Propagation d'événements dans un graphe économique

Jocelyn Bernard, Julien Goncalves, Hamamache Kheddouci

Propagation d'événements dans un graphe économique

Jocelyn Bernard[*,**] Julien Goncalves[*]
Hamamache Kheddouci[**]

[*]ReportLinker, 21 Quai Antoine Riboud, 69002, Lyon, France
prenom.nom@reportlinker.com, https://www.reportlinker.com/
[**]Université Lyon 1, LIRIS CNRS 5205, 69100, Villeurbanne, France
prenom.nom@univ-lyon1.fr

Résumé. Les modèles de diffusion dans les réseaux sociaux sont beaucoup étudiés ces dernières années. Les études concernent notamment les diffusions de maladies et de rumeurs dans les réseaux sociaux ou de risques financiers dans les réseaux bancaires. Nous proposons dans cet article de répondre au problème de diffusion des événements au sein de réseaux économico-sociaux. En particulier, nous proposons d'étudier un nouveau problème de diffusion appelé *Influence Classification Problem (ICP)* dont l'objectif est de classifier automatiquement quels noeuds sont impactés pour un événement donné. Nous proposons également deux modèles de propagation basés sur un seuil calculé en fonction des attributs du graphe et de l'événement. Nous testons nos modèles sur deux événements connus : l'ouragan Katrina et l'acquisition de Monsanto par Bayer.

1 Introduction

Les modèles de propagation actuels proposent de représenter la diffusion d'informations dans un réseau de données. Les problèmes de diffusions de rumeurs ou de maladies sont étudiés pour comprendre et modéliser la propagation d'informations (Kempe et al. (2003); Kermack et McKendrick (1932)). Ces modèles de diffusion ont également été adaptés aux réseaux économiques et notamment bancaires pour étudier les faillites.

Les événements économiques et sociaux semblent donc pouvoir être représentés de la même façon : une baisse de production de l'orge peut augmenter le prix des céréales, ce qui peut impacter les producteurs et distributeurs de bières. Notre idée est de proposer un modèle de propagation représentant ce genre de scénarios. Pour cela nous proposons un problème appelé *Infection Classification Problem (ICP)*. Le but d'*ICP* est de déterminer pour chaque élément du réseau s'il est concerné par l'événement. Dans cette optique nous avons étudié deux événements connus : L'ouragan Katrina qui a frappé la Nouvelle-Orléans et la tentative d'acquisition de Monsanto par Bayer. Pour répondre à *ICP* nous proposons deux modèles de diffusions appelés *Hybrid Linear Threshold (HLT)* et *Adapted Threshold (AT)* qui sont basés sur le modèle de seuil *Linear Threshold*. Pour chacun des modèles nous utilisons un seuil pour déterminer à partir de quel instant un noeud est impacté par un événement. Les modèles diffèrent par la manière dont le seuil est calculé, pour *HLT*, nous utilisons la valeur des poids des arcs entrants tandis que nous utilisons la valeur du poids du noeud pour *AT*.

2 État de l'art

Un réseau est représenté par un graphe dirigé $G = (V, E)$ où les noeuds V sont des entités et les arcs E représentent les interactions entre les entités. Dans le cas de réseaux sociaux, les noeuds représentent des personnes et les arcs les interactions sociales entre les personnes.

Parmi les problèmes étudiés d'infection de réseaux, Kempe et al. (2003) proposent d'étudier l'*Influence Maximization Problem*, où le but est de déterminer quels noeuds initiaux impactent le plus d'autres noeuds. Pour cela ils proposent deux modèles, *Independent Cascade Model (IC)* et *Linear Threshold Model (LT)*. Ces modèles, à partir de noeuds initiaux, propagent l'information dans le graphe. Ces noeuds initiaux, ou noeuds sources, transmettent l'information de la contagion à leurs voisins qui peuvent alors se retrouver infectés et la transmettre à leur tour.

Independent Cascade Model (IC) : Goldenberg et al. (2001) proposent un modèle où, à chaque étape t, un noeud impacté u peut aléatoirement impacté un ou plusieurs voisins relié par l'arc (u, v) via une fonction probabiliste $p_{(u,v)}$. Si l'information est transmise, le noeud v devient impacté et peut alors transmettre l'information au temps $t + 1$.

Linear Threshold Model (LT) : Granovetter (1978) proposent un modèle basé sur un seuil. Chaque noeud v a un seuil $s(v)$ et il y a un poids w sur chacun de ses arcs entrants. À une étape t, si la somme du poids des arcs venant des voisins impactés est supérieure au seuil $s(v)$, v devient impacté :

$$\exists t, v \in S_t,\ \forall u \in I_t, (u, v) \in E, \quad \sum w(u, v) > s(v) \Rightarrow v \in I_{t_{+1}} \tag{1}$$

Dans notre papier, nous utiliserons le modèle *LT* car il est déterministe pour un seuil fixe (Lu et al. (2011)), ce qui nous permet de garantir un résultat toujours identique pour les expérimentations, contrairement au modèle *IC*.

D'autres problèmes proposent d'étudier la minimisation de l'influence ou l'influence compétitive (Yang et al. (2017); Caliò et Tagarelli (2018)). Mehmood et al. (2016) proposent le *Typical Cascade Problem* où le but est de trouver quels noeuds sont généralement impactés pour des noeuds sources donnés.

Les modèles de diffusions ont également permis de représenter les risques de faillites dans les milieux bancaires (Chinazzi et Fagiolo (2015); Kenett et Havlin (2015)), la simulation des risques du aux échanges entre banques (Montagna et Kok (2016)) ou encore la formation de réseaux économiques (Kantemirova et al. (2018)).

3 Problème de classification de noeuds impactés

Nous voulons déterminer, pour un graphe et un événement donné, quels noeuds sont impactés. Nous définissons pour cela un nouveau problème appelé *Infection Classification Problem (ICP)*. Soit :

- $L_{IV} \subset V$ la liste des noeuds initialement impacté par l'infection i
- $L_{TV} \subset V$ une liste de noeuds cibles dans laquelle :
  - $L_{RV} \subset L_{TV}$ sont les noeuds devant être impactés à la fin de la propagation
  - $L_{AV} \subset L_{TV}$ sont les noeuds ne devant pas être impactés à la fin de la propagation

— $L_{RV} \cap L_{AV} = \emptyset$

— Lorsque la propagation s'arrête pour un modèle de diffusion nous avons :
  — L_{TP} la liste des noeuds impactés par le modèle et qui devaient l'être
  — L_{FP} la liste des noeuds impactés par le modèle et qui ne devaient pas l'être
  — L_{TN} la liste des noeuds non impactés par le modèle et qui ne devaient pas l'être
  — L_{FN} la liste des noeuds non impactés par le modèle et qui devaient l'être

Avec ces listes nous pouvons calculer la précision et le rappel :

$$Précision = \frac{|L_{TP}|}{|L_{TP}| \cup |L_{FP}|} \qquad Rappel = \frac{|L_{TP}|}{|L_{TP}| \cup |L_{FN}|} \tag{2}$$

La précision offre une évaluation du bruit en calculant le pourcentage de noeuds bien classé parmi les noeuds impactés. Le rappel offre une évaluation du silence en calculant le pourcentage de noeuds impactés parmi ceux qui sont censés l'être. La moyenne harmonique de ces valeurs, appelée *F-measure*, permet d'évaluer *ICP*. Le but des modèles appliquer à *ICP* étant de maximiser cette valeur.

$$F - measure = 2 \cdot \frac{précision \cdot rappel}{précision + rappel} \tag{3}$$

4 Modèles de diffusion

Nous proposons de résoudre *ICP* à l'aide de 2 modèles dont le développement a été motivé par l'adaptation d'un graphe à un événement typé. En effet, l'*acquisition d'une société* n'impactera pas les noeuds de la même manière que la sortie d'un *nouveau produit*, même si les sources sont identiques. Pour cela nous définissons une infection $i = (L_{IV}, type)$ tel que :

— $L_{IV} \subset V$ la liste des sources : $\forall v \in L_{IV}, v \in I_{t_0}$

— $type$, le type de l'infection

Nous utilisons les données présentes dans le graphe (voir Section 5) tel que le poids, le *type* des liens ou des noeuds ou encore la distance aux sources pour adapter la propagation.

Calcul du seuil : Pour déterminer pour chacun des noeuds leur seuil respectif nous utilisons leur *type*, le *type* de l'infection et soit le poids du noeud soit celui des arcs entrant.

Poids des arcs : Le poids des arcs est adapté selon le type des noeuds de l'arc, du lien et de l'événement. Selon ces types un pourcentage est appliqué au poids initial de l'arc.

Valeur d'infection : Nous utilisons le rapport entre la somme des infections provenant des voisins et le seuil propre au noeud pour donner une valeur à l'infection, transmise aux voisins. L'idée repose sur la virulence pour représenter un impact plus ou moins fort.

Fonction d'utilité : Nous utilisons la distance entre le noeud infecté et la source de l'infection pour diminuer la valeur d'infection : plus un noeud est distant des sources moins l'information transmise sera virulente.

Nous avons créé deux modèles de propagation qui utilisent la notion de seuil. Comme pour le modèle *LT* les noeuds ont un seuil et deviennent impactés lorsque l'information transmise par leurs voisins devient supérieure à ce seuil. La différence avec le modèle *LT* est que nous adaptons l'infection du graphe à un événement donné à l'aide des fonctionnalités. Ces deux modèles se distinguent par la façon dont est calculé le seuil :

Hybrid Linear Threshold (HLT) : Nous utilisons le *type* de l'événement et du noeud pour déterminer un pourcentage qui est appliqué sur le total du poids des arcs entrants pour déterminer la valeur du seuil.

Adapted Threshold (AT) : Le *type* de l'événement et du noeud sont utilisés pour déterminer un pourcentage qui est appliqué sur le poids du noeud (qui est calculé indépendemment des arcs).

5 Expérimentations

Pour tester notre modèle, nous utilisons un multi-graphe composé de 3,8 millions de noeuds et de 9,6 millions d'arcs. Ce graphe représente une situation économico-sociale en 2016. Nos noeuds sont typés *(organisations, lieux, personnes, etc.)*.

Le graphe est construit par agrégations d'hypothèses détectées à l'aide d'un traitement automatique du langage sur différents documents (news, rapports économiques). Basiquement, si deux entités nommées (= noeuds) sont présents dans une même phrase, ils se retrouvent liés par un arc dans le graphe. Plus une information est présente, plus son poids est important dans le graphe. Un label est utilisé pour désigner le type de chaque arc afin de représenter l'interaction entre deux noeuds. Les arcs sont dirigés et peuvent être multiples (par exemple pour représenté des situations de coopérations et de compétitions entre deux mêmes noeuds).

Pour tester nos modèles, nous avons défini avec des analystes économiques deux événements : *Katrina* et *Bayer-Monsanto*.

Katrina : Nous avons simulé l'impact de l'ouragan *Katrina* qui a frappé la *Nouvelle-Orléans* en 2005. Nous avons défini l'événement comme une infection avec *Nouvelle-Orléans* et *Louisiane* comme sources L_{IV} et *catastrophe naturelle* comme type d'événement.

Bayer-Monsanto : Nous avons simulé l'impact de l'acquisition de *Monsanto* par *Bayer*. Nous avons défini l'événement comme une infection avec *Bayer* et *Monsanto* comme sources L_{IV} et *acquisition compagnie* comme type d'événement.

Pour évaluer la qualité des modèles, nous avons défini avec les analystes économiques une liste de 3792 noeuds considérés comme importants : ce sont nos noeuds cibles dans *ICP*. Ces noeuds ont été choisis à partir du poids des noeuds et des arêtes entrantes. Pour chaque événement les analystes ont défini quels noeuds cibles étaient concernés, 55 pour *Katrina* et 92 pour *Bayer-Monsanto*.

Nous avons testé quatre modèles pour résoudre le problème *ICP* :

Linear Threshold Model (LT) : Basé sur l'article de Granovetter (1978). Chaque noeud à son propre seuil fixé par un pourcentage s'appliquant sur la somme du poids des arcs entrants.

Dynamic Linear Threshold Model (DLT) : Basé sur l'article de Litou et al. (2016), *DLT* calcule un seuil par un pourcentage appliqué sur les poids des arcs entrants. Le poids des arcs change dans le temps en exploitant une distribution suivant une Loi de Poisson. Les noeuds peuvent également renoncer à l'infection et modifier leurs seuils en fonction du temps.

Hybrid Linear Threshold Model (HLT) : Les poids des arcs et l'information transmise sont calculés en fonction du *type* de l'événement, des noeuds et des arcs. Le seuil est calculé selon un pourcentage, défini par le *type* de l'événement et du noeud, s'appliquant sur le poids des arcs entrants.

Adapted Threshold Model (AT) : Le seuil est calculé selon un pourcentage, défini par le *type* de l'événement et du noeud, s'appliquant sur le poids initial du noeud.

Les résultats des données sont présentés en Table 1, ceux de *ICP* dans la Table 2.

Alg.	$\|I\|$	TP	FP	FN
LT	750	4	1	51
DLT	518	4	0	51
HLT	405	7	5	48
AT	89	18	21	37

Alg.	$\|I\|$	TP	FP	FN
LT	1 691	16	9	76
DLT	1 244	14	4	78
HLT	1 693	32	27	60
AT	328	50	22	42

TAB. 1 – *Résultats pour Katrina (gauche) et Bayer-Monsanto (droite).*

La Table 1 présente les résultats des différents modèles pour les événements *Katrina* et *Bayer-Monsanto*. La colonne $|I|$ donne le nombre de noeuds impactés. Les colonnes TP, FP et FN représentent le nombre de noeuds correctement classifiés parmi les noeuds cibles.

Nous notons que *DLT*, avec le changement du poids des arcs, impacte moins de noeuds que *LT* tandis que *AT* et *HLT* impactent une plus grande fraction de noeuds cibles.

Alg.	Précision	Rappel	F-Measure
LT	0.8	0.07	0.13
DLT	1.0	0.07	0.14
HLT	0.58	0.13	0.21
AT	0.47	0.33	0.38

Alg.	Précision	Rappel	F-Measure
LT	0.64	0.17	0.27
DLT	0.78	0.15	0.25
HLT	0.54	0.35	0.42
AT	0.69	0.54	0.61

TAB. 2 – *Résultats de classification pour Katrina (gauche) et Bayer-Monsanto (droite).*

La Table 2 présente les résultats pour le problème *ICP*. Les colonnes Précision et Rappel donnent respectivement une évaluation du bruit et du silence (voir Section 3). La colonne F-Measure donne l'évaluation de chaque algorithme pour le problème *ICP*.

Pour *ICP*, les modèles *LT* et *DLT* donnent tous deux de bons résultats pour la précision car ils impactent un petit nombre de noeud cibles voisins des noeuds sources L_{IV}, qui se retrouvent généralement être dans les cibles désirées L_{TP}. En conséquence le bruit est très bas. Cependant, ils n'impactent pas suffisamment de noeuds cibles et se retrouvent avec un silence important. *AT*, qui produit de meilleurs impacts sur les noeuds cibles, donne de meilleurs résultats que les autres modèles. Les fonctionnalités adaptés aux attributs du graphe permettent également à *HLT* de surpasser le standard *LT* duquel il est tiré.

6 Conclusion et ouvertures

Dans ce papier nous proposons un nouveau problème de diffusion, *ICP*, dont le but est de classifier, pour un graphe et un événement donnés, les noeuds impactés par cet événement. Nous proposons deux modèles de diffusions inspirés des modèles de seuil. Le premier modèle, *HLT*, calcule le seuil à partir du poids des arcs entrants tandis que le second, *AT*, calcule le seuil à partir du poids du noeud. Nous utilisons ces modèles pour simuler des événements économiques sur un graphe typé et pondéré. Nos premières expérimentations montrent que nos modèles surpassent ceux de la littérature.

Plusieurs pistes restent à explorer pour améliorer nos résultats. Étant donné que *ICP* est un problème de classification, avec plus d'événements, nous pourrions utiliser des techniques

d'apprentissage pour optimiser les valeurs de pondérations. Nous pourrions également évaluer l'impact d'un événement, regarder du côté de l'analyse des sentiments ou des cas d'événements compétitifs pour proposer des interactions proches de la réalité.

Remerciement Nous tenons à remercier Benjamin Carpano et Marine Gonzalez pour leur expertise dans la définition des événements.

Références

Caliò, A. et A. Tagarelli (2018). Trust-based dynamic linear threshold models for non-competitive and competitive influence propagation.

Chinazzi, M. et G. Fagiolo (2015). Systemic risk, contagion, and financial networks : A survey.

Goldenberg, J., B. Libai, et E. Muller (2001). Talk of the network : A complex systems look at the underlying process of word-of-mouth.

Granovetter, M. (1978). Threshold models of collective behavior.

Kantemirova, M., Z. Dzakoev, Z. Alikova, S. Chedgemov, et Z. Soskieva (2018). Percolation approach to simulation of a sustainable network economy structure.

Kempe, D., J. Kleinberg, et É. Tardos (2003). Maximizing the spread of influence through a social network.

Kenett, D. Y. et S. Havlin (2015). Network science : a useful tool in economics and finance.

Kermack, W. O. et A. G. McKendrick (1932). Contributions to the mathematical theory of epidemics. ii.—the problem of endemicity.

Litou, I., V. Kalogeraki, I. Katakis, et D. Gunopulos (2016). Real-time and cost-effective limitation of misinformation propagation.

Lu, Z., W. Zhang, W. Wu, B. Fu, et D. Du (2011). Approximation and inapproximation for the influence maximization problem in social networks under deterministic threshold model.

Mehmood, Y., F. Bonchi, et D. García-Soriano (2016). Spheres of influence for more effective viral marketing.

Montagna, M. et C. Kok (2016). Multi-layered interbank model for assessing systemic risk.

Yang, L., A. Giua, et Z. Li (2017). Minimizing the influence propagation in social networks for linear threshold models.

Summary

The diffusion models of infections in social networks are intensively studied these last years. The existing studies concern in particular disease and rumor diffusions in social networks or financial risk in banking networks. We propose in this paper to study the diffusion problem of events within social and economic networks. In particular, we define a new problem of diffusion called the *Influence Classification Problem*. The objective is to find the set of nodes which are impacted by a given network. We also propose two diffusion models based on a computed threshold according to the graph and event attributes. We test our models on two real and known events : the hurricane Katrina and the fusion of Bayer and Monsanto.