# Phénomènes de diffusion dans les réseaux dynamiques : simulation et modélisation

Alice Albano*

* LIP6
alice.albano@lip6.fr

**Résumé.** Les phénomènes de diffusion sont présents dans de nombreux contextes : diffusion d'épidémies, de virus informatiques, d'information dans des réseaux sociaux, etc. Bien que les réseaux où se produit la diffusion soient souvent dynamiques, cette dynamique n'est pas prise en compte dans la plupart des modèles existants. L'objectif de ces travaux est de proposer des modèles de diffusion, et d'étudier l'impact de la dynamique du réseau sur la diffusion.

## 1 Introduction et Contexte

Les réseaux d'interaction issus du monde réel, ou graphes de terrain, sont utilisés dans de nombreux contextes pour modéliser des objets divers : topologie de l'Internet, réseaux sociaux, réseaux pair-à-pair, etc. La plupart des graphes de terrain sont dynamiques, c'est-à-dire que le nombre de nœuds et de liens évolue au fil du temps. Les travaux sur les graphes de terrain dynamiques, comme Magnien (2010), sont toutefois assez récents.
Les études sur la diffusion sont par contre assez nombreuses, car les phénomènes de diffusion sont abordés dans plusieurs disciplines : en informatique (virus informatique, diffusion d'information dans un réseau social) Girvan et al. (2001), en biologie (épidémie) Cauchemez et al. (2010), en physique, etc.
Toutefois, l'étude de la diffusion sur un graphe dynamique est un domaine extrêmement nouveau, et il existe à ce jour peu de travaux regroupant ces deux aspects. Nous pouvons toutefois citer Li et al. (2004), Stehlé et al. (2011) ou encore Yoneki et al. (2008).

## 2 Objectif

Notre objectif ici est d'étudier des phénomènes de diffusion sur des graphes dynamiques. La première étape consiste à proposer un modèle de diffusion reflétant aussi bien que possible les diffusions observées dans les réseaux réels. Ensuite, nous appliquons ce modèle à un jeu de données, et nous tentons grâce aux observations obtenues et à une comparaison avec une diffusion réelle, d'affiner le modèle utilisé. Il s'agit donc d'une approche empirique. Par conséquent, le modèle obtenu sur un jeu de donnée particulier risque de ne pas être pertinent sur d'autres données. Toutefois, étant donné que les graphes de terrain possèdent des propriétés communes, comme par exemple une composante connexe géante, ou une densité faible, on peut espérer que certains de ces modèles peuvent être étendus à différents contextes.

# 3 Modélisation

Le modèle de diffusion le plus simple est le modèle SI (Sain/Infected), expliqué par exemple par Girvan et al. (2001). Dans ce modèle, tous les nœuds du graphe sont soit sains, soit infectés. Un nœud sain est contaminé avec une certaine probabilité s'il a un voisin contaminé. Une fois qu'un nœud est contaminé, il reste dans cet état, et ne peut pas guérir.
Il existe plusieurs variations de ce modèle : nous pouvons par exemple citer le modèle SIR (Sain/Infected/Recovered) expliqué par Kermack et McKendrick (1927). Dans ce modèle, les nœuds peuvent prendre trois états différents : sain, infecté, et immunisé. La contamination se fait sur le même principe que le modèle SI, à la différence qu'un nœud immunisé ne peut pas être infecté. De plus, les nœuds infectés peuvent guérir et passer dans un état immunisé.
De nombreuses variantes de ces modèles peuvent être imaginées : par exemple, en ayant des probabilités de contamination variables selon les noeuds, ou en ayant des noeuds qui, même infectés ne contamineront pas les autres (par exemple pour un virus informatique).
Pour nos travaux, nous avons choisi d'utiliser un jeu de données d'un réseau pair-à-pair provenant de l'enregistrement de l'activité d'un serveur eDonkey pendant deux jours. Le fonctionnement du réseau est le suivant : un pair client effectue une requête pour un fichier. Le serveur lui répond en donnant les adresses IP de fournisseurs potentiels. Le pair télécharge ensuite le fichier auprès de ces fournisseurs. Les données contiennent les réponses du serveur aux clients : on a donc l'instant de la réponse, l'adresse IP du client, le fichier demandé et les adresses IP des fournisseurs. Les requêtes des clients qui n'ont pas de réponse ne sont pas présentes dans le jeu de données. De plus, les données contiennent des informations de session sur les nœuds : on sait quand un nœud se connecte ou se déconnecte, ce qui reflète la dynamique des nœuds dans le réseau. Au cours de la mesure, le serveur a reçu 210 millions de requêtes, et a géré environ 1,5 millions de connexions et déconnexions. Comme la mesure dure deux jours, nous pouvons observer un phénomène jour/nuit sur le nombre de nœuds connectés au serveur : dans la soirée, l'activité baisse et le jour, l'activité augmente.
Sur ce jeu de données, nous supposons que la diffusion d'un fichier a lieu dès lors que le serveur donne une réponse au client, car alors le client va demander le fichier aux fournisseurs qui vont l'envoyer au client. Ces données nous permettent donc tout d'abord d'observer la diffusion réelle d'un fichier, sous l'hypothèse précédente.
De plus, pour simuler une diffusion sur ce jeu de données, nous avons besoin d'un graphe sur lequel simuler la diffusion. Pour cela, nous construisons un graphe dynamique dans lequel deux nœuds sont voisins s'ils ont manifesté, à un instant donné, un intérêt commun pour un fichier (par exemple, si un client fait une requête, et que le serveur lui indique trois fournisseurs, le client et les trois fournisseurs seront tous liés dans le graphe). Nous appelons ce graphe le graphe d'*intérêt*. Les informations de session, qui indiquent les instants précis où un nœud est connecté, nous permettent de rendre compte de la dynamique réelle de ce graphe d'intérêt : nous aurons au fil du temps des nœuds et des liens qui disparaissent et apparaissent. Le graphe d'intérêt comporte environ 470 000 nœuds et 768 millions de liens.

# 4 Résultats

En premier lieu, nous nous intéressons à l'influence de la dynamique du graphe sur la diffusion. Pour étudier ceci, nous avons appliqué le modèle SI sur le graphe d'intérêt introduit

précédemment, en prenant en compte les informations de session dans le premier cas, et sans les prendre en compte dans le deuxième cas. Les résultats de ces simulations sont montrés sur la figure 1 (à gauche).

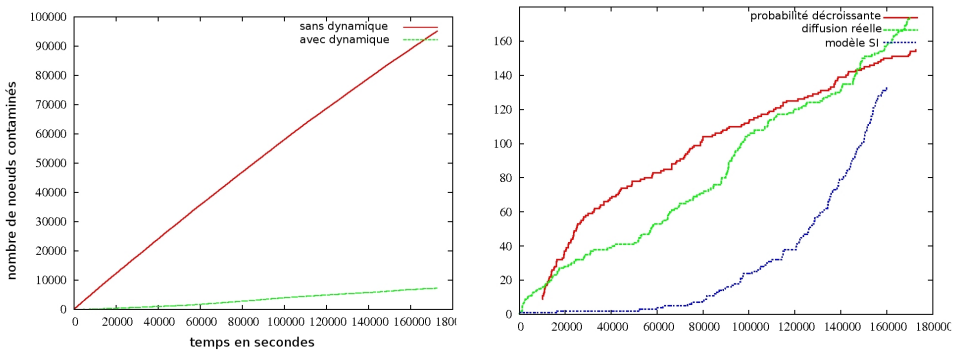


FIG. 1 – *A gauche : Impact de la dynamique du réseau sur la diffusion. A droite, comparaison des simulations de diffusion avec la diffusion réelle.*

Nous observons sur cette figure une différence marquée selon que l'on prend en compte la dynamique du graphe ou pas, la dynamique limitant fortement la diffusion dans un rapport d'environ un pour dix. Nous pouvons en déduire que la dynamique du réseau a un impact important sur la diffusion, et que le graphe que nous considérons dans le cas du jeu de données d'un réseau pair-à-pair a une dynamique importante : de nombreux nœuds et liens apparaissent et disparaissent à chaque instant.
Nous avons ensuite observé le comportement d'une diffusion réelle d'un fichier. Nous considérons qu'une diffusion a lieu entre pairs dès lors que le serveur répond à la requête d'un client avec les adresses des fournisseurs. Nous avons comparé le comportement d'une diffusion réelle d'un fichier avec celui d'une diffusion simulée (figure 1, à droite), sur le graphe d'intérêt dynamique.

La courbe verte (au milieu) représente la diffusion réelle d'un fichier choisi au hasard dans le jeu de données, la courbe bleue (en bas) réprésente le nombre de contaminés en fonction du temps pour un modèle SI avec un paramètre de contamination de $\frac{1}{220000}$ (choisi pour arriver environ au même nombre de contaminés que celui de la diffusion réelle), et la courbe rouge (en haut) représente un modèle avec une probabilité d'être contaminé décroissante au fil du temps. Nous pouvons tout d'abord observer que la diffusion réelle d'un fichier peut se découper en phases. Pendant certains intervalles de temps, la croissance de la courbe est élevée (entre 50 000 et 105 000, et entre 150 000 et 172 000), et à d'autres instants, elle ralentit. Ce phénomène s'explique par le fait que la mesure a été réalisée sur deux jours. La nuit, le nombre de requêtes et de gens connectés diminue fortement.
D'autre part, nous pouvons observer que la simulation avec le modèle SI n'est pas très réaliste. En effet, si nous arrivons au même nombre de nœuds contaminés à la fin de la mesure, le comportement des courbes est très différent. La croissance est étalée dans le temps, mais la diffusion réelle ayant une croissance linéaire alors que SI a une croissance exponentielle, si la simulation durait plus longtemps, la croissance de SI dépasserait très rapidement la diffusion

réelle.
D'après ces constatations, nous avons proposé une variante du modèle SI (courbe du haut) : il s'agit de faire diminuer la probabilité de contamination des nœuds au fil du temps. Ce modèle semble plus vraisemblable par rapport à la diffusion d'un fichier, qui sera beaucoup téléchargé à les premiers jours, puis qui sera de moins en moins demandé. Nous observons des meilleurs résultats avec ce modèle, qui présente une croissance similaire à celle de la diffusion réelle.

# 5 Conclusion et perspectives

Nous avons vu des exemples de simulation de diffusion sur un graphe dynamique, et nous avons effectué une comparaison de ces simulations avec un exemple de diffusion réelle. Une perspective possible est de simuler un phénomène de diffusion sur un réseau aléatoire avec une dynamique très typée pour étudier avec plus de précision l'impact de la dynamique du réseau sur la diffusion.
Par ailleurs, nous appliquerons aussi ces modèles de diffusion présents ici à d'autres jeux de données, afin de voir dans quel cas ces modèles sont valides dans d'autres contextes.

# Références

Cauchemez, S., A. Bhattarai, T. L. Marchbanks, R. P. Fagan, S. Ostroff, N. M. Ferguson, et Swerdlow (2010). Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *PNAS*.

Girvan, M., D. S. Callaway, M. Newman, et S. H. Strogatz (2001). A simple model of epidemics with pathogen mutation. *Physical Review E*.

Kermack, W. O. et A. G. McKendrick (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of London*.

Li, S., M. Meng, et H. Ma (2004). Epidemic spreading in dynamic small world networks.

Magnien, C. (2010). *Intégrer mesure, métrologie et analyse pour l'étude des graphes de terrain dynamiques*. Habilitation à diriger des recherches.

Stehlé, J., N. Voirin, A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Régis, J.-F. cois Pinton, N. Khanafer, W. V. den Broeck, et P. Vanhems (2011). Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees. *BMC Medicine*.

Yoneki, E., P. Hui, et J. Crowcroft (2008). Wireless epidemic spread in dynamic human networks. *LNCS 5151*, 116–132.

# Summary

Spreading phenomena are present in many contexts : virus spreading, computer virus, spreading of information in social networks, etc. In most cases, these phenomena occur in dynamic networks. This dynamic will be very important for the study of spreading. The aim of my work is to offer spreading models, and to study the effect of network dynamics on spreading.