Biais dans les mesures obtenues par un réseau de capteurs sans fil

Adrien Friggeri, Guillaume Chelius

Biais dans les mesures obtenues par un réseau de capteurs sans fil

Adrien Friggeri[1] and Guillaume Chelius[2]

[1] *LIP UMR 5668/ENS de Lyon, DNET/INRIA, Université de Lyon*
[2] *DNET/INRIA, LIP UMR 5668/ENS de Lyon, Université de Lyon*

Dans le domaine des réseaux complexes, la recherche a été stimulée par la disponibilité de grands jeux de données obtenus de manière automatique. Dans cet article, nous nous intéressons en particulier à des données d'interactions au sein d'un service hospitalier, récoltées à l'aide d'un réseau de capteurs sans fil. Nous mettons en avant le biais introduit par le système de mesure et proposons une méthode de reconstruction du signal d'origine permettant de mettre en évidence des phénomènes qui n'étaient pas visibles sur les données brutes.

Keywords: réseaux complexes, réseaux de capteurs, applications médicales, reconstruction de données, biais dus à la mesure

1 Introduction

Les réseaux complexes apparaissent dans des domaines aussi variés que la sociologie, les réseaux informatiques, ou la médecine. Les deux dernières décennies ont connu une forte croissance de leurs études, après qu'il eut été montré [FFF99, NBW06, WS98] que la plupart des réseaux extraits de la réalité partagent des propriétés non triviales. Dans de nombreux domaines d'études, la notion de relation entre entités du réseau découle d'une interaction sociale ou d'une proximité physique qui sont traditionnellement mesurées à l'aide d'audits et d'interviews. Non seulement ces méthodes ne sont pas exhaustives mais leur fiabilité dépend de facteurs humains, ce qui conduit à des limitations dans l'étude des domaines cités.

Les avancées techniques effectuées dans le domaine des réseaux mobiles offrent une opportunité prometteuse de récolter en continu des données sur ces relations en fournissant une méthode de mesure passive et automatique [ASC+09, CBVdB, HCS+05]. Il est cependant important de noter que la fiabilité d'une telle approche n'est pas garantie car la relation entre distance radio (atténuation du signal) et distance physique est loin d'être déterministe. Ce biais est rarement pris en compte dans l'étude de ces réseaux ce qui conduit à une analyse de données qui ne reflète pas la réalité. Se pose alors la problématique de transformer les données d'une mesure non fiable en utilisant une estimation de l'erreur due au système de mesure lui même pour obtenir une représentation fidèle des interactions originales.

Dans cet article, nous décrivons tout d'abord le contexte et les méthodes utilisées (Section 2), nous estimons l'erreur due à la mesure (Section 3) et proposons une méthode de reconstruction du signal d'origine (Section 4). Nous décrivons les résultats obtenus dans la Section 5.

2 Récolte de données

Ces travaux sont issus du projet AFFSET TubExpo, dont le but est l'étude de l'exposition du personnel hospitalier à la tuberculose dans son environnement de travail. Ce projet se concentre sur l'évaluation de l'intensité et la fréquence des contacts entre personnel soignant et patients infectés au sein du *Service des Maladies Infectieuses et Tropicales* (SMIT) de l'hôpital Bichat-Claude Bernard à Paris.

De manière à mesurer l'exposition du personnel soignant au bacille de la tuberculose, nous avons mesuré le temps de présence de chaque individu dans chacune des chambres du service, sur une période de trois mois. Pour cela, deux approches différentes ont été utilisées.

Audits Effectuer des audits est une méthode classique d'obtention de données sur les interactions sociales ou les habitudes de travail, en particulier dans un environnement médical. Étant donnée la nature de la mesure, il est garanti qu'il n'y a aucun faux positif. Cependant elle dépend de la présence d'un observateur au sein du service et n'est donc pas exhaustive. Dans notre cas, 48 visites ont été enregistrées. Pour chacune, les numéros de la chambre et du visiteur, l'heure d'entrée et la durée de visite ont été notées. Dans cet ensemble de visites, la durée moyenne est de 3min 26s, la plus courte étant de 10s et la plus longue de 18min.

Réseau de capteurs Pour automatiser la détection de présence du personnel au sein du service, les 32 chambres ont été équipées de capteurs fixes et un capteur mobile a été donné à chaque membre de l'équipe hospitalière. Toutes les 5 secondes, chaque capteur mobile envoyait un paquet signalant sa présence et le (ou les) capteur fixe le recevant enregistrait la présence du capteur mobile à cet instant. L'ensemble de donnée résultant consiste en 16 066 096 contacts sur une période de 98 jours, entre 56 personnels soignants et 32 chambres. Étant donné la nature discrète de la mesure, nous agrégeons les contacts successifs et appelons *visite* une suite contiguë de contacts.

Nous travaillons dans la suite sur un sous ensemble des données récoltées, correspondant aux interactions entre le personnel soignant et les chambres dans lesquelles se trouvaient 63 patients suivis car étant des cas suspectés de tuberculose.

3 Modèle d'erreur & méthode de reconstruction

3.1 Faux négatifs

Mise en évidence de la perte de paquets On appelle *intercontact* la durée d'absence entre deux visites successives entre un même couple de capteurs (fixe, mobile). Cette durée représente le temps d'absence entre deux visites dans une chambre du même personnel soignant.

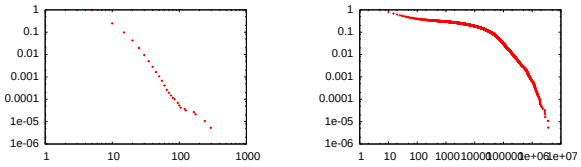

FIGURE 1: Distributions de durées de visites (gauche) et d'intercontacts (droite).

En observant les distributions des durées de visites et d'intercontacts (Figure. 1), on remarque que 90% des visites durent au plus 10s et qu'il y a par ailleurs plus de 30% des intercontacts qui durent 15s ou moins. Nous expliquons ces deux faits par la volatilité du médium radio utilisé pour transporter l'information de contact entre les capteurs fixes et mobiles.

Nous faisons l'hypothèse d'une perte de paquet uniforme et stationnaire et nous proposons une méthode basée sur les intercontacts et validée grâce aux données d'audits, pour déterminer la probabilité p de mesurer un contact sachant qu'il a lieu.

Probabilité de mesure d'un contact On suppose qu'il existe une *durée minimale d'intercontact*. Ceci correspond au fait qu'un personnel soignant ne sort pas d'une chambre pour y rentrer quelques secondes plus tard. Nous supposons qu'il est légitime de considérer que cette durée minimale est supérieure à 20s.

Ceci signifie que les intercontacts de moins de 20s sont la conséquence d'une perte de paquets, et donc la probabilité d'obtenir un intercontact de durée $5k$ est exactement égale à la probabilité de perdre k paquets successifs. Soit, en faisant l'hypothèse d'uniformité et de stationnarité, $(1-p)^k$. La mesure nous permet d'évaluer $p \sim 0.13$.

Validation grâce aux audits En comparant les audits et les contacts obtenus pendant la durée des audits, on peut évaluer p comme étant la fraction de temps de présence mesuré par temps de présence effectif.

De cette manière, on estime $p \sim 0.145$, ce qui est cohérent avec la valeur obtenue par observation des intercontacts courts.

3.2 Faux positifs

Notons par ailleurs que, toujours en raison de la volatilité du médium radio, l'existence d'un contact ne garantit pas une présence effective dans la chambre. En effet, un personnel soignant passant devant une porte ouverte, ou étant dans une pièce adjacente peut être perçu avec une probabilité faible. Cependant, il ne faut pas perdre de vu qu'en raison du grand nombre de contacts mesurés, même si cette probabilité est faible, elle peut engendrer un nombre important de faux positifs. Nous supposons que la probabilité d'avoir un tel faux positif est stationnaire et uniforme. Ainsi, la probabilité d'avoir une succession de k faux positif successifs diminue exponentiellement en fonction de la durée.

4 Méthode de reconstruction

Ces observations nous permettent d'élaborer la méthode de reconstruction suivante :

- Faire apparaître les faux négatifs
  - déterminer la durée minimale d'intercontact d_i ;
  - agréger les contacts consécutifs séparés d'au plus d_i secondes ;
- Supprimer les faux positifs
  - déterminer la durée minimal de visite d_v ;
  - conserver uniquement les visites d'au moins d_v secondes (suppression des faux positifs).

4.1 Durée minimale d'intercontact

Comme expliqué précédemment, les intercontacts courts sont uniquement un effet de la perte de paquets. En observant la distribution des durées des intercontacts pour chaque chambre, on observe un changement de comportement (inflexion de la dérivée de la distribution cumulative inversée) aux alentours de 180s pour chacune des chambres (Figure. 2). Ce changement de comportement traduit le fait qu'au delà de cette durée, les intercontacts ne sont pas uniquement dus à la perte de paquet mais aussi au comportement du personnel soignant et donc aux intercontacts réels. On pose donc $d_i = 180$s.

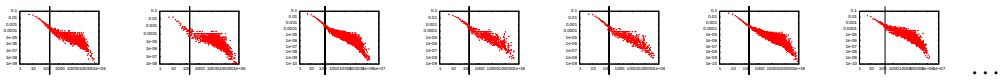

FIGURE 2: Dérivée à droite de la distribution cumulative inversée des durées d'intercontacts dans différentes chambres. La barre verticale indique 180s.

4.2 Durée minimale de visite

On agrège dans un premier temps les contacts séparés d'au plus d_i secondes et on trace, pour chaque durée d et chaque chambre c, la fraction de la durée cumulée dans la chambre c due aux visites de durées d en fonction de la proportion de visites de durée d dans cette chambre. Sur la figure (Figure. 3) on observe que les visites de 5s (encadrées sur la figure) ont une contribution radicalement différente des visites d'autres durées. Nous affirmons que cette différence traduit le fait que les visites de 5s sont dues à des faux positifs et que donc la durée minimale de contact est d'au moins 10s. Étant donné qu'il n'y a pas de séparation claire entre les contributions des contacts de plus de 10s, on pose dans la suite $d_v = 10$.

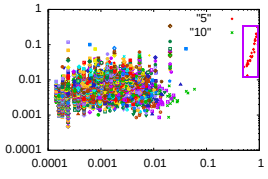

FIGURE 3: Contributions des visites, par chambre et durée, au total des visites et des durées

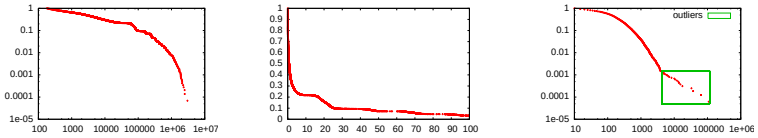

FIGURE 4: Distribution des durées d'intercontacts après reconstruction (gauche) et zoom sur les 100 premières heures (milieu). Distribution des durées de visites après reconstruction (droite).

5 Résultats

Après reconstruction, on obtient la courbe d'intercontacts donnée en Figure 4. L'irrégularité de celle ci est expliquée par le rythme de travail du personnel soignant et notamment par le fait que la durée de travail maximale sur une journée (respectivement de repos) est de l'ordre de 8h (resp. 16h). Ainsi, par exemple, il y a peu d'intercontacts d'une durée compris entre 8h et 16h.

Cette observation nous conforte dans l'idée que la méthode de reconstruction est correcte, car elle met en évidence *a posteriori* un comportement réel qui n'était pas explicite dans la distribution originelle et qui n'est pas lié à un biais de la reconstruction.

La distribution des durées des visites (Figure. 4) met clairement en évidence la présence d'outliers, visites dont la durée est anormalement élevées. Celles ci, au nombre de 23, se décomposent de la manière suivante : 19 visites de médecins dans la chambre 19 (adjacente à une salle utilisée par les médecins pour utiliser un ordinateur), 3 visites d'une infirmière dans la chambre 14 (adjacente à la pièce où étaient stockés les capteurs la nuit) et une visite d'infirmière dans la chambre 8 (adjacente à la salle de soin). L'existence de ces outliers pouvant être expliquée par la topographie du service ou l'oubli d'une blouse dans la pièce, nous considérons qu'ils ne constituent pas une erreur dans la reconstruction et proposons donc de les supprimer du jeu de donnée, en tant qu'aberrations dues à la configuration du lieu et à son usage.

6 Conclusion

En utilisant des données obtenues au travers d'audits et par l'utilisation d'un réseau de capteurs mobiles, nous avons mis en évidence le biais introduit par ce dernier, principalement dû à la fragilité du lien radio entraînant une perte de paquets. De manière à exploiter ces données, nous avons proposé un procédé de reconstruction simple et montré que celui ci permettait d'obtenir un signal à la fois cohérent avec la réalité et faisait apparaître des phénomènes absents des données brutes.

À présent que la problématique de la fiabilité de la mesure a été identifiée, un grand nombre d'extensions à ces travaux reste possible. Des approches purement analytiques pour déterminer l'impact de la perte de paquet sur le phénomène réel sont envisageables.

Références

[ASC+09] H. Alani, M. Szomsor, C. Cattuto, W. Van den Broeck, G. Correndo, and A. Barrat. Live social semantics. In *8th International Semantic Web Conference (ISWC2009)*, Washington DC, USA, October 2009.

[CBVdB] C. Cattuto, A. Barrat, and W. Van den Broeck. http ://www.sociopatterns.org/.

[FFF99] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM*, 1999.

[HCS+05] P. Hui, A. Chaintreau, J. Scott, R. Glass, J. Crowcroft, and C. Diot. Pocket switched networks and human mobility in conference environments. In *SIGCOMM*, Philadelphia, USA, August 2005. ACM.

[NBW06] Mark Newman, Albert-Laszlo Barabási, and Duncan J. Watts. *The Structure and Dynamics of Networks*. Princeton University Press, Princeton, USA, 2006.

[WS98] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393 :440–442, June 1998.