Inférence d'interactions sociales par colocalisation discrète

Roberto Pasqua, Matthieu Roy, Gilles Trédan

Inférence d'interactions sociales par colocalisation discrète

Roberto Pasqua[1,2], Matthieu Roy[1,2], Gilles Tredan[1,2]

[1]CNRS ; LAAS ; 7 avenue du colonel Roche, F-31400 Toulouse, France
[2]Univ de Toulouse ; LAAS ; F-31400 Toulouse, France
E-mail : {rpasqua,mroy,gtredan}@laas.fr

Keywords

Privacy, Indoor Human Mobility, Social Interaction, Co-Location.

Résumé

La liste des personnes avec qui nous passons du temps constitue assurément une donnée privée. Néanmoins, nous sommes de plus en plus nombreux à transporter des téléphones intelligents équipés de capacités de communication à courte portée (de l'ordre du mètre) permettant de détecter notre présence. Cette capacité de localisation permet de fournir, par exemple, des publicités ciblées (*Location Based Advertising*, voir *e.g.* [2]).
En déployant des balises dans l'environnement (Bluetooth par exemple), il est ainsi possible d'enregistrer la colocalisation de deux utilisateurs. Cet article explore l'utilisation d'un tel dispositif afin d'inférer (attaquer) la liste des interactions des individus ciblés.
Plus précisément, la question est la suivante : en supposant que l'on dispose de balises enregistrant la présence de tous les individus dans un rayon de 1m, combien faut-il déployer de telles balises pour inférer avec précision qui passe son temps avec qui ?
Après avoir défini le problème, nous présentons les résultats expérimentaux obtenus lors d'un cocktail ayant rassemblé 45 personnes dans une pièce de 10x10m. Ceux-ci montrent qu'il est possible d'obtenir, par exemple, environ 50% des contacts en déployant judicieusement moins de 20 balises.

1 Introduction

Le développement et la diffusion des technologies de micro-localisation permet aujourd'hui la collecte de vastes ensembles de données sur des phénomènes jusqu'alors inaccessibles, comme la mobilité et les interactions au sein de groupes d'individus. Ces données sont d'une extrême utilité pour comprendre les dynamiques qui règlent des comportements humains qui peuvent être objet de recherches dans plusieurs disciplines. Dans ces disciplines ne sont pas exclues celles qui utilisent les informations de notre vie privé pour aboutir à leur objectif, comme le marketing, le management ou l'espionnage (*e.g. NSA-Gate*). Si l'utilisateur n'est pas apte à vérifier et maîtriser ce flux d'informations sur sa position géographique, la protection de sa vie privée peut être mise en danger.
Dans le cas de la micro-localisation dans un périmètre restreint, de plus en plus d'applications web permettent de partager (ou rendre publique) la localisation des nos contacts sociaux proches. Ce mécanisme de colocalisation peut donc nous permettre d'avoir beaucoup plus d'informations, pas seulement sur la position géographique d'un groupe d'utilisateur (information locale), mais aussi sur leur réseau social qui contient forcement des informations qui n'ont rien à voir avec leur position géographique qu'ils ont partagée (information globale).
Ces nouvelles technologies peuvent être utilisées dans le domaines ses systèmes intelligents pour l'aide à l'achat [3], car des informations sur les produits en vente et les informations personnelles des potentiels acheteurs qui sont équipés des dispositifs *smart* seront utiles pour décider quelles informa-

tions publicitaires proposer aux clients et en quel moment.
Pour ce qui concerne la colocalisation, dans [4] il y a une analyse quantitative par rapport à son impact sur la confidentialité de la localisation géographique. Le mécanisme de colocalisation est étudié en correspondance de la popularité des *social networks* vu que ce mécanisme est de plus en plus utilisé par ses utilisateurs sans qu'ils se rendent compte de l'atteinte à la vie privé des potentiels attaquants.

Contexte et problématique

Population, contacts : Nous supposons qu'une population de N individus, constante, évolue dans un espace délimité (*e.g.* un aéroport, un centre commercial, un cocktail), pendant une durée T. Les individus composant cette population interagissent, et nous supposons que cette interaction r peut être mesurée par unités de temps : $r^t(i,j) = 1 \Leftrightarrow i$ et j sont en interaction au temps t. La somme des interactions s'étant déroulé dans l'espace peut-être représenté par une une matrice R avec

$$\forall i,j \in [1,N]^2, r_{i,j} = \sum_{t=1}^{T} r^t(i,j) = r_{j,i}.$$

Interaction : La notion d'interaction est très variable, et en général difficile à mesurer. Dans la suite de cet article, nous considérerons que deux individus sont en interaction s'ils passent du temps ensemble, formellement à une distance de moins d'1m : $r^t(i,j) = (d^t(i,j) < 1m)$. Cette valeur est difficilement mesurable, car elle nécessite d'avoir accès à la position de i et j avec une bonne précision.

Colocalisation : Nous supposons qu'il est possible de déployer dans l'espace un ensemble de K capteurs de colocalisation. Ces capteurs, d'une portée p, ont la capacité d'enregistrer la liste des individus à portée au temps t. L'implémentation d'un tel capteur peut par exemple être réalisée au moyen d'une Arduino équipée d'une pile Bluetooth et enregistrant les adresses mac détectées. Nous supposons que nous n'avons pas accès à la position précise du capteur. L'objectif étant de reconstruire la matrice d'interactions, nous nous intéresserons uniquement à la colocalisation : la présence simultanée de deux individus à portée du capteur ou non. Formellement chaque capteur k produit pour chaque individu i un vecteur $v_k^i(t) = 1 \Leftrightarrow d^t(i,k) < p$.

Bien que nous n'exploitons pas la position des capteurs, nous verrons que celle-ci a un impact : les capteurs déployés dans des zones peu peuplées apporteront peu d'information, les capteurs déployés dans des zones sur-peuplées colocaliseront beaucoup d'individus.
Dans la section suivante nous présenterons la méthodologie utilisée pendant notre travail en décrivant d'abord la procédure pratique pour enfin présenter nos résultats de simulation.

2 Méthodologie

Pour présenter nos idées et nos résultats nous utiliserons une collection des données qui contient les trajectoires de 45 utilisateurs dans un contexte *indoor*.

2.1 Inférence des contacts

Pour chaque utilisateur nous générons des vecteurs binaires qui indiquent leur présence ou leur absence dans le rayon de captation de chaque capteur. Pour chaque couple d'utilisateurs i,j et sur chaque capteur k nous avons calculé la similarité entre leurs échantillons binaires en utilisant l'indice de similarité de Jaccard [5]

$$J_k(i,j) = \frac{\|v_k^i \wedge v_k^j\|}{\|v_k^i \vee v_k^j\|}.$$

La figure 1 illustre l'intérêt d'utiliser l'indice de Jaccard au lieu d'un simple « et » logique : A et B ont passé 3 rondes ensemble dans la zone, tout comme A et C. Cependant, alors que la corrélation des trajectoires de A et B paraît bien présente, la colocalisation de A et C s'explique surtout par l'immobilité de C dans la zone. Le diviseur de Jaccard permet de corriger cet effet : $J(A,B) = \frac{3}{4} > J(A,C) = \frac{3}{7}$.

La somme des indices de similarité du couple i,j sur l'ensemble des capteurs nous donnera le niveau de contact entre les deux utilisateurs, *i.e.* l'élément $m_{i,j}$ de la matrice M des voisinages :

$$m_{i,j} = \sum_k J_k(i,j).$$

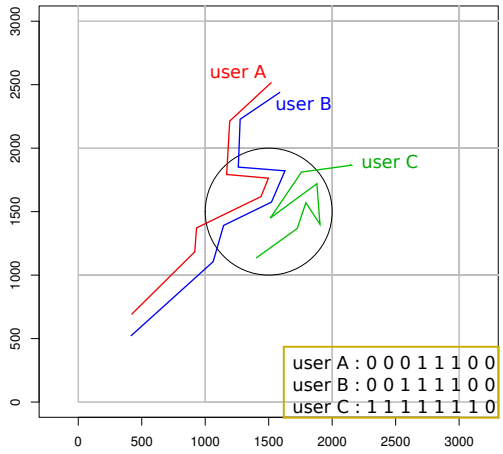

FIGURE 1 – L'indice de Jaccard favorise le couple (A, B) par rapport au couple (A, C)

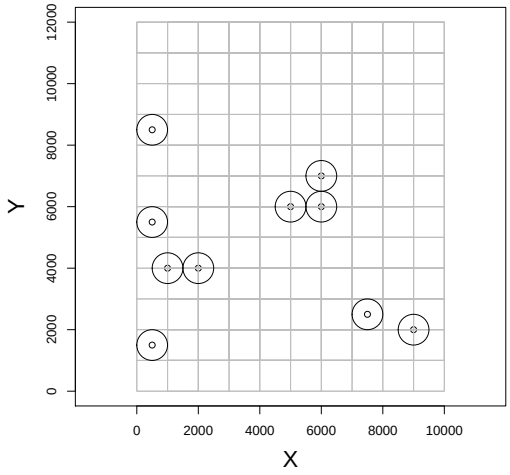

FIGURE 2 – Discrétisation de la pièce d'expérience

Une fois identifiée la matrice nous utiliserons le coefficient de corrélation des rangs de Spearman [5], défini par

$$cor = \frac{cov(r, m)}{s_r \, s_m}.$$

où r et m sont respectivement les variables des classements des contacts réels et estimés avec les capteurs tandis que s_r et s_m sont leurs écarts-type (r et m sont respectivement issus des matrices R et M). Le coefficient de corrélation sera l'indice de la distance entre l'analyse globale et l'analyse locale de l'information collectée.

2.2 Validation de l'approche

En connaissant les trajectoires des utilisateurs, nous avons simulé la présence de capteurs capables de détecter la présence d'un individu dans son rayon à partir d'un potentiel signal émis par un appareil portable apte à révéler l'identité de son propriétaire. La position des capteurs, de forme circulaire et rayon préfixé, est générée de façon aléatoire dans un ensemble de centres fini, suivant une discrétisation de l'espace de la pièce considérée, voir figure 2.

À partir des résultats de la captation, nous avons reconstruit les voisinages de chaque participant pour pouvoir vérifier si ils correspondent à leur liens sociaux réels.

2.3 Résultats

Nous considérons que le système de localisation est a *privacy* nulle, c'est-à-dire que toute l'information relative à la mobilité et l'interaction (les caractères, les comportements qui nous intéressent) est contenue dans les données en notre possession. En ayant toute l'information, pouvons-nous reconstruire cette information à partir d'une analyse locale du comportement en examen ? En termes pratiques, avec une technologie accessible et réaliste, que pouvons-nous inférer sur la vie privé des utilisateurs ?

Pour pouvoir capturer de façon efficace des informations sur la mobilité et l'interaction d'une foule humaine nous avons développé une plate-forme expérimentale qui nous a fourni une collection de données (*the ground truth*) pendant un buffet à l'occasion de l'inauguration d'un bâtiment [1].

A partir de l'analyse des trajectoires des participants à l'expérience nous avons construit la matrice R des voisinages (ou contacts) réels en considérant un rayon de 1m. En déployant ensuite des capteurs du même rayon en nombre variable et en position aléatoire nous avons généré la matrice M des contacts mesurés. La corrélation entre ces deux matri-

ces est le résultat des nos simulations. La figure 3 montre le résultat des simulations en déployant des capteurs sur la surface totale de la pièce d'expérience. Pour chaque étape il y a un nombre croissant de capteurs et chaque étape est répétée 20 fois. La ligne noire sur le tracé est l'allure moyenne de la corrélation.

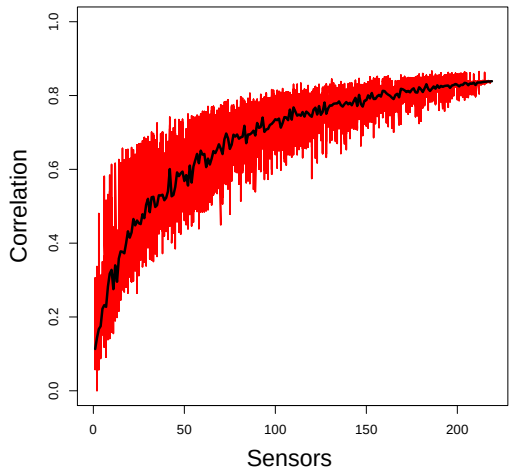

FIGURE 3 – Résultat sur la surface totale

La courbe résultante montre bien comment il est impossible d'avoir une corrélation maximale entre les deux mesures, car il y a toujours des contacts que ne sont pas pris en compte à cause de la nature locale et discrète de l'information collectée par captation (même si nous avons permis aux balises de se superposer pour une meilleur précision). En revanche les oscillations dans les valeurs de corrélation nous confirment que la distribution de densité dans la pièce n'est pas uniforme donc on peut améliorer l'estimation en ne considérant que les zones à forte densité.

Une fois identifiées les zones à forte densité (dues à la présence de points d'intérêt comme les comptoirs ou les tables) nous avons répété la simulation tirant aléatoirement les capteurs dans cette sous-zone à densité élevée d'utilisateurs, voire figure 4.

La figure 5 montre comment, en ayant une information supplémentaire pour pouvoir placer nos balises, on obtient un résultat comparable au précédent en terme de corrélation maximale avec moins de la moité des capteurs.

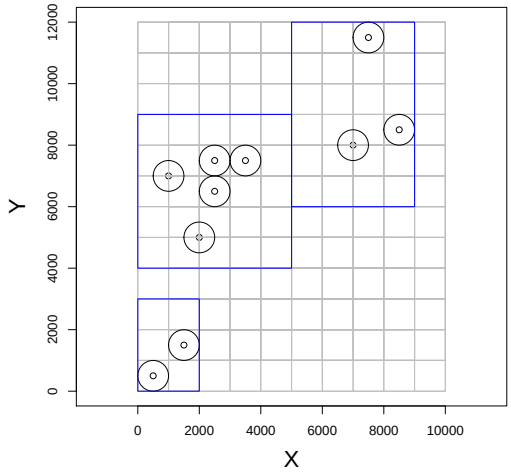

FIGURE 4 – Captation dans les zones à forte densité

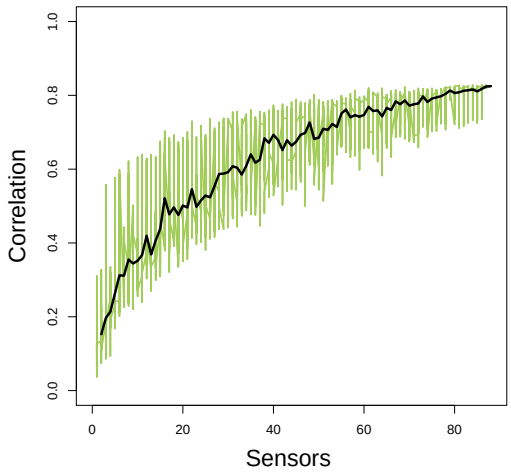

FIGURE 5 – Résultat sur les points d'intérêt

2.4 Analyse

En superposant les deux traces moyennes de corrélation nous pouvons mettre en évidence le résultat obtenu en plaçant nos balises dans des points de la salle à forte densité. Dans la figure 6 le point 1 nous montre que il est possible d'avoir une corrélation entre les contacts réels et les contacts estimés

du 52% avec 16 balises. Si on considère comme espace de mesure la surface totale de la pièce on aurait eu besoin de 30 balises pour avoir la même corrélation (point 3 sur la figure). En augmentant le nombre des capteurs, si on considère 55 points de mesure dans les zones à forte densité on a une corrélation du 75% (point 2 sur la figure) alors qu'il faut 108 capteurs dans l'ensemble de l'espace pour obtenir cette corrélation de 75% (point 4 sur la figure) La prise en considération d'une simple information, comme la présence des points d'intérêts, permet donc d'avoir des résultats d'estimation acceptables avec des ressources de captation réduites (environ moitié moins de capteurs).

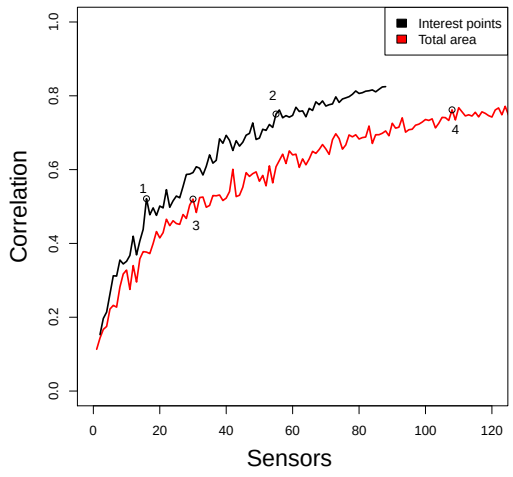

FIGURE 6 – Comparaison des résultat

Pour étudier encore les inférences sur la vie privée par colocalisation, notre analyse continue en montrant jusqu'à quel point il est possible de détecter les principaux contacts de chaque participant. Pour cette analyse nous considérons la matrice M correspondant à la simulation plaçant un maximum de capteurs dans les points d'intérêt. Pour chaque participant, nous avons trié en ordre décroissant les listes de contacts issues de R et de M, pour comparer les préfixes de ces listes. La figure 7 représente l'intersection moyenne de ces deux listes en faisant varier la taille du préfixe. Ainsi il est possible de voir combien de contacts réels on peut détecter en moyenne pour chaque utilisateur parmi ses plus proches contacts. Si on considère les 10 plus proches contacts mesurés on trouve en moyenne 6 des plus importants contacts réels.

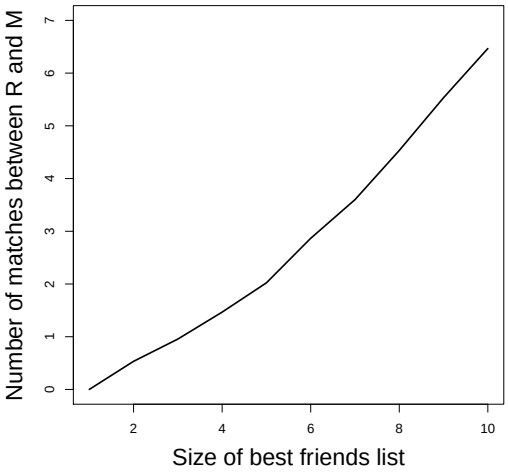

FIGURE 7 – Prévision des principaux contacts

3 Conclusions

Nous avons mis en évidence comment le mécanisme de colocalisation peut nous permettre de reconstruire les interactions sociales des individus qui partagent leur position dans un contexte de microlocalisation. On a aussi montré en quelle mesure il est possible (qualitativement avec la corrélation et quantitativement avec l'estime des meilleurs contacts) d'attaquer la *privacy* des individus dans une foule avec un système physique de captation légère. La prise en considération d'une information supplémentaire (la densité dans la pièce de l'expérience) nous a permis d'améliorer notre écart entre la réalité à reconstruire et la mesure menée par captation. Jusqu'à quel point l'analyse d'une information locale peut nous permettre de reconstruire l'information globale ?
Il est pertinent d'envisager donc, des nouveaux scénarios de simulations qui nous permettront de mieux comprendre les risques auxquels s'expose un utilisateur de certaines mécanismes de partage d'informations.

Références

[1] M.-O. Killijian, M. Roy, G. Trédan, and C. Zanon. Souk : social observation of human kinetics. In F. Mattern, S. Santini, J. F. Canny, M. Langheinrich, and J. Rekimoto, editors, *UbiComp*, pages 193–196. ACM, 2013.

[2] B. Kim, J.-Y. Ha, S. Lee, S. Kang, Y. Lee, Y. Rhee, L. Nachman, and J. Song. Adnext : A visit-pattern-aware mobile advertising system for urban commercial complexes. In *Proceedings of the 12th Workshop on Mobile Computing Systems and Applications*, HotMobile '11, pages 7–12, New York, NY, USA, 2011. ACM.

[3] S. Longo, E. Kovacs, J. Franke, and M. Martin. Enriching shopping experiences with pervasive displays and smart things. In F. Mattern, S. Santini, J. F. Canny, M. Langheinrich, and J. Rekimoto, editors, *UbiComp (Adjunct Publication)*, pages 991–998. ACM, 2013.

[4] A. M. Olteanu, K. Huguenin, R. Shokri, and J.-P. Hubaux. Quantifying the Effect of Co-Location Information on Location Privacy. Technical report, 2014.

[5] G. Saporta. *Probabilités Analyse des Données et Statistique*. Paris, 1990.