Diffusion dans les réseaux dynamiques

Rapport de Stage
ENS de Lyon

Jérémie Dumas

Maître de Stage : Christophe Crespelle

Juin-Juillet 2010

Laboratoire d'Informatique de Paris 6
4 place Jussieu
75005 Paris

Au cours des 6 semaines de recherche (intensive) que constituent ce stage, on a eu l'occasion d'étudier la diffusion d'information dans des graphes dynamiques, i.e. des graphes dont les arêtes évoluent au cours du temps. En se basant sur des données recueillies lors d'expériences, on mesure certaines propriétés de nos graphes dynamiques, afin d'en tirer des informations structurelles sur celui-ci. Ce papier retrace un peu les expériences étudiées, la démarche abordée, et les mesures effectuées.

Remerciements

Un tel stage n'aurait pas été rendu possible sans la présence bienveillante de Christophe Crespelle. Je tiens également à remercier les autres chercheurs du labo', avec qui on n'a malheureusement pas eu le temps de discuter suffisamment : Mathieu Latapy, Clément Magnien, Fabien Tarissan, et j'en passe.

Merci aussi à Éric Fleury pour être passé dire bonjour à l'occasion du HDR de Clémence, et de nous avoir accordé un peu de temps.

On remercie également les autres stagiaires qui étaient avec moi dans l'aquarium : les deux zouaves Alexandre Isoard et François Gindraud, ainsi que le bien aimable Antoine Mazières.

Table des matières

1 Présentation générale

1.1 Introduction

Le but de ces 6 semaines de stage était d'étudier la diffusion d'information dans un graphe dynamique (oui bon ok c'est déjà marqué dans le résumé). Par graphe dynamique on entends ici un ensemble de sommet fixé, avec des arêtes (liens symétriques) qui apparaissent et disparaissent au cours du temps. On se donne par exemple un ensemble S^0 de sommets détenteurs de l'information à une date $t = t_0$, et l'on voudrait pouvoir dire à partir de quand tel ou tel sommet sera informé, quelle quantité d'information lui sera transmise, etc.

On fini donc par étudier diverses propriété du graphe afin de répondre à ces attentes : nombre de chemin allant d'un sommet u à un sommet v entre deux dates t_0 et t_1, nombre de fois où les deux acteurs ont été en contact, "flot" qui peut passer de l'un à l'autre des sommets, etc.

On a enfin implémenté les divers algo' proposés (ça casse pas trois pattes à un canard non plus), sur quelques données expérimentales que l'on peut ensuite analyser (ou du moins faire semblant).

1.2 Description des expériences

1.2.1 I-Bird

C'est l'expérience sur laquelle on devait se pencher au départ. I-Bird entre dans le cadre du projet européen Mosar (pour *Mastering hOSpital Antimicrobial Resistance in Europe*). L'expérience a eu lieu sur le site de l'Hôpital Maritime de Berck-sur-Mer ; pendant une durée de 6 mois a été mesuré l'évolution des contacts entre quelques 800 personnes, grâce à l'aide de capteur qui leur ont été distribués. En parallèle des prélèvements hebdomadaires ont été fait sur cette population, afin d'évaluer la propagation de diverses souches bactériologiques.

L'idée était donc de pouvoir corréler la structure de graphe dynamique de cette population avec la diffusion des bactéries au sein de ce réseau. Malheureusement les données issues des prélèvements n'étaient pas encore utilisables, aussi a-t-on dû travailler sur d'autres jeux de donnés – sans informations sur des souches bactériologiques – déjà étudiés dans de précédents papiers [SBF+08].

1.2.2 Infocom

Expérience réalisée sur 3 jours à l'occasion de la conférence Infocom 2005 [KH05]. On distribue des capteurs aux gens, capteurs qui émettent régulièrement un signal (toutes les 120s) et écoutent les signaux des capteurs alentours. Quand deux capteurs sont suffisamment proches ils enregistrent un contact, qui s'étale entre deux dates t_1 et t_2.

Les graphes obtenus en traçant les liens entre tous ces gens ont pour caractéristiques d'être creux et de présenter un nombre élevé de triangles. De plus la distribution des

degrés est assez hétérogène et évolue au cours du temps (fort pic pendant les repas, sommets isolés la nuit, etc.).

Dans le cas présent, on a pris un échantillons toutes les 300s, en regroupant les liens sur cet intervalle, et ce sur toutes la durée de l'expérience. On obtient donc une série de quelques 900 graphes sur lesquels on a fait nos mesures.

1.2.3 Reality Mining

Cette expérience a été menée sur une durée de 9 mois sur des étudiants au MIT [EPL09]. A l'aide d'applications installées sur leur téléphone portable, on a pu relever l'évolution des liens entre une centaine de personnes. Les appareils émettant cette fois un signal bluetooth toutes les 300s, on se retrouve avec un nombre beaucoup plus important de contact à étudier.

On notera que les comportements de ces deux séries de graphes sont assez similaires vis-à-vis des propriétés étudiées dans [SBF+08]. Cependant du faire de l'importance du nombre de lien, on a étudié cette série de graphe en prenant un échantillon toutes les heures, en regroupant les liens entre deux heures successives.

1.3 Évolution du projet

Avant de passer à la partie "technique" qui présente les notions retenues dans ce stage, et leur application aux diverses données, attardons-nous un peu sur le cheminement qu'a suivi nos pensées pendant ce stage.

Rappelons donc que l'on désirait étudier le processus de diffusion dans un graphe dynamique (ici un réseau de personnes/sommets fixés dont les contacts évoluent). Pour ce faire il y a plusieurs axes d'études possibles : un aspect analyse, on l'on cherche à définir un "pouvoir de diffusion", à corréler ensuite à une diffusion réelle ; un aspect modélisation, où l'on peut comparer la dynamique d'un modèle à la dynamique d'un graphe réel, etc.

Au début on a un peu cherché à définir des modèles de diffusion des bactéries, que l'on aurait aimer comparer avec les données du projet I-Bird par exemple. Mais il a beaucoup de paramètre à caractériser "empiriquement" pour cela : temps d'incubation d'une bactérie avant qu'elle devienne contagieuse, "virulence" de la bactérie une fois réveillée, temps avant que le système immunitaire ne l'élimine, temps d'exposition nécessaire avant d'être contaminé, etc.

Finalement on a vite oublié la partie modélisation pour s'intéresser à la partie analyse, à savoir définir plusieurs critères qui rendent compte du pouvoir diffuseur de certains acteurs, à certains moments de l'expérience. Ont fini par émerger plusieurs notions simples et parfois complémentaires, portant sur le temps que mettrait une information pour passer d'un sommet u à une date t_1 à un sommet v à une date t_2, le nombre de contact de v avec des sommets porteurs de l'information, etc. Bref pas la peine de détailler plus et passons aux choses sérieuses :

2 Définitions théoriques

2.1 Généralités

2.1.1 Graphe dynamique, transitions

Définition 2.1 (Graphe dynamique)**.** C'est une séquence de graphes $(G_t = (V_t, E_t))_{0 \leqslant t < t_m}$, éventuellement orientés.

Dans la suite on supposera toujours que $\forall t \in]\!]0, t_m[\![, V_t = V$. On notera alors $n = |V|$

Mais il peut être utile de ne travailler que sur un seul (gros) graphe, on définit alors la notion le

Définition 2.2 (Graphe de transition)**.** Soit un graphe dynamique $(G_t)_{0 \leqslant t < t_m}$. Son graphe de transition, noté $\mathcal{G} = (V^{\mathrm{T}}, E^{\mathrm{T}})$, est un graphe **orienté** tel que :

- $V^{\mathrm{T}} = \{(x, t), x \in V, 0 \leqslant t \leqslant t_m\}$
- $E^{\mathrm{T}} = \{(x, t)(y, t+1), x = y \vee (x, y) \in E_t, 0 \leqslant t < t_m\}$

Remarque. Le graphe $\mathcal{G}$ est un DAG (voir Fig.1).

Définition 2.3 (Voisinage)**.** Soit $t \in I$.

L'ensemble des voisins d'un sommet u dans G_t est noté $\mathcal{N}_t(u)$.

On étend cette définition à un ensemble de sommets S : $\mathcal{N}_t(S) = \bigcup_{u \in S} \mathcal{N}_t(u)$.

2.1.2 Chemins dynamiques

Définition 2.4 (Chemin)**.** Rappel : un chemin dans un graphe $G = (V, E)$ est une séquence de sommets $u_0, \ldots, u_p$ telle que $\forall i \in [\![1, p]\!], u_{i-1}u_i \in E$, et chaque arête apparaisse au plus une fois : $|\{u_{-1}u_i, 1 \leqslant i \leqslant p\}| = p$ Un tel chemin est de durée p.

Définition 2.5 (Chemin dynamique)**.** On étend la notion à un graphe dynamique $(G_t = (V, E_t))_t$. C'est une séquence de sommets $c_{t_0} = u_0, \ldots, u_p$ telle que $\forall i \in [\![1, p]\!], u_{i-1}u_i \in E_{i-1+t_0} \vee u_{i-1} = u_i$.

- La durée d'un tel chemin est $d(c_{t_0}) = p$
- Longueur du chemin = nombre de sauts effectués : $l(c_{t_0}) = |\{(u_{i-1}u_i, i), u_{i-1} \neq u_i\}|$.
- La date de départ de ce chemin est $t_d(c_{t_0}) = t_0$.
- Sa date d'arrivée est $t_a(c_{t_0}) = t_0 + p$.

On parlera de plus bref chemin pour la durée, et de plus court chemin pour la longueur.

Remarque. Un chemin dynamique dans $(G_t)_t$ coïncide avec la notion de chemin (orienté) dans le graphe de transition $\mathcal{G}$.

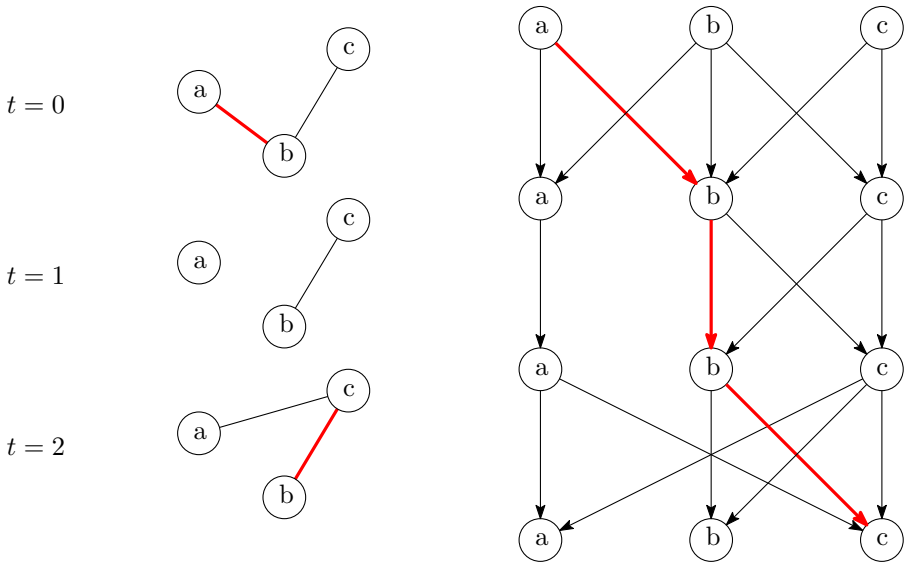


FIGURE 1: Exemple de graphe dynamique avec son graphe de transition associé. En rouge, $c_0 = abbc$, chemin dynamique de date de départ $t_d = 0$.

2.1.3 Structure de données

On se donne une liste de contacts $\langle u, v, t_a, t_b \rangle$ signifiant qu'il y a contact entre u et v entre les dates t_a et t_b. A partir de là, on construit un graphe dynamique qui dépend de deux paramètres ζ et δ : on prélève un échantillons toutes les ζ secondes, en mettant un lien entre deux sommets u et v à la date i s'ils ont été connectés dans l'intervalle de temps $[i \times \zeta, (i+1) \times \zeta[$, et que ce lien dure depuis plus de δ secondes.

Remarque. Avec $\delta = \zeta = 1$, la liste donnée est la liste des intervalles de présence de chacune des arêtes du graphe dynamique.

À partir de là, on génère la liste des arêtes à modifier aux étapes $0, \ldots, p$. On obtient donc une liste L de taille $p+1$, où $L[i][0]$ est la liste des arêtes ajoutées à la date i, et $L[i][1]$ la liste des arêtes supprimées entre G_{i-1} et G_i (en prenant G_{-1} le graphe vide).

Complexité On notera $n = |V|$ le nombre de sommets du graphe. m_{tot} le nombre total d'arêtes listées dans L. m_{max} le nombre maximum d'arête d'un des graphes G_t : $m_{max} = \max |E_t|$. Enfin on notera $p+1$ le nombre de graphe de la séquence $(G_0 \ldots G_p)$.

Dans la suite, on considèrera un ensemble de départ S^0 de sommets porteurs d'information à la date $t = 0$. On souhaite étudier la diffusion de cette information dans le graphe au cours du temps. On va donc définir plusieurs propriétés à étudier, en fonction de l'aspect de la diffusion qui nous intéresse.

2.2 Plus brefs chemins

2.2.1 Durée des plus brefs chemins

On considère un graphe dynamique $(G_t = (V, E_t))_t$ et un sommet $v \in V$.

Définition 2.6 (Temps de diffusion)**.** C'est une fonction du temps qui à une date t associe la plus petite durée d'un chemin de date d'arrivée t terminant sur le sommet v et commençant sur un sommet de S^0. Formellement, en notant $\mathcal{C}h(S^0, v)$ l'ensemble des chemins dynamiques de S^0 à v :

$$\mathcal{T}_{S^0}(v)(t) = \min\{d(c), c \in \mathcal{C}h(S^0, v), t_a(c) = t\}$$

2.2.2 Nombre de plus brefs chemins

On s'intéresse maintenant au nombre de ces plus brefs chemins, entre un sommet de S^0 et un sommet v donné.

Définition 2.7. On définit $\mathcal{C}_{S^0}$ comme une fonction qui à une date t associe le nombre de plus brefs chemins d'un sommet $u \in S^0$ à v de date d'arrivée $t_a = t$. Soit tout simplement :

$$\mathcal{C}_{S^0}(v)(t) = |\{c \in \mathcal{C}h(S^0, v), d(c) = \mathcal{T}_{S^0}(v)(t)\}|$$

Proposition 2.1. *Avec les définitions proposées, on a :*

$$\forall v, t, \mathcal{T}_{S^0}(v)(t+1) = \min\{\mathcal{T}_{S^0}(u)(t), v \in \mathcal{N}_t(u) \vee u = v\} + 1$$
$$\forall v, t, \mathcal{C}_{S^0}(v)(t+1) = \sum_{u \in X_{S^0}(v,t+1)} \mathcal{C}_{S^0}(u)(t)$$

où $X_{S^0}(v, t+1) = \{u \in V, u = v \vee v \in \mathcal{N}_t(u), \mathcal{T}_{S^0}(u)(t)+1 = \mathcal{T}_{S^0}(v)(t+1)\}$ *est l'ensemble des sommets par lesquels arrive un plus bref chemin de* S^0 *à* v *à la date* $t+1$.

Proposition 2.2 (Conditions initiales)**.**

$$\forall v, \mathcal{T}_{S^0}(v)(0) = \mathcal{L}^*_S(v)(0) = \begin{cases} 0 & si\ v \in S^0 \\ +\infty & sinon \end{cases}$$
$$\forall v, \mathcal{C}_{S^0}(v)(0) = \mathcal{E}^*_S(v)(0) = \begin{cases} 1 & si\ v \in S^0 \\ 0 & sinon \end{cases}$$

2.2.3 Algorithme, complexité

Programme 1 Plus brefs chemins

Entrée : La liste L des arêtes modifiées. L'ensemble S^0. L'entier n.
Sortie : Une séquence de tableaux T_t et C_t des valeurs pour chaque sommet.

```
1: C ← Créer_tableau(n, 0) // Valeurs courantes
2: T ← Créer_tableau(n, +∞)
3: Pour i ∈ S^0 faire
4:    C[i] ← 1; T[i] ← 0
5: Fin Pour
6: C_0 ← C; T_0 ← T
7: Pour t = 1 à p + 1 faire
8:    G ← G − L[t][1] + L[t][0] // Mise à jour du graphe
9:    C_t ← C; T_t ← T // Création des nouvelles valeurs
10:   Pour u ∈ V\S^0 faire
11:      ++T_t[u] // Initialisation de T_t
12:   Fin Pour
13:   Pour u ∈ V faire
14:      Pour v ∈ N(u) faire
15:         Si T[u] + 1 < T_t[v] alors // Maj de la durée du plus bref chemin
16:            T_t[v] ← T[u] + 1
17:            C_t[v] ← C[u]
18:         Sinon Si T[u] + 1 = T_t[v] alors // Ajout des chemins issus de u
19:            C_t[v] ← C_t[v] + C[u]
20:         Fin Si
21:      Fin Pour
22:   Fin Pour
23:   C ← C_t; T ← T_t // Maj des valeurs courantes
24: Fin Pour
25: Retourner (E_t)_t
```

Complexité : linéaire, $\mathcal{O}((n + m_{max})p)$.

2.3 Nombre de contacts

2.3.1 Contacts simples

Soit un graphe dynamique $(G_t = (V_t, E_t))_{0 \leqslant t < t_m}$. Notons $I = [\![0, t_m - 1]\!]$

On mesure le nombre de contacts qu'un sommet v a eu avec des sommets porteurs d'information pendant une fenêtre de temps donnée de $\hat{\gamma}$ secondes. On note $\gamma = \hat{\gamma}/\zeta$ le nombre de graphes de la séquence pendant cet intervalle de temps. On peut alors définir nos fonctions :

Définition 2.8 (Indice de contact)**.** Il est défini pour un sommet $v \in V$ comme étant

une fonction du temps, qui à une date t associe le nombre de contact qu'un sommet v a eu avec un sommet de S^0 dans l'intervalle $[\![t-\gamma, t]\!]$.

$$\mathrm{I}_{S^0,\gamma}(v)(t) = \sum_{i=t-\gamma+1}^{t} \left|S^0 \cap \mathcal{N}_i(v)\right|$$

2.3.2 Contacts étendus

On étend maintenant les considérations de contacts avec des sommets pour lesquels il existe un chemin dynamique de S^0 à celui-ci entre les dates $t-\gamma$ et t. Cela revient à dire que quand un sommet porteur d'information est en contact avec un autre, il lui transmet cette information. On formalise cela par une fonction "de diffusion" comme suit :

Définition 2.9 (Diffusion canonique)**.** On note $f_c(S^0, t_d, t)$ l'ensemble obtenu à l'instant t, en diffusant à partir de la date t_d avec les sommets sources de S^0. Concrètement :

$$f_c(S^0, t_d, t) = \begin{cases} \emptyset & \text{si } t < t_d \\ S^0 & \text{si } t = t_d \\ f_c(S^0, t_d, t-1) \cup \mathcal{N}_{t-1}(f(S^0, t_d, t-1)) & \text{sinon} \end{cases}$$

Cette fois on va donc mesurer le nombre de contact avec l'ensemble S^0, étendu à chaque pas de temps avec les nouveaux sommets. On utilise donc la fonction f_c sus-définie :

Définition 2.10 (Indice de contact étendu)**.** On le définit pour un sommet $v \in V$, comme étant une fonction du temps qui à une date t associe le nombre de contact avec la séquence d'ensembles qui convient dans l'intervalle de temps $[\![t-\gamma, t]\!]$:

$$\mathrm{I}^*_{S^0,\gamma}(v)(t) = \sum_{i=t-\gamma+1}^{t} \left|f(S^0, t-\gamma+1, i) \cap \mathcal{N}_i(v)\right|$$

2.3.3 Algorithme, complexité

Programme 2 Indices d'exposition

Entrée : La liste L des arêtes modifiées. L'ensemble S^0. L'entier n.
Sortie : Une séquence de tableaux I_t et I_t^* des indices de contact simple et étendus.

1: **Pour** $t = 0$ à $p - \gamma$ **faire**
2: $I, I^* \leftarrow$ Créer_tableau$(n, 0)$ // Les indices courants, départ à $t_d = t$
3: $S \leftarrow S^0$
4: $G_{copy} \leftarrow G$
5: **Pour** $i = t$ à $t + \gamma$ **faire** // Fenêtre de temps de γ étapes
6: $G \leftarrow G - L[t][1] + L[t][0]$ // Mise à jour du graphe
7: **Pour** $u \in V$ **faire**
8: **Pour** $v \in \mathcal{N}(u)$ **faire**
9: $++I[v]$ **si** $v \in S^0$
10: $++I^*[v]$ **si** $v \in S$
11: **Fin Pour**
12: **Fin Pour**
13: **Pour** $u \in S$ **faire** // Mise à jour de S
14: **Pour** $v \in \mathcal{N}(u)$ **faire**
15: $S \leftarrow S \cup \{v\}$
16: **Fin Pour**
17: **Fin Pour**
18: **Fin Pour**
19: $G \leftarrow G_{copy}$
20: $G \leftarrow G - L[t][1] + L[t][0]$ // Mise à jour du graphe
21: $I_t \leftarrow I; I_t^* \leftarrow I^*$
22: **Fin Pour**
23: **Retourner** $(I_t)_t, (I_t^*)_t$

Complexité : le coût des lignes 2 à 4 et 19 à 21 est dominé par le coût de la boucle intermédiaire. La boucle lignes 7 à 12 a un coût $\mathcal{O}(n + m_{max})$, car les tests lignes 9 et 10 peuvent se faire en temps constant. De même pour la boucle lignes 13 à 18. Finalement on a bien un coût total $\mathcal{O}((n + m_{max})p \times \gamma)$.

2.4 Flot dynamique

2.4.1 Généralités

Les deux mesures précédentes offrent en quelques sortes une borne inférieure et une borne supérieure de l'information qui pourrait transiter entre les sources dans S^0 et un sommet v du graphe. La notion qui vient alors naturellement à l'esprit consiste en un compromis entre les deux : le flot maximal, entre deux sommets du graphe à deux instants différents.

Définition 2.11 (Graphe dynamique pondéré)**.** C'est une séquence $(G_t = (V, E_t, c_t))_t$ où $(V_t, E_t)_t$ est un graphe dynamique (éventuellement orienté), et $(c_t)_t$ est une séquence de fonctions de $V \times V \to \overline{\mathbb{R}}_+$ qui vérifie $\forall t, \forall x, y \in V, xy \notin E_t \wedge x \neq y \Rightarrow c_t(x, y) = 0$ (la fonction est nulle en dehors des arêtes et des sommets eux-mêmes).

On distingue également une source (s, t_0) et un puits (a, t_1) (avec $t_0 \leqslant t_1$).

Définition 2.12 (Flot dynamique)**.** Soit $(G_t)_t$ un graphe dynamique pondéré, ainsi qu'une source (s, t_0) et un puits (a, t_1). Un flot dynamique est une séquence de fonction $(\varphi_t)_{t_0 \leqslant t \leqslant t_1}$ qui vérifie :

1. **Positivité** : $\forall t, \forall x, y \in V, \varphi_t(x, y) \geqslant 0$
2. **Contrainte de capacité** : $\forall t, \forall x, y \in V_t, \varphi_t(x, y) \leqslant c_t(x, y)$
3. **Conservation** : $\forall t_0 \leqslant t < t_1, \forall x \in V, \sum_y \varphi_t(y, x) = \sum_y \varphi_{t+1}(x, y)$
4. **Conditions aux limites** :
   a) $\forall x, y \in V, x \neq s \Rightarrow \varphi_{t_0}(x, y) = 0$
   b) $\forall x, y \in V, y \neq a \Rightarrow \varphi_{t_1}(x, y) = 0$

Remarque. En quoi est-ce bien un flot ? Tout simplement car la notion se transporte aisément au graphe de transition $\mathcal{G}$, dans lequel le flot dynamique devient un flot au sens usuel. D'où là :

Proposition 2.3. *Il existe une bijection canonique entre un flot dynamique sur $(G_t)_t$ et un flot sur son graphe de transition $\mathcal{G}$.*

Démonstration. Pour le sens $(G_t)_t \Rightarrow \mathcal{G}$:

Soit un graphe dynamique pondéré $(G_t = (V, E_t, c_t))_{0 \leqslant t < t_m}$. Notons $I = [\![0, t_m - 1]\!]$ et $\mathcal{G} = (V^{\mathrm{T}}, E^{\mathrm{T}})$ le graphe de transition de $(G_t)_t$.

On adapte la fonction de capacité aux sommets de $\mathcal{G}$:

$$c^{\mathrm{T}} : V^{\mathrm{T}} \times V^{\mathrm{T}} \to \overline{\mathbb{R}}_+$$
$$(x, t_x), (y, t_y) \to \begin{cases} c_{t_x}(x, y) & \text{si } t_y = t_x + 1 \\ 0 & \text{sinon} \end{cases}$$

Il est facile de voir d'après sa définition (cf. 2.2) que le graphe $\mathcal{G}$, munie de cette fonction de capacité, définit bien un graphe pondéré. On distingue d'ailleurs dans $\mathcal{G}$ les sommets source (s, t_0) et puits (a, t_1).

Considérons maintenant un flot dynamique $(\varphi_t)_{t_0 \leqslant t \leqslant t_1}$ pour $(G_t)_t$. On construit alors un flot ψ pour $\mathcal{G}$ de la manière suivante :

$$\psi(x, t_x, y, t_y) = \begin{cases} \varphi_{t_x}(x, y) & \text{si } t_y = t_x + 1 \wedge t_x \in [\![t_0, t_1]\!] \\ -\varphi_{t_y}(y, x) & \text{si } t_x = t_y + 1 \wedge t_y \in [\![t_0, t_1]\!] \\ 0 & \text{sinon} \end{cases}$$

Vérifions que ψ respecte les contraintes de flot usuelles :

Pseudo-symétrie $\forall x, y \in V, \forall t_x, t_y \in I, \psi(x, t_x, y, t_y) = -\psi(y, t_y, x, t_x)$: découle immédiatement de la définition de ψ

Contrainte de capacité Comme il n'y a d'arc qu'entre des sommets des la forme $(x, t)(y, t + 1)$, le seul cas à vérifier est celui où $t_y = t_x + 1$. On a alors par définition :

$$\psi(x, t_x, y, t_y) = \varphi_{t_x}(x, y) \leqslant c_{t_x}(x, y) = c^{\mathrm{T}}(x, t_x, y, t_y)$$

Conservation On veut montrer que pour tout sommet (x, t_x) différent de (s, t_0) et (a, t_1), on a $\sum_{(y,t_y)} \psi(x, t_x, y, t_y) = 0$. Notons que c'est déjà le cas si $t_x \notin [\![t_0, t_1 + 1]\!]$. Il reste 3 cas à traiter :

1. **Cas** $t_x = t_0$. Immédiat d'après les conditions aux limites :

$$\begin{aligned}\sum_{(y,t_y)} \psi(x, t_x, y, t_y) &= \sum_y \psi(x, t_0, y, t_0 + 1) \\ &= \sum_y \varphi_{t_0}(x, y) = 0 \text{ car } x \neq s\end{aligned}$$

2. **Cas** $t_x = t_1 + 1$. Idem.
3. **Cas** $t_0 < t_x \leqslant t_1$. On décompose :

$$\begin{aligned}\sum_{(y,t_y)} \psi(x, t_x, y, t_y) &= \sum_y \psi(x, t_x, y, t_x + 1) + \sum_y \psi(x, t_x, y, t_x - 1) \\ &= \sum_y \varphi_{t_x}(x, y) \qquad\qquad - \sum_y \varphi_{t_x - 1}(y, x)\end{aligned}$$

Pour ce qui du sens réciproque, on notera seulement que la démarche est assez similaire, le flot dynamique se déduisant facilement grâce à la définition de $\mathcal{G}$. □

Définition 2.13 (Valeur d'un flot dynamique)**.** On définit la valeur $|\varphi|$ du flot de manière analogue au cas statique : $|\varphi| = \sum_y \varphi_{t_0}(s, y)$.

2.4.2 Cas particulier

Maintenant qu'on dispose d'un "cadre" formel pour parler de flot dans un graphe dynamique, précisons les mesures qui nous intéressent vis-à-vis de notre problème de diffusion. Tout simplement, on souhaite que sur une arête d'un graphe G_t puisse transiter 1 unité d'information, et que cette information puisse attendre à volonté sur n'importe quel sommet. Formellement, on définit donc une :

Définition 2.14 (Capacité canonique)**.** C'est la séquence de fonction $(c_t)_t$ qui, pour t donné, vaut 1 sur E_t, $+\infty$ sur $\{(x, x), x \in V\}$, et 0 sinon. On la notera $(\vec{c}_t)_t$.

Revenons à notre problème de diffusion à partir d'un ensemble S^0 à l'instant $t_0 = 0$. On peut facilement se ramener au problème de flot maximum (flot dynamique) an ajoutant un sommet factice s dans V ainsi que les arêtes $\{sx, x \in S^0\}$ aux ensembles E_t.

On ajoute également un graphe G_{-1} à la série (qui ne contient que les arêtes partant de s). Il suffit alors de chercher un flot maximum entre $(s, t_0 - 1)$ et (v, t_1) pour un sommet v à une date t donnée. On peut donc définir une nouvelle fonction :

Définition 2.15 (Flot d'information)**.** C'est une fonction qui à un sommet v associe la fonction qui à une date t associe la valeur du flot maximal arrivant sur v à la date t et partant d'une source s à une date $t - \gamma$. Formellement :

$$\mathcal{Q}_{s,\gamma}(v)(t) = \max\{|\varphi|, \varphi \text{ flot dynamique de } (s, t - \gamma + 1) \text{ à } (v, t)\}$$

NB : Les fonctions de capacités utilisées sont celles de la séquence $(\vec{c}_t)_t$.

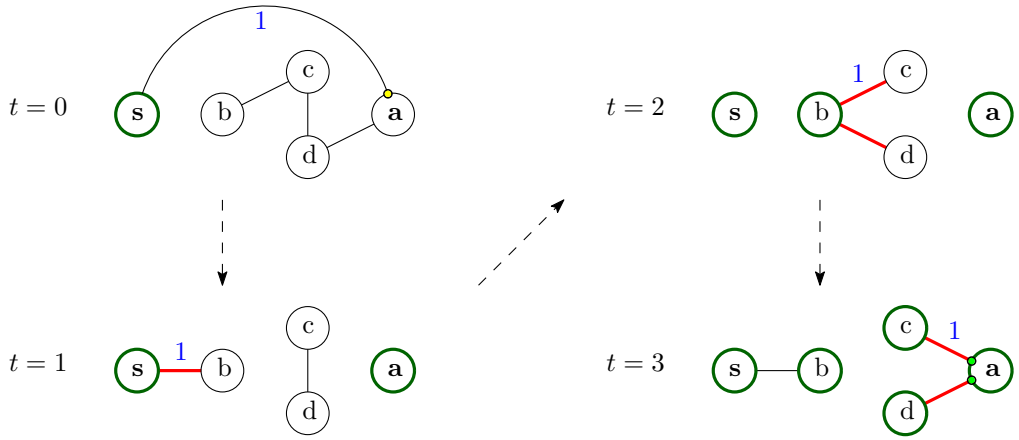


FIGURE 2: Illustration des différentes notions abordées.

Propriété	Valeur	Légende
Temps de diffusion	$\mathcal{T}_{\{s\}}(a)(4) = 2$	Chemins en rouge
Nb de plus brefs chemins	$\mathcal{C}_{\{s\}}(a)(4) = 2$	Chemins en rouge
Contacts simples	$\mathrm{I}_{\{s\},\gamma=4}(a)(3) = 1$	Points en jaune
Contacts étendus	$\mathrm{I}^*_{\{s\},\gamma=4}(a)(3) = 3$	Points en jaune ou vert
Flot dynamique	$\mathcal{Q}_{s,\gamma=4}(a)(3) = 2$	Valeurs indiquées en bleu

Avec $S^0 = \{s\}$, et $f_c(S^0, 0, t)$ indiqué par les cercles vert foncé.

2.4.3 Structure de donnée

En pratique, on calculera un flot maximum sur le graphe de transition, entre deux sommets de celui-ci. Ce graphe étant potentiellement assez grand, on le représente de manière compacte par un liste d'adjacence, en utilisant deux tableaux :

A : tableau de taille m_{tot}, contient la liste des arêtes (u, v) dans l'ordre lexicographique

I : tableau de taille $(n+1) \times p$, contient les intervalles d'indice de A correspondant aux différents sommet

Exemple 2.1. $I[2] = 5$ signifie que $A[5] = (2, \cdots)$ mais que $A[3] \neq (2, \cdots)$.

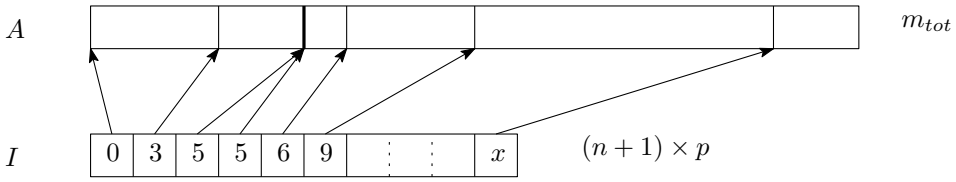


FIGURE 3: Le tableau I permet de s'y retrouver dans le tableau A

Remarque. On ne représente pas les arêtes "verticales" de la forme $(u, u+n)$.

2.4.4 Algorithme, complexité

L'algorithme utilisé est celui d'Edmonds-Karp, que l'on fait tourner sur un sous-graphe de $\mathcal{G}$. Il y a sans doute plus efficace (préflot ? algorithme spécifique à notre graphe ?), aussi on ne s'étonnera pas que cet algorithme ait mis plus de temps à tourner en pratique.

Notons que la fonction de capacité étant connue, elle n'a pas besoin d'être représentée en dur dans la mémoire. De même, le flot courant que l'on maintient vaut soit 0 soit 1 sur les arêtes entre deux sommets distincts dans V, on utilise donc un tableau de booléen de taille m_{tot}. Pour le flot courant qui "attend" sur un sommet de V (et donc qui transite sur une arête de type $(u, u+n)$), on utilise un tableau de $n \times p$ entiers.

On note alors que le sous-graphe de $\mathcal{G}$ utilisé contient $n \times \gamma$ sommets, et au plus $(m_{max} + n) \times \gamma$ arêtes. Soit une complexité temporelle $\mathcal{O}((n \cdot m_{max}^2 + n^3) \cdot \gamma^3)$. La complexité spatiale est quant à elle $\mathcal{O}(m_{tot} + n \cdot p)$

3 Résultats expérimentaux

3.1 Valeurs mesurées

Après avoir implémenté les différentes fonctions présentées plus haut, nous avons pu effectuer des mesures sur les expériences Reality Mining [EPL09] et plus particulièrement Infocom 2005 [KH05], décrites au début de ce papier.

Les fonctions décrites dans les sections antérieures sont de la forme $f_{S^0}(v)(t)$, $f_{S^0,\gamma}(v)(t)$ ou encore $f_{s,\gamma}(v)(t)$. Dans nos calculs, on s'est limité au cas où $S^0 = \{u\}$ est un singleton.

On a ensuite fixé arbitrairement un paramètre γ. Au final on a calculé les valeurs prises par $f_u(v)(t)$ pour tout triplet $(u, v, t) \in V \times V \times [\![0, t_m - 1]\!]$.

On obtient une suite de valeurs sous la forme `date cible source valeur`. À partir de là on peut ensuite filtrer les données et établir des scores :

- Sur tous les triplets (u, v, t)
- Sur les sommets cibles v (moyenne sur les sommets source + temps)
- Sur les dates t (moyenne sur les couples de sommets cibles/source)
- Sur les couples de sommets (u, v) (moyenne sur le temps)

On peut alors tracer leur **distribution simple** (nombre de fois que telle valeur est atteinte en fonction de cette valeur), **cumulative** (nombre de fois qu'une valeur inférieure ou égale est atteinte), ou **cumulative inverse** (nombre de fois qu'une valeur supérieure ou égale est atteinte).

Enfin, on peut aussi regarder les **corrélations** entre les différents scores, en affichant le nuage des points de coordonnées $(f(x), g(x))$, où x est le paramètre à faire varier, et f, g les fonctions dont on veut observer la corrélation.

Remarque. En pratique, pour éviter d'obtenir des mesures aberrante sur certaines valeurs (temps de diffusion infini car un sommet est isolé par exemple), on a considéré les données comme "cycliques". C'est à dire que l'on a fait tourner l'algorithme sur les mêmes données jusqu'à ce que tous les $f_u(v)(t)$ soient définis, avant de lancer l'enregistrement. Ce fut possible car dans chaque cas la version non orientée du graphe de transition $\mathcal{G}$ s'avère être connexe.

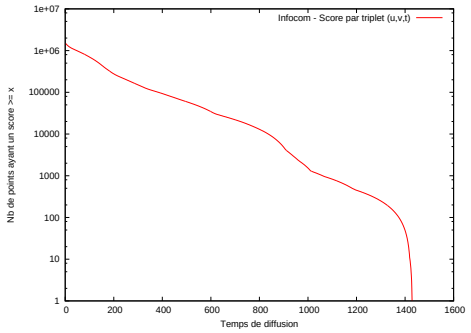
Infos pratiques Sur les courbes présentées, les données d'Infocom sont étudiées avec les paramètres $\zeta = 300s$ (temps entre deux échantillons), $\delta = 1s$ (durée minimum d'un contact valide), et $\hat{\gamma} = 50 \times 300s$ (durée de la mesure pour le nombre de contact et le flot dynamique). Les paramètres pour Reality Mining (MIT) sont $\zeta = 3600s$, $\delta = 1s$ et $\hat{\gamma} = 24 \times 3600s$.

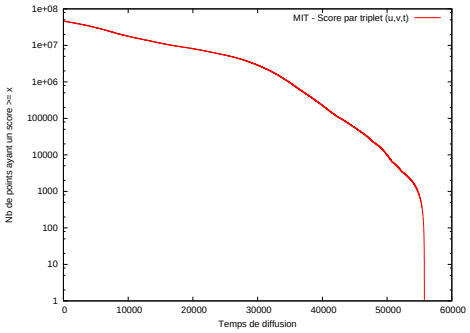
3.2 Distributions obtenues

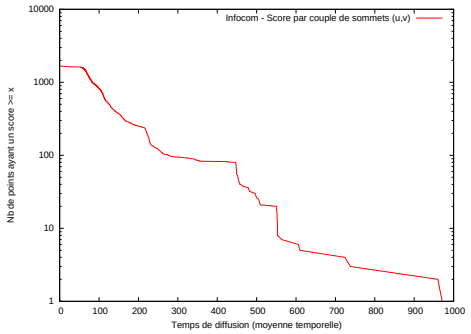
3.2.1 Plus brefs chemins

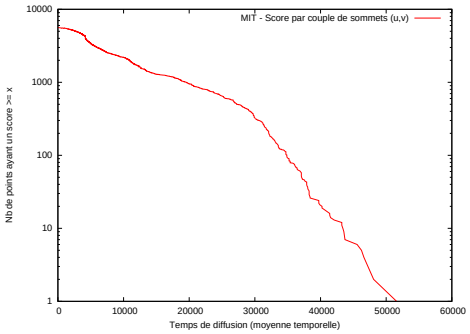
Cf. Fig.4 : les distributions cumulatives inverses pour le temps de diffusion ainsi que la durée du plus bref chemin entre les couples de sommets suit une loi vaguement exponentielle. En revanche on ne peut pas dire grand chose des valeurs moyennées sur le temps, qui n'ont pas vraiment la même allure pour Infocom et MIT.

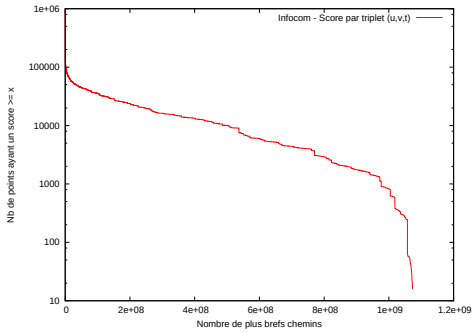
FIGURE 4: Distributions cumulatives inverses, Infocom (Gauche) et MIT (Droite).
Temps de diffusion par triplet / couple de sommets (moyenne temporelle)
Nombre de plus brefs chemins par triplet / couple de sommets (moyenne temporelle)

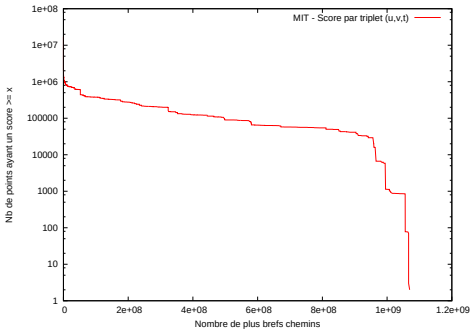


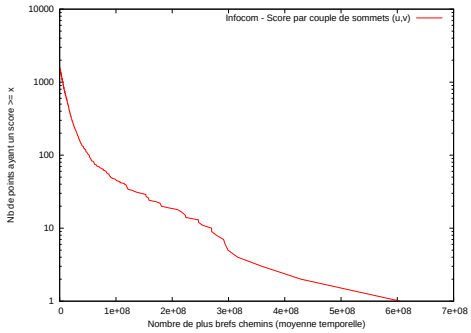


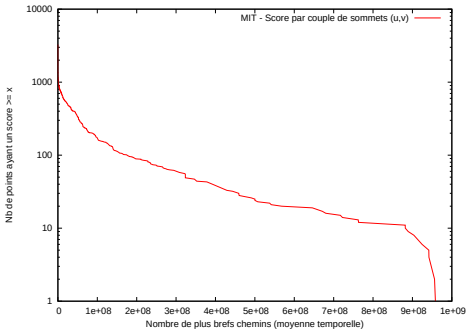












Remarque. Afin d'éviter les overflows, on a tronqué à partir d'une certaine valeur dès que le nombre de plus brefs chemins devenait trop important (explosion combinatoire : il suffit par exemple qu'une composante connexe reste isolée de la source pour que la durée du plus court chemin ne soit plus mise à jour).

Rétrospectivement, il aurait sans doute mieux valut regarder les chemins de longueur minimale, ce qui aurait évité une trop grande dispersion due à l'explosion combinatoire des valeurs.

3.2.2 Nombre de contacts, flot dynamique

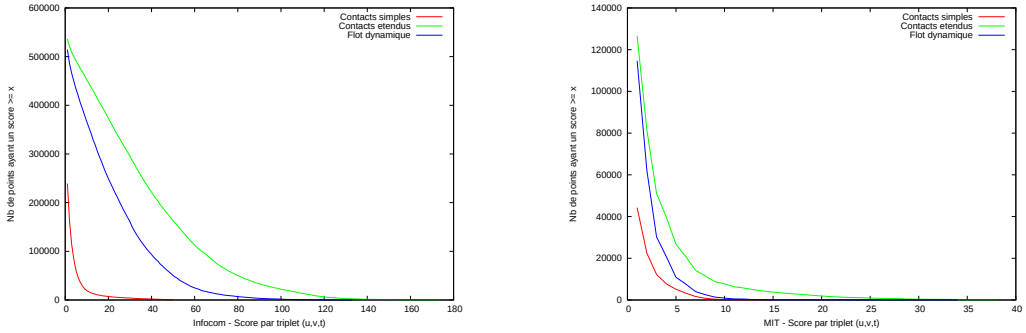


FIGURE 5: Distributions cumulatives inverses, Infocom (Gauche) et MIT (Droite).

La Fig.5 présente sur un même graphe les distributions pour les trois valeurs : nombre de contacts simples (I), contacts étendus (I^*), et flot dynamique (φ), enregistrés sur une fenêtre de γ graphes. On observe bien l'inégalité $I \leqslant \varphi \leqslant I^*$, qui découlent des différentes définitions.

Globalement, on observe que la mesure des contacts étendus approche mieux la valeur du flot dynamique que celle — plus élémentaire — des contacts simples. C'est intéressant d'un point de vue complexité, car le flot dynamique est actuellement assez long à calculer.

3.3 Corrélations entre les valeurs

On peut maintenant s'intéresser à voir si certaines de ces valeurs sont effectivement corrélées, ou si elles n'ont rien à voir entre elles. Il y a beaucoup de croisement que l'on pourrait faire, certains paraissant plus pertinents que d'autres. La Fig.6 présente le lien entre les moyennes temporelles du temps de diffusion et du nombre de plus brefs chemins entre les couples de sommets. Comme on pouvait s'y attendre, on ne voit pas vraiment de corrélation intéressante émerger.

Toutefois on peut noter que pour Infocom le nombre de chemins est assez élevé, alors qu'il est plus étalé même sur les petites valeurs dans le cas de MIT. Peut-être est-ce dû aux paramètres de la mesure ? En tout cas cela semble indiquer qu'il y a plus de brefs trajets dans le graphe des étudiants du MIT que chez les participants d'Infocom.

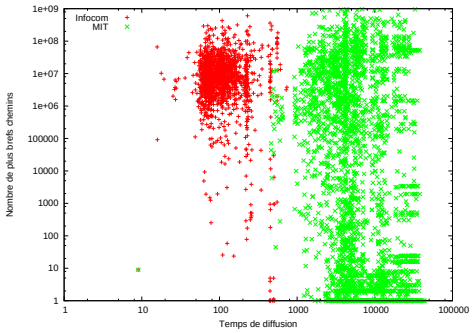


FIGURE 6: Des données pas franchement corrélées. Infocom (Rouge) et MIT (Vert).

Une autre corrélation plus intéressante à établir concerne la relation entre les contacts simples, étendus et le flot dynamique, comme on pouvait s'en douter plus haut. La Fig.7 présente ces corrélations dans le cas des données Infocom. Le nuage de point a globalement la même allure pour les données MIT, mais à moindre échelle. On notera bien entendu les valeurs extrêmes imposées par la fenêtre, ainsi que l'inégalité $I \leqslant \varphi \leqslant I^*$ qui explique les demi-plan vide. Bon dans le cas contacts étendus/flot dynamique, l'éventail est un peu plus étalé et nous laisse parfois une marge de 50% (cf. un flot entre 5 et 15 pour un nombre de contacts étendus mesuré de 20), mais c'est déjà ça !

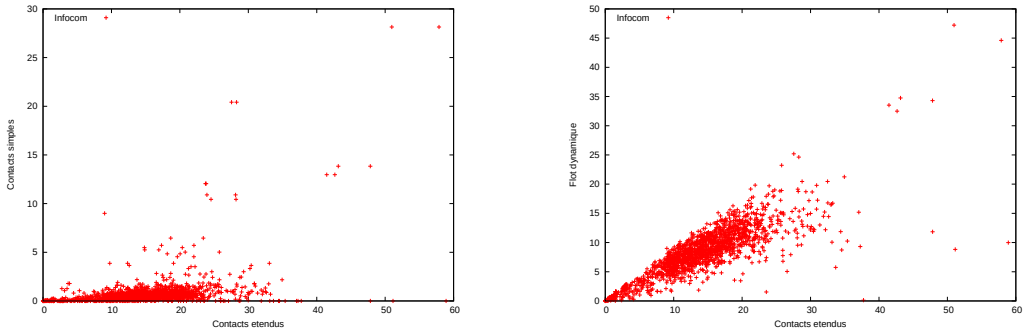


FIGURE 7: Corrélations entre contacts étendus/contacts simples (Gauche) et contacts étendus/flot dynamique (Droite).

Conclusion, Perspectives

On conclura ce rapport en soulignant les résultats prometteurs liés à la notion de flot dynamique dans le graphe de nos relations, mais il faut quand même insister sur la nécessiter d'étudier certains de ces paramètres plus en détail avant de pouvoir en retenir quelque chose.

La qualité des données sur lesquelles on travaille joue beaucoup : l'étude des transmissions bactériologiques réelles grâce au projet I-Bird par exemple, ouvre de nouveaux axes d'études des notions présentées ici. Les artefacts inhérents aux expérimentations peuvent aussi nous jouer des tours si l'on n'est pas attentif (par exemple pour Infocom, des capteurs dans des pièces adjacentes peuvent détecter leur présence mutuelle, ou à la fin de l'expérience les boitiers rangés dans les cartons continuer d'enregistrer, etc.).

On peut également envisager d'approfondir l'étude des mesures présentées ci-dessus selon différents axes : calcul de médiane, quartile, écart-type, etc. Mais aussi comparaisons des courbes obtenues avec des mesures effectuées sur des graphes dynamiques totalement aléatoires, ou qui modélisent déjà des communautés vis-à-vis de certaines propriétés (nombre de triangles par exemple), tels que présentés dans [SBF+08].

Une autre voie qui n'a pas eu le temps d'être explorée faute de temps, aurait été de faire du **clustering** sur ces valeurs. Le partitionnement de données vise à identifier des sous-ensembles de sommets partageant des caractéristiques communes, vis à vis d'une certaine propriété. Par exemple on peut vouloir regrouper dans un même ensemble les sommets qui ont deux à deux une moyenne de contacts entre eux plus élevée.

Références

[BXFJ02] B. Bui Xuan, Afonso Ferreira, and Aubin Jarry. Computing shortest, fastest, and foremost journeys in dynamic networks. 0 RR-4589, INRIA, 10 2002.

[EPL09] Nathan Eagle, Alex Pentland, and David Lazer. Inferring social network structure using mobile phone data. *Proceedings of the National Academy of Sciences (PNAS)*, 106(36) :15274–15278, Sep 2009.

[KH05] David Kotz and Tristan Henderson. CRAWDAD : A Community Resource for Archiving Wireless Data at Dartmouth. *IEEE Pervasive Computing*, 4(4) :12–14, Oct–Dec 2005.

[SBF+08] Antoine Scherrer, Pierre Borgnat, Eric Fleury, Jean-Loup Guillaume, and Céline Robardet. Description and simulation of dynamic mobility networks. *Computer Networks*, 52(15) :2842–2858, 10 2008.