08], [CAR 08] où, la décision de démarcher un ensemble d'individus est basée non seulement sur le *marketing direct* mais aussi sur le *marketing viral*. Le paradigme du marketing viral [KEM 03] s'énonce comme suit : en démarchant initialement un petit nombre d'individus influents dans un réseau, une cascade récursive se déclenche par des amis qui recommandent le produit à leurs amis, ... Le réseau d'influence peut être vu comme un graphe orienté $G = (V, E, w)$ où, V est l'ensemble des individus et E l'ensemble d'arcs $(u, v) - u$ influe v. Une estimation de l'influence est fournie $w : E \to R$, telle que pour tout arc $(u, v) \in E, w(u, v)$ est l'influence de l'individu u sur v. L'estimation des $w(u, v)$ dans les réseaux est un domaine d'activité en plein essor qui suit de près les nouvelles technologies de communication [EAG 11]. Tout au long du papier on notera $m = |E|$ et $n = |V|$.

Soit $f(I)$ – l'espérance du nombre de personnes influencées à partir d'un ensemble inital I suivant une méthode de diffusion D dans le graphe G. Un sommet (individu) $v \in V$ est *actif* s'il est influencé et *inactif* autrement. Pour tout modèle D de diffusion d'influence, soit $p(v \mid I, D, G)$ la probabilité que v devienne actif. Lorsque G et D sont connus, on notera cette probabilité par $p(v \mid I)$. On définit ainsi, pour tout sommet v, une variable aléatoire $x(v) \in \{0, 1\}$, 1 si v devient actif et 0 sinon, avec $p(x(v) = 1) = p(v \mid I)$, $p(x(v) = 0) = 1 - p(v \mid I)$. La cardinalité de l'ensemble actif final est une variable aléatoire $X = \sum_{v \in V} x(v)$ et son espérance est $f(I) = E(X) = \sum_{v \in V} E\big(x(v)\big) = \sum_{v \in V} p(v \mid I)$. Le problème central de ce papier est la recherche efficace de l'ensemble initial I^f maximisant $f(I)$.

Par la suite nous utiliserons le modèle de diffusion *Linear Threshold* (LT) proposé par Granovetter [GRA 78] et Schelling [SCH 71]. Ce modèle est au cœur de plusieurs généralisations ultérieures. L'idée de base de LT consiste à calculer, pour un sommet inactif v, la somme des influences des voisins actifs de v : $infl(v) = \sum_{u \in A \cap N^-(v)} w(u, v)$, où A est l'ensemble actuel de sommets actifs et $N^-(v)$ est l'ensemble de voisins entrants de v. Ensuite, v devient actif si $infl(v)$ est supérieure au seuil d'activation $t(v)$ de v.

La maximisation de $f(I)$ est un problème d'optimisation combinatoire NP-Difficile [KEM 03]) pour la plupart des modèles de diffusion standard (dont LT) utilisés dans l'analyse des réseaux sociaux. D'autres modèles de diffusion, centré utilisateur, nommés RUC [LAG 13] donnent des problèmes d'optimisation NP-Difficiles aussi. Il est alors nécessaire d'approximer I^f par des méthodes heuristiques. Dans [KEM 03] les auteurs démontrent que $f(I)$ est une fonction sous modulaire pour LT (et aussi pour d'autres modèles dont IC – *Independent Cascade* [SCH 78]). Pour le modèle RUC, $f(I)$ n'est pas sous modulaire [LAG 13]. L'intérêt

de la sous modularité vient du fait qu'un simple algorithme glouton, donne $I^f_{glouton}$ tel que $f(I^f_{glouton}) \geq 0.63\, f(I^f)$ [NEM 78]. C'est une bonne garantie sur la qualité de $I^f_{glouton}$ mais, elle ne peut pas être utilisée pour des modèles de diffusion comme RUC. En dehors de cette garantie, le problème de l'estimation efficace de $f(I)$ reste ouvert [CHE 11], [XIA 13], [SAI 08], [KEM 03]. L'évolution de l'ensemble actif par les méthodes de diffusion étant itérative, il parait difficile de trouver une formule fermée pour l'espérance $f(I)$. A ce jour, le procès de diffusion est initialisé R fois avec des seuils aléatoires et $f(I)$ est approximée par la valeur moyenne des résultats de chaque lancement. Nous estimons, pour I et LT fixés, la complexité de calcul de $f(I)$: $N_{LT} \leq R(nm + 2n^2 + 2n + 1) \sim O(R\, n\, m)$. On montre (voir section 2) que le calcul de l'approximation de I^f par l'algorithme glouton est de complexité HN_{LT} avec $H = \sum_{k=0}^{|I|-1}(n-k)$.

Notre objectif est de calculer un « bon » ensemble initial avec une moindre complexité. Notre approche est basée sur l'observation simple que pour tout ensemble initial I, les probabilités $p(v \mid I), v \in V$ peuvent être vues comme des degrés d'appartenance flous de $v \in V$. La manipulation par des opérateurs flous de ces ensembles nous permettra de produire une borne inférieure exacte $g(I) \leq f(I)$. Le calcul de I^g maximisant $g(I)$ est basé sur l'estimation, pour chaque sommet $u \in V$, des probabilités d'activation $p(v \mid \{u\}), v \in V$. Le calcul de cette matrice « d'ordre un » $\boldsymbol{P}$, $\boldsymbol{P}(v,u) = p(v \mid \{u\})$, est commun pour la recherche de I^f ainsi que de I^g et nécessite $n\, N_{LT}$ opérations. La recherché de I^g est NP-Difficile. Nous démontrons que $g(I)$ est sous modulaire *indépendamment* de la méthode de diffusion et utilisons un algorithme glouton pour la maximiser. Etant donnée la matrice $\boldsymbol{P}$, le calcul de $I^g_{glouton}$ ne passe plus par le procès de diffusion et nous montrons que seulement $|I|n^2$ opérations supplémentaires seront nécessaires, alors que le calcul de $I^f_{glouton}$ nécessitera $N_{LT}(H-n)$ opérations supplémentaires. Le gain en complexité est $O(R\, m)$.

Nous testons la validité de notre approche sur des graphes jouets mais aussi sur un graphe réel extrait de co-publication de 82999 auteurs sur 43000 articles dans le domaine biomédical PubMed (http://www.ncbi.nlm.nih.gov/pubmed).

2. Complexité de modèles de diffusion et algorithme glouton

Le résultat principal de [KEM 03] est que pour une grande classe de modèles de diffusion D (LT, IC et leurs généralisations), l'espérance de la cardinalité de l'ensemble actif final est une fonction monotone croissante et sous modulaire. Ces fonctions $F : 2^V \rightarrow R$ vérifient : (i) $F(\emptyset) = 0$; (ii) (monotonie) $A \subseteq B \subseteq V, F(A) \leq F(B)$; (iii) (sous modularité) $\forall A, B \subseteq V, F(A \cup B) + F(A \cap B) \leq F(A) + F(B)$.

Le fait important est que pour les fonctions sous modulaires, l'algorithme glouton retourne $I^f_{glouton}$ avec un bon ratio d'approximation $\frac{f(I^f)}{f(I^f_{glouton})} \leq \frac{e}{e-1}$. Ce

ratio est déduit des travaux de [NEM 78]. Ici, $e = 2.7183\ldots$ est la base du logarithme naturel. Cette inégalité garantit que le résultat de l'algorithme glouton sera supérieur à 63% du maximum pour les fonctions sous modulaires et croissantes.

La maximisation de $f(I)$ étant un problème combinatoire NP-Difficile, une heuristique naturelle pour approximer l'optimum I^f est l'algorithme glouton ci-dessous :

EnsembleInitialViralGlouton(G, D, k)
Entrée: G – graphe d'influence ; D – méthode de diffusion ; k – cardinalité d'ensemble initial;
Retourne: I – ensemble viral initial

1. $I \leftarrow \emptyset$
2. **tantque** $|I| < k$
3. **trouver** $u^* \leftarrow argmax_{u \in V \setminus I}(f(I \cup u))$
4. $I \leftarrow I \cup u^*$
5. **fintantque**

Figure 1. Pseudo code de l'algorithme glouton pour maximiser $f(I)$.

Calculons maintenant la complexité de l'estimation de $f(I)$ pour un ensemble initial I. Comme nous l'avons signalé en introduction, le calcul efficace de $f(I)$ reste ouvert. Il s'agit d'estimer l'espérance d'une variable aléatoire. On fixe alors un nombre R d'essais et pour tout $r, r = 1, \ldots, R$ un modèle de diffusion D initialisé par I, itère pendant S étapes. Au pire des cas S peut atteindre $n - |I|$. Soit I_r l'ensemble actif final trouvé à l'essai r. La loi des grands nombres donne un bon estimateur de $f(I)$: $\tilde{f}(I) = \frac{1}{R} \sum_{r=1}^{R} |I_r|$, la moyenne des cardinalités de I_r, $r = 1, \ldots, R$. On sait aussi que la vitesse de convergence de $\tilde{f}(I)$ vers $f(I)$ est de l'ordre de $\frac{1}{\sqrt{R}}$. Par conséquent, si l'on cherche une estimation $\left|\tilde{f}(I) - f(I)\right| < 10^{-2}$ il faut choisir $R \approx 10^4$.

La complexité de EnsembleInitialViralGlouton dépend du modèle de diffusion D. Nous donnons ci-dessous la complexité de LT. Notons les voisins entrants de u par $N^-(u) = \{v \in V \mid (v, u) \in E\}$. Tout modèle de diffusion fournit, à partir d'un ensemble initial A_0, une suite croissante d'ensembles $A_0 \subseteq A_1 \subseteq A_2 \ldots$. Ici, A_s contient les sommets de A_{s-1} et les voisins de A_{s-1} qui viennent de s'activer par le modèle D. Ce procès itératif s'arrête lorsqu'il n'y a plus de nouveaux sommets activés. Pour LT, soit u un sommet inactif à l'étape s, $u \notin A_s$. L'influence globale exercée sur u est une fonction $infl(u, N^-(u) \cap A_s) = \sum_{v \in N^-(u) \cap A_s} w(v, u)$. Si $infl(u, N^-(u) \cap A_s)$ est supérieures au seuil d'activation $t(u)$ de u alors, u devient actif à l'étape $s + 1$.

Une analyse simple montre que la complexité d'estimation de $f(I)$ pour LT est $O(Rnm)$ (voir le pseudo code ci-après). La ligne 3 nécessite n opérations. La ligne 4 demande 1 seule opération. La boucle 6 – 8 nécessite au plus $N^-(u)$ additions pour chaque $u \in V \setminus A_s$. Ainsi, on aura au plus $\sum_{u \in V \setminus A_s} N^-(u) \leq m$ additions pour réaliser cette boucle. La boucle 9 – 11 nécessite au plus 2 opérations pour chaque $u \in V \setminus A_s$, et comme $|V \setminus A_s| < n$, cette boucle demande au plus $2n$ opérations. La ligne 12 demande 3 opérations. Par conséquent, la boucle 5 – 13 nécessitera au plus $S(m + 2n + 3)$ opérations et le nombre total d'opérations est : $N_{LT} \leq R(n + 1 + S(m + 2n + 3))$. Comme le nombre d'étapes de diffusion $S < n$, on obtient $N_{LT} \leq R(nm + 2(n + 1)^2 - 1) \sim O(Rnm)$.

LinearThresholdDiffusion(G, b, R, S, A)
1. $\tilde{f}(A) \leftarrow 0$
2. **faire pour** $r \leftarrow 1 \ldots R$
3. $t \leftarrow rand(1, n)$ // seuil aléatoire uniforme de chaque sommet
4. $A_s \leftarrow A$
5. **faire pour** $s \leftarrow 1 \ldots S$
6. **faire pour** $u \in V \setminus A_s$
7. $infl(u) \leftarrow \sum_{v \in N^-(u) \cap A_s} w(v, u)$
8. **finpour**
9. **faire pour** $u \in V \setminus A_s$
10. **si** $infl(u) \geq t(u)$ **alors** $A_s \leftarrow A_s \cup u$
11. **finpour**
12. $\tilde{f}(A) \leftarrow \tilde{f}(A) + |A_s|/R$
13. **finpour**
14. **finpour**

Figure 2. Pseudo code de Linear Threshold et l'estimation de $f(I)$.

Nous pouvons maintenant fournir la complexité de l'algorithme EnsembleInitialViralGlouton pour le calcul de $I^f_{glouton}$. L'opération nécessitant le plus de calcul est l'estimation de $f(I \cup u)$ (**trouver** $u^* \leftarrow argmax_{u \in V \setminus I}(f(I \cup u))$). Le nombre d'opérations est majoré par HN_{LT} , où $H = \sum_{k=0}^{|I|-1}(n - k)$. Pour LT cela donne : $T_{LT} = \left(|I|n - \frac{1}{2}|I|(|I| - 1)\right) O(Rnm) \sim O(|I|\, R\, n^2\, m)$. Nous avons vu que pour une approximation de $\left|\tilde{f}(I) - f(I)\right| < 10^{-2}$ il faut choisir $R \approx 10^4$. Par la suite nous présentons la recherche de I dans un graphe avec $n, m \approx 10^5$. Ainsi, l'algorithme glouton, bien que polynomial, nécessiterait $O(10^{20})$ opérations pour LT. Le besoin de réduire cette complexité nous a conduits à chercher une méthode plus efficace et qui ne détériore pas la qualité de l'approximation.

3. Influence comme un problème de couverture par ensembles flous

La recherche de $I^f_{glouton}$ par EnsembleInitialViralGlouton passe par la recherche du premier meilleur sommet u et nécessite l'estimation de $f(u), u \in V$. Par conséquent, cette première étape fournit les vecteurs de probabilité $p(v \mid \{u\}), v \in V$ pour tout $u \in V$. Soit $\boldsymbol{P}$ la matrice $n \times n$ contenant les colonnes $\boldsymbol{P}_u = p(v \mid \{u\}), v \in V$. Le calcul de cette matrice nécessite nN_{LT} opérations. Dans cette section nous cherchons un « bon » ensemble initial I_1, l'indice '1' se référant aux informations d'ordre un contenues dans $\boldsymbol{P}$. Notre objectif est de réduire la complexité de l'approximation de I^f. De l'autre coté, il doit être clair que $f(I_1)$ sera inférieur au $f(I^f)$. Sans résultat théorique sur la relation entre $f(I_1)$ et $f(I^f)$, nous donnons dans la section suivante une estimation de la validité de $f(I_1)$ sur des graphes jouets et un réseau réel.

Soit donné pour un ensemble $I \subseteq V$ le vecteur $p(v \mid I), v \in V$. La définition d'un ensemble flou nécessite un ensemble de base, ici V, et un degré d'appartenance pour tout $v \in V$, $p(v \mid I) \in [0,1]$. L'ensemble V étant fixé, nous identifierons les ensembles flous avec les vecteurs $\mathbf{p} = p(v \mid I), v \in V$. La théorie des ensembles flous ([DUPR 80], [YAG 82], [ZAD 65]) fournit des opérateurs pour ensembles flous, analogues aux opérateurs classiques. Le complément flou *standard* de $\mathbf{p}$ est $NOT_{flou}(\mathbf{p}) = 1 - \mathbf{p}$. L'union et l'intersection standards de $\mathbf{p}$ et $\mathbf{q}$ sont $OR_{flou}(\mathbf{p}, \mathbf{q}) = \max(\mathbf{p}, \mathbf{q})$ et $AND_{flou}(\mathbf{p}, \mathbf{q}) = \min(\mathbf{p}, \mathbf{q})$. Notre intérêt des ensembles flous est justifié par les propriétés simples qui suivent.

Propriété 1. Pour tout ensemble $I \subseteq V$, $p(v \mid I) \geq OR_{flou}\{p(v \mid \{i\}), i \in I\}$.

Preuve. Pour tout $i \in I, p(v \mid \{i\}) \leq p(v \mid I)$. Par conséquent, $\max\{p(v \mid \{i\}), i \in I\} \leq p(v \mid I)$. Ici, max n'est rien d'autre que le OR_{flou} standard.

Définissons maintenant la fonction :

$$g(I) = \sum_{v \in V} OR_{flou}\{p(v \mid \{i\}), i \in I\} \quad (1)$$

Propriété 2. Pour tout ensemble $I \subseteq V$, $f(I) \geq g(I)$, et cette borne est exacte.

Preuve. Immédiate par la propriété 1.

Propriété 3. La fonction $g(I)$ est sous modulaire et monotone croissante *indépendamment* de la méthode de diffusion ayant généré la matrice $\boldsymbol{P}$.

Preuve. Vérifions les propriétés (i), (ii), (iii) des fonctions sous modulaires pour $g(I)$. Il est clair que $g(\emptyset) = 0$ et si $A \subseteq B \subseteq V$, alors $p(v \mid A) \leq p(v \mid B)$ pour tout $v \in V$. Par conséquent, $OR_{flou}\{p(v \mid \{a\}), a \in A\} \leq OR_{flou}\{p(v \mid \{b\}), b \in B\}$. Cette monotonie élément par élément, donne immédiatement $g(A) \leq g(B)$. Il reste à prouver que g est sous modulaire. Rappelons qu'une définition équivalente de la propriété de sous modularité (iii) est :

(iv) $A \subseteq B \subseteq V, u \in V \backslash B, F(A \cup u) - F(A) \geq F(B \cup u) - F(B)$;

Nous démontrons une propriété plus forte pour g : pour tout $v \in V$ et $A \subseteq B \subseteq V, u \in V \backslash B$,

$$OR_{flou}\{p(v \mid A \cup u)\} - OR_{flou}\{p(v \mid A)\} \geq$$
$$OR_{flou}\{p(v \mid B \cup u)\} - OR_{flou}\{p(v \mid B)\} \qquad (2)$$

En d'autres termes, chaque terme de $g(I)$ est sous modulaire.

Soit $m_A = OR_{flou}\{p(v \mid A)\}$ et $m_B = OR_{flou}\{p(v \mid B)\}$. Par la propriété de monotonie, on a $m_A \leq m_B$. L'inégalité (2) peut se démonter facilement en considérant trois cas :

Cas1. $p(v \mid \{u\}) \leq m_A \leq m_B$.

$OR_{flou}\{p(v \mid A \cup u)\} = m_A$ et $OR_{flou}\{p(v \mid B \cup u)\} = m_B$.

Cas 2. $m_A \leq p(v \mid \{u\}) \leq m_B$.

$OR_{flou}\{p(v \mid A \cup u)\} \geq m_A$ et $OR_{flou}\{p(v \mid B \cup u)\} = m_B$.

Cas 3. $m_A \leq m_B \leq p(v \mid \{u\})$. $p(v \mid \{u\}) - m_A \geq p(v \mid \{u\}) - m_B$

Par conséquent, g est sous modulaire comme somme de fonctions sous modulaires, ce qui prouve la propriété 3.

L'idée principale de notre approche est de chercher un ensemble initial I^g qui maximise $g(I)$. Ce problème d'optimisation combinatoire est NP-Difficile puisque il se réduit au problème classique SET COVER dans le cas particulier où $p(v \mid I)$ est un vecteur binaire $\{0, 1\}$. La monotonie et la sous modularité de $g(I)$ garantissent que l'ensemble initial $I^g_{glouton}$ trouvé par l'algorithme glouton ci-dessous donne une valeur $g\left(I^g_{glouton}\right)$ telle que :

$$0.63 g(I^g) \leq g\left(I^g_{glouton}\right) \leq g(I^g) \leq f(I^g) \leq f(I^f)$$

Il est facile d'adapter l'algorithme glouton pour le SET COVER classique dans le cas des opérateurs flous pour obtenir un algorithme de COUVERTURE PAR ENSEMBLE FLOU. Une définition standard de la cardinalité d'un ensemble flou **p** est $CARD_{flou}(\mathbf{p}) = \sum_{v \in V} \mathbf{p}(v)$.

Nous nous intéressons à la complexité de calcul de $I^g_{glouton}$ par l'algorithme ci-dessous. La ligne 3 de cet algorithme nécessite au plus n comparaisons (autant que de colonnes dans $\boldsymbol{P}$) pour trouver u^*. Chacune des opérations AND_{flou}, OR_{flou}, NOT_{flou} et $CARD_{flou}$ nécessite n opérations. Ainsi, la matrice $\boldsymbol{P}$ étant donnée, la recherche de $I^g_{glouton}$ nécessite au plus $|I|n^2$ opérations. Ce résultat est à comparer avec $(H - n)N_{LT}$ – le nombre d'opérations pour chercher le $2^{ième}$, ..., $|I|^{ième}$ sommet de $I^f_{glouton}$. Rappelons que $H - n = \sum_{k=1}^{|I|-1}(n - k) \sim O(|I|\, n)$ et $N_{LT} \sim O(R\, m\, n)$. Par conséquent, le gain en complexité est $O(R\, m)$.

GloutonCouvertureEnsembleFlou$(\boldsymbol{P}, k)$
Entrée: $\boldsymbol{P}$ – Colonnes $\boldsymbol{P}_u = p(v \mid \{u\}$; k – cardinalité d'ensemble initial;
Retourne: I – ensemble viral initial

1. $I \leftarrow \emptyset$; $\mathbf{v} \leftarrow \mathbf{1}_{n\times 1}$; $\mathbf{c} \leftarrow \mathbf{0}_{n\times 1}$
2. **tantque** $|I| < k$
3. **trouver** $u^* \leftarrow argmax\left(CARD_{flou}\ \left(AND_{flou}(\mathbf{v}, \boldsymbol{P}_u)\right)\right), u \in V \backslash I$;
4. $I \leftarrow I \cup u^*$;
5. $\mathbf{c} \leftarrow OR_{flou}(\mathbf{c}, \boldsymbol{P}_{u*})$
6. $\mathbf{v} \leftarrow AND_{flou}\left(\mathbf{v}, NOT_{flou}(\mathbf{c})\right)$
7. **fintantque**

Figure 3. Algorithme glouton pour COUVERTURE PAR ENSEMBLE FLOU.

4. Expérimentations

Afin de valider notre approche, nous avons construit un ensemble de graphes planaires et aussi nous avons extrait et étudié un graphe d'influence basé sur les publications de co-auteurs du site PubMed[1].

4.1. Graphes jouets planaires

Pour illustrer la méthode nous avons créé un générateur de graphe planaire sur une grille de $a \times b$ sommets. Les sommets de ces graphes sont les couples $(i, j), i = 1, \ldots, a, j = 1, \ldots, b$. Chaque sommet peut être 4-connecté (Nord, Sud, Est Ouest) ou 8-connecté avec ses voisins de la grille. La densité des arcs est contrôlée par un pourcentage de densité que nous avons fixé à $De = 0.44$. Ces graphes ont été enregistrés avec les noms génériques $G\_a \times b\_C\_4\_De\_44$ afin d'indiquer leur taille, connectivité et densité. Les influences $w(u, v)$ pour tout arc (u, v) sont générées aléatoirement avec la contrainte que la somme des influences entrant dans un sommet soit inférieure ou égale à 1.

4.2. Graphe de co-auteurs

Nous avons extrait les publications avec co-auteur de 82999 auteurs et plus de 43000 articles sur le cancer. Ces articles sont dans le site de PubMed (http://www.ncbi.nlm.nih.gov/pubmed). Les données contiennent pour chaque

[1] Pour obtenir le code et les données demander à stefan.janaqi@mines-ales.fr

auteur A la liste $L(A)$ (les indices) de ses publications. Après, nous avons calculé l'influence d'un auteur A sur un auteur B comme suit :

$$w(A,B) = k(A)\frac{|L(A)\cap L(B)|}{|L(B)|} \qquad (3)$$

Ici, $k(A)$ est un indicateur de la productivité d'un auteur A. Il est proportionnel au nombre de publications de A. Ainsi, si $L(A) \cap L(B) \neq \emptyset$ les deux arcs (A,B) et (B,A) existent dans le graphe mais les influences $w(A,B)$ et $w(B,A)$ peuvent être différentes. Finalement, toutes les influences $w(A,B)$ ont été normalisées pour que la somme des influences entrantes dans un sommet soit inférieure ou égale à 1. Ce graphe sera appelé $AUTH\_82999$.

Nous avons expérimenté notre approche avec LT et avec des cardinalités de l'ensemble initial I allant de $|I| = 1$ à $|I| = 30$. Afin de comparer nos résultats avec les approches classiques, nous avons calculé $f(I)$ pour I choisi par chacune des cinq méthodes suivantes :

(i) Random – I est un ensemble aléatoire de k sommets, $k = 1, \ldots 30$. Le générateur aléatoire a été lancé 10000 fois et le meilleur résultat est retenu.

(ii) MaxDeg – Les $k = 1, \ldots 30$ sommets de I ont été choisis parmi ceux de degré sortant maximum.

(iii) MinDist – Pour chaque sommet u, ses distances vers tous les autres sommets sont calculées. La moyenne $m(u)$ de ces distances est un indicateur d'influence du sommet u. On peut raisonnablement argumenter que l'influence de u à v décroit avec la distance de u à v. L'ensemble initial est alors choisi selon l'ordre croissant des moyennes $m(u)$;

(iv) L'ensemble initial I est choisi par l'algorithme glouton pour $f(I)$;

(v) L'ensemble initial I est choisi par l'algorithme glouton pour la borne inférieure $g(I)$.

Tableau 1. Espérance $f(I)$ pour $|I| = 30$ avec le modèle de diffusion LT.

	Random	MaxDeg	MinDist	Glouton $f(I)$	Glouton $g(I)$
G_10x10_C_4_D_44	33.29	45.25	38.04	61.48	58.53
G_10x10_C_8_D_44	37.65	46.04	36.63	59.55	58.97
G_20x20_C_4_D_44	49.38	56.14	42.60	88.99	88.88
G_20x20_C_8_D_44	61.23	85.47	44.98	114.40	109.69
AUTH_82999	492.59	1234.79	1223.60	1256.29	1254.80

La proximité des résultats des deux dernières colonnes du Tableau 1 est remarquable. L'erreur relative $\left|f\left(I^f_{glouton}\right) - f\left(I^g_{glouton}\right)\right| / f\left(I^f_{glouton}\right) \leq 5\%$. Sur

AUTH_82999 MaxDeg et MinDist donnent des résultats satisfaisants aussi alors que pour les graphes jouets ces deux méthodes donnent des résultats très éloignés de $f(I^f_{glouton})$ et $f(I^g_{glouton})$.

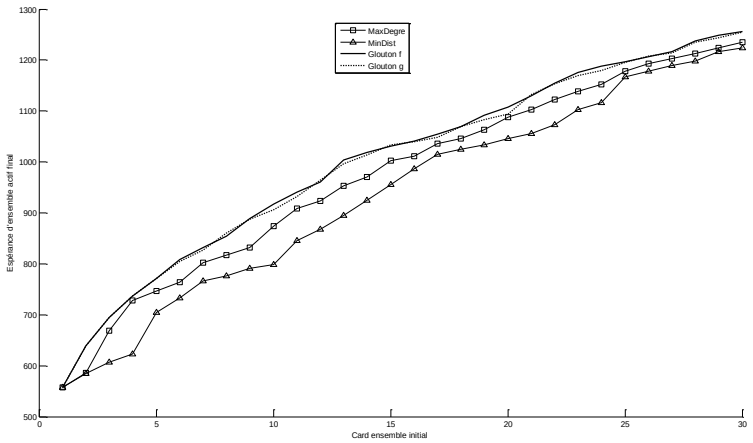

Figure 4. Valeurs de $f(I)$ pour $|I| = 1, \dots, 30$ pour le graphe AUTH_82999. La ligne en pointillé représente la borne $g(I)$ suivant de très près $f(I)$.

La figure 4 présente l'évolution des espérances pour les méthodes de choix d'ensemble initial mentionnées pour le graphe $AUTH\_82999$. La ligne continue au-dessus des autres représente les valeurs de $f(I^f_{glouton})$. La ligne en pointillé juste en dessous représente $f(I^g_{glouton})$. Nous nous attendions à un décrochage de ces deux courbes lorsque $|I|$ augmente. Ce n'est pas le cas et une étude plus poussée est nécessaire pour voir l'effet de la structure du graphe dans les résultats.

5. Conclusion et perspectives

Nous avons présenté une méthode de recherche de l'ensemble viral initial basée sur une borne inférieure exacte de la fonction mesurant l'influence. La complexité de calcul d'un ensemble initial par notre méthode est de complexité bien inférieure que la complexité de l'algorithme glouton appliqué directement sur la fonction d'influence. Notre borne inférieure est monotone et sous modulaire indépendamment de la méthode de diffusion choisie. Ceci, donne une garantie de bons résultats de l'algorithme glouton sur cette borne.

Les résultats sur des graphes jouets et sur un graphe réel sont encourageants et semblent confirmer notre intuition sur la bonne qualité de notre approximation.

Néanmoins, il est nécessaire d'élargir ce travail en cours en menant des expérimentations sur d'autres graphes (jouets et réels) et avec d'autres méthodes de

diffusion. Ceci afin de mesurer la dépendance des résultats de la structure du graphe ou des particularités de la méthode de diffusion.

Nous envisageons la recherche d'autres bornes intéressantes de faible complexité et leurs tests sur des bancs d'essai de réseaux d'influence.

Bibliographie

[ARA 12] Aral, S., D. Walker, Identifying Influential and Susceptible Members of Social Networks. Science, 337 (6092) 337-341, 2012

[BAS 69] F. Bass. A new product growth model for consumer durables. Management Science 15(1969), 215-227.

[BRU 08] Bruyn, A. D., G. L. Lilien, A Multi-Stage Model of Word-of-Mouth Influence through Viral Marketing. International Journal of Research in Marketing 25(3) 151-163, 2008.

[CAR 08] Carr, D. How Obama Tapped into Social Networks' Power. The New York Times, November 9, 2008.

[CHE 11] Chen, W., C. Wang, Y. Wang, Scalable Influence Maximization for Prevalent Viral Marketing in Large-Scale Social Networks. Proceedings of the 16th ACM SIGKDD International Conf on Knowledge Discovery and Data Mining, 1029-1038, 2010.

[DOM 01] P. Domingos, M. Richardson. Mining the Network Value of Customers. Seventh International Conference on Knowledge Discovery and Data Mining, 2001.

[DUPR 80] Dubois D, Prade H., Fuzzy Sets and Systems: Theory and Applications. Academic Pres, New York, 1980.

[EAG 09] Eagle, N., A. Pentland, D. Lazer. 2009. Inferring Social Network Structure using Mobile Phone Data. *Proc. of the National Academy of Sciences* **106**(36) 15274-15278.

[GOL 01] J. Goldenberg, B. Libai, E. Muller. Using Complex Systems Analysis to Advance Marketing Theory Development. Academy of Marketing Science Review 2001.

[GRA 78] Granovetter M., Threshold Models of Collective Behavior, American Journal of Sociology, vol. 83, no 6, 1978, p. 1420-1443, The University of Chicago Press.

[KEM 03] Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence in a social network. In: Proc. 9th Intl. Conf. on Knowledge Discovery and Data Mining. (2003), 137–146

[KIM 07] Kimura M., Saito K., Nakano R., « Extracting influential nodes for information diffusion on a social network », Proceedings Of The National Conference On Artificial Intelligence, vol. 22, no 2, 2007

[LAG 13a] Lagnier C., Gaussier E., Etude de la maximization de l'influence dans les réseaux sociaux, MARAMI 2013, 4ième conférence sur les modèles e l'analyse des réseaux : Approches Mathématiques et informatiques, oct 2013, St Etienne, France.

[LAG 13b] Lagnier C., Denoyer L., Gaussier E., Gallinari P., Predicting Information Diffusion in Social Networks using Content and User's Profiles, IN ECIR, 2013.

[LES 09] Leskovec J., Backstrom L., Kleinberg J., « Meme-tracking and the dynamics of the news cycle », Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data mining, KDD '09, ACM, 2009, p. 497-506.

[NEM 78] Nemhauser G. L., Wolsey L. A., Fisher M. L., « An analysis of approximations for maximizing submodular set functions-I », Mathematical Programming, vol. 14, no 1, 1978, p. 265-294, Springer Berlin / Heidelberg.

[RIC 02] Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing. In: Proc. 8th Intl. Conf. on Knowledge Discovery and Data Mining. (2002) 61–70.

[SAI 08] Saito K., Nakano R., Kimura M., « Prediction of Information Diffusion Probabilities for Independent Cascade Model », Proceedings of the 12th international conference on Knowledge-Based Intelligent Information and Engineering Systems, Part III, KES '08, Springer-Verlag, 2008, p. 67-75.

[SCH 78] T. Schelling. Micromotives and Macrobehavior. Norton, 1978.

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

[XIA 13] Xiao F., Hu P., Li Z., Tsai W., Predicting Adoption Probabilities in Social Networks, University of Utah, arxiv.org 2013.

[YAG 82] Yager R. R., Fuzzy Sets and Possibility Theory, Pergamon Press, Oxford 1982.

[ZAD 65] Zadeh L. A., Fuzzy Sets, Information and Control, 8, pp 338-353.