Inférence et modèles de données personnelles : mobilité sociale, proximité spatiale

Roberto Pasqua

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *17/11/2016* **par :**

ROBERTO PASQUA

Inférence et modèles de données personnelles : mobilité sociale, proximité spatiale

---

JURY

ROBERTO BALDONI	Professore ordinario	Rapporteur
CLÉMENCE MAGNIEN	Directeur de recherche	Rapporteur
JOSIANE MOTHE	Professeur des universités	Examinateur
SONIA BEN MOKHTAR	Chargé de recherche	Examinateur
MOHAMED KAÂNICHE	Directeur de recherche	Directeur de thèse
GILLES TREDAN	Chargé de recherche	Directeur de thèse
MATTHIEU ROY	Chargé de recherche	Invité

École doctorale et spécialité :
EDSYS : Informatique 4200018

Unité de Recherche :
LAAS-CNRS (UPR 8001)

Remerciements

Je remercie premièrement Roberto Baldoni et Clémence Magnien pour avoir accepté de rapporter sur ce manuscrit, fruit de mon travail de thèse. Je suis également reconnaissant à Josiane Mothe et Sonia Ben Mokhtar qui siègent dans le jury de thèse en tant qu'examinateurs. J'aimerais aussi exprimer ma gratitude envers Matthieu Roy, Gilles Tredan et Mohamed Kaâniche qui ont dirigé ces travaux. En particulier, leur ténacité et leur passion pour la recherche ont fait en sorte de me conduire au bout de ce chemin, prêt à continuer sur ma voie dans une carrière académique. Je garderai dans mon bagage le souvenir de leurs conseils, leurs indications et leurs remarques. Merci de m'avoir écouter, j'ai appris beaucoup en vous écoutant.

Parmi tous les personnes qui ont travaillé avec moi pendant ces trois années, j'ai une mention spéciale pour Jesus Friginal et Kévin Huguenin, toujours disponibles pour partager leurs temps et leurs connaissances. Dans la même catégorie, je remercie tous les membres de l'équipe TSF pendant ces derniers trois ans, surtout le sous-ensemble qui a suivi ou, au moins, essayé de suivre, quotidiennement, mes monologues à table ou autour de la machine à café. Moussa, Benoit, Carla, Joris, Mathilde et Pierre pourront se reconnaitre dans ce groupe. En dehors du couloir TSF, j'ai une pensée pour toute les rencontres que j'ai faites au LAAS et à l'INSA et que j'espère renouveler encore.

Je remercie ma famille et mes amis pour n'avoir pas encore bien compris ce que je fais dans mes recherches, ça sera toujours un plaisir de leur expliquer à nouveau. Leur support est pour moi fondamental.

Tous les jours à mes côtés, Céline et Étienne ont une place spéciale dans cette page.

Table des matières

Table des figures

Liste des tableaux

Introduction

La diffusion massive des dispositifs portables, de plus en plus utilisés pour le traitement et la communication de l'information, permet la collecte d'importantes masses de données liées à l'activité des utilisateurs sur des applications mobiles. Les données proviennent des systèmes de captation embarqués, *i.e.* des capteurs de mouvement comme le gyromètre et l'accéléromètre, des capteurs de localisation comme le récepteur GPS ou des systèmes d'interface de communication comme le Wi-Fi et le Bluetooth. Ces données sont souvent nécessaires pour l'obtention d'un service particulier, par exemple, la navigation routière ou le monitorage de l'activité sportive. L'explosion de la demande pour ces services et de la capacité des dispositifs portables à produire de l'information engendre de gros volumes de données (*datasets*). Conjointement, ces données sont de plus en plus partagées, dans certains cas le partage est nécessaire pour l'obtention du service requis, tandis que dans d'autres cas le partage est discutable. Cette dynamique de partage permet aux fournisseurs de services de collecter, entre autres, les données provenant des capteurs embarqués.
Les données produites dans ces systèmes ont différentes formes et contenus, ils ouvrent la voie à de nouvelles recherches dans une multitude de disciplines. En outre, une approche interdisciplinaire est possible sur des données qui relèvent de plus en plus d'informations sur le comportement humain au quotidien.
Nous allons nous intéresser aux données de localisation et de proximité, c'est-à-dire les traces de mobilité, qui sont issues des systèmes mobiles formés par un groupe d'utilisateurs. Les traces de mobilité contiennent des informations concernant le mouvement géographique ainsi que la dynamique temporelle dans le déplacement et le réseau social des utilisateurs. Les enjeux économiques, sociétaux et scientifiques des données de mobilité sont prouvés et permettent l'utilisation de ces données dans différentes applications. Nous nous intéressons à l'étude des données de mobilité dans le développement des systèmes de communication mobiles et dans l'impact que ces données ont sur la protection de la vie privée des utilisateurs. En conséquence, les données de mobilité produites par les utilisateurs à l'intérieur d'un système mobile sont étudiées suivant deux axes :

- l'utilisation des modèles de mobilité est à la base du développement d'algorithmes de communication dédiés aux systèmes mobiles. Les traces de mobilité réelles vont nous permettre de comparer les traces de mobilité synthétiques utilisées dans la simulation avec la réalité qu'ils sont censés décrire.
- la manipulation des traces de mobilité réelles implique une réflexion sur les conséquences que les informations extraites de ces données ont, relativement à la protection de la vie privée des utilisateurs.

La caractérisation des réseaux de communication basés sur des systèmes mobiles conduit naturellement à l'utilisation de certains modèles de mobilité capables de générer des traces de mobilité assimilables à des agents mobiles. La paramétrisation de ces modèles est un aspect crucial dans la réussite de la simulation et elle varie selon la définition des différents modèles. La possibilité d'analyser des traces de mobilité réelles issues d'une collecte à haute precision,

où la granularité de captation est inférieure à la portée de communication envisagée dans la simulation, nous permet une fine estimation des paramètres nécessaires. Nous allons pouvoir donc vérifier le niveau de confiance des modèles de mobilité dans la génération des traces synthétiques. L'analyse des propriétés sociales et spatiales des traces réelles et synthétique mettra en évidence l'absence de la prise en considération de la "mobilité sociale" de la part des modèles.

Les informations contenues dans les traces de mobilité peuvent avoir un impact sur la protection de la vie privée des utilisateurs. La possibilité de collecter l'ensemble des traces dans une foule d'utilisateurs à un moment particulier et le traitement qui peut en dériver peuvent générer des pertes dans la protection de la vie privée par le biais d'inférences révélant des comportements locaux (concernant un utilisateur particulier) ou globaux (concernant l'ensemble de la foule) des utilisateurs. L'inférence, et en conséquence la protection, des informations personnelles à partir des données de localisation fait déjà l'objet de récentes études. Nous montrons qui il est possible de mener des inférences à partir des informations de mobilité des utilisateurs sans besoin de leur localisation. Le distinguo entre les informations de localisation et celles de "proximité spatiale", contenues dans des traces de mobilité, est au centre de la définition de la co-localisation dans notre approche. Le concept de co-localisation nous permet aussi de quantifier l'impact des informations probabilistes concernant la mobilité humaine sur l'inférence de données de localisation.

Dans le premier chapitre de cette thèse nous allons présenter l'état de l'art des études concernant les deux axes présentés dans cette introduction. Dans le deuxième chapitre, nous proposons une analyse fine des données de mobilité issues d'une série d'expérimentations (expériences de collecte) afin d'évaluer l'écart entre les modèles de mobilité et la réalité, dans le cadre de l'évaluation d'algorithmes de communication pour systèmes mobiles. Le troisième chapitre présentera un scénario inédit d'attaque sur des données de mobilité en utilisant le concept de co-localisation indépendamment de la localisation des utilisateurs. Le concept de co-localisation sera au centre du quatrième chapitre où on évaluera la possibilité d'une attaque par inférence sur la localisation à partir d'informations probabilistes concernant la co-localisation. En conclusion, nous présentons le bilan de nos contributions en proposant des perspectives futures aux travaux présentés dans cette thèse.

Chapitre 1

État de l'art

Sommaire

Ce chapitre présente l'état de l'art des travaux en relation avec notre étude des systèmes mobiles. Nous définissons les systèmes mobiles comme des dispositifs de traitement et de communication de l'information non géographiquement statiques, munis d'interfaces de communication diverses. On peut donc parler de systèmes mobiles comme des systèmes cyber-physiques où des entités de traitement et de communication de l'information (*i.e.*, des systèmes informatiques) interagissant avec des phénomènes physiques, dans notre cas spécifique nous considérerons la mobilité humaine. L'avènement des dispositifs portables *smart* et des objets connectés (*i.e.* internet des objets), permet la collecte de masses de données, provenant de ces dispositifs, qui caractérisent des phénomènes de la réalité jusqu'alors quantitativement inobservables. Dans le cas de notre étude, nous nous intéressons à l'impact que l'utilisation des systèmes mobiles, en terme des données de mobilité produites par ces mêmes dispositifs, a :

1. dans la *caractérisation des Mobile Ad hoc NETworks (MANETs)* à travers l'analyse des modèles de mobilité nécessaires pour le développement d'algorithmes de communication efficaces ;
2. dans la *protection de la vie privée (privacy)* des utilisateurs à travers l'exploration de moyens d'inférence issus de la co-localisation des utilisateurs.

Après avoir introduit les généralités des réseaux *Ad Hoc*, notre illustration de l'état actuel des travaux sera organisée selon trois axes :

– La *collecte des données*, avec la présentation des différentes jeux des données de mobilité, en particulier dans l'étude des réseaux opportunistes (*Opportunistic Networks*).

– La présentation des *modèles de mobilité* utilisés dans la caractérisation des réseaux mobiles.
– L'introduction des enjeux concernant la protection de la vie privée des utilisateurs et l'influence des données geo-localisées dans *l'inférence par co-localisation*.

1.1 Introduction aux réseaux Ad Hoc

Une vaste littérature concerne les réseaux *Ad Hoc*, deux références significatives [Per08] [Toh01] présentant un aperçu des définitions et problématiques abordées dans ce paragraphe. *Ad Hoc* est une locution latine traduisible comme "à cet effet" et qui, appliqué à une chose, indique que cette chose est "adaptée à tel usage précis". Dans le cas des réseaux de communication, on définit des réseaux Ad Hoc comme un ensemble de dispositifs capables de communiquer à un instant (un moment, une circonstance, une situation, un contexte) défini, sans infrastructure physique globale (par exemple, internet). L'absence d'infrastructure globale rend le réseau adaptatif envers les nouveaux scénario d'utilisation qui se profilent avec la diffusion massive des dispositifs portables et objets connectés. La possibilité d'une infrastructure dynamiquement configurable permet d'optimiser la communication des informations qui ont des supports divers, par exemple dans le cas des dispositifs portables ce sont les utilisateurs eux mêmes qui forment le support de communication de l'information. Ce sont donc des systèmes distribués de communication. Chaque dispositif peut communiquer directement avec tous autres dispositifs à sa portée et, plus généralement avec tout dispositif dans le système si il existe un ou plusieurs dispositifs relais entre lui et le dispositif destinataire. Ces systèmes adaptatifs peuvent être fixes (*wired*, par exemple un système des capteurs environnementaux chargés de la mesure d'un certain phénomène) ou mobiles (*wireless*, par exemple un système des véhicules autonomes chargés d'une mission particulière), composés par des dispositifs hétérogènes ou homogènes. La nature des dispositifs composant le réseau peut avoir un impact important sur les performances du réseau même (capacité de calcul, portée et délai de communication, capacité énergétique, etc.). Nous allons nous focaliser sur les réseaux Ad Hoc mobiles.
Les réseaux mobiles Ad Hoc sont généralement nommés MANET [MC98], acronyme de *Mobile Ad hoc NETwork*. Différentes communautés scientifiques ont étudié les réseaux mobiles Ad Hoc. Les différentes perspectives apportées par ces communautés ont conduit à la définition de plusieurs architectures. Les définitions que nous avons rencontrées pendant notre étude sont les suivantes :

– Les reseaux tolerants aux délais (*Delay Tolerant Networks*, DTN) [Fal03], architecture fortement asynchrone pour la tolérance des pertes des connexions dans un réseau Ad Hoc mobile.
– Les reseaux opportunistes (*OPPortunistic NETworks*, OPPNET) [PPC06], évolution du concept de MANET qui permet à chaque nœud de pouvoir participer à la communication sans avoir connaissance de la topologie du réseau.
– Le reseaux AdHoc mobiles Mesh (*Mobile Mesh Ad-Hoc Networking*, MMAN) [Nag+05], réseau hybride qui utilise les propriétés des réseaux Ad Hoc conjointement aux infras-

tructures des réseaux fixes et mobiles traditionnels.
- Les reseaux Switches de poche (*Pocket Switched Networks*, PSN) [Hui+05], réseaux opportunistes basés sur l'utilisation des dispositifs "de poche".
- Le reseaux sociaux basés sur la proximité (*Proximity based Social Networking*, PSN) [Dob14], communication à courte portée basée sur les réseaux sociaux des utilisateurs.

La différence dans les définitions des diverses architectures n'implique pas forcément des différences substantielles dans leur fonctionnement (interface). Certaines problématiques dans la mise en œuvre de ces systèmes sont donc partagés entre les différentes architectures. Dans la caractérisation des réseaux mobiles Ad Hoc et dans le développement des protocoles de communications dédiés à ces réseaux, il est fondamental de pouvoir simuler le comportement des nœuds dans différents contextes d'utilisation et de mobilité. À ce sujet, des scénarios de simulation peuvent être générés à partir des traces de mobilité synthétiques (*i.e.* issues des modèles de mobilité) ou à partir des traces de mobilité réelles (*i.e.* contenues dans des collections de données de mobilité capturées sur le terrain).

1.2 Collections de données de mobilité

Les données de mobilité sont de plus en plus utilisées dans de nombreuses disciplines, des sciences sociales à l'informatique théorique, en passant par l'épidémiologie [Sal+10] et l'informatique décisionnelle [Ant+12]. Il est possible de collecter des données de mobilité à partir de différentes situations (scénarios, contextes, cas réels) et par le biais de multiples dispositifs. Dans cette section nous allons expliciter comment certains jeux de données sont construits et comment ils sont utilisés dans la caractérisation des réseaux Ad Hoc [Cha+07b] et dans la quantification de leurs impacts sur la vie privée [BSM10].

1.2.1 Collecte des données

Depuis la massification de l'utilisation de dispositifs portables équipés avec différentes interfaces de communication, de plus en plus des données de mobilité ont été collectées afin de créer des collections exploitables pour des exigences distinctes. On peut classer deux catégories d'informations concernant les traces de mobilité :

1. Les informations de localisation, c'est-à-dire la succession des positions géographiques qui déterminent une trace de mobilité. Les positions seront relatives à un système de coordonnées prédéfini.
2. Les information de proximité de dispositif à dispositif, qui nous permettent d'avoir une information sur la position relative entre les utilisateurs des dispositifs concernés.

Les informations susmentionnées dépendent de l'interface de communication (la technologie) utilisée pour leur génération. Dans le cas des informations de localisation, les technologies majoritairement utilisées dans le cas *outdoor* sont le *Global Positioning System* (GPS) comme

dans [Rhe+09] et le *Global System for Mobile Communications* (GSM) comme dans [EPL09a]. Pour ce qui concerne la localisation *indoor*, la technologie Wi-Fi (IEEE 802.11) est souvent utilisée pour inférer la position des dispositifs par rapport aux points d'accès (*access points*, AP) présents, en mesurant la puissance du signal de transmission (*received signal strength indicator* RSSI) ou la trilatération entre differents points d'accès [Bil+13], [MV05] et [CKB14]. Pour collecter des données de proximité, de dispositif à dispositif, les capacités du standard Bluetooth répondent bien au problème. La possibilité de pouvoir détecter des dispositifs dans un rayon plus ou moins étendu permet d'enregistrer la proximité d'un dispositif par rapport à un autre ce qui, dans le cas d'un dispositif fixe dont on connait la position, nous permettra d'inférer sa position. Le Bluetooth peut être intégré dans les téléphones portables comme dans [EPL09a] ou dans des dispositifs dediés comme dans [Hui+05] (iMote, Intel).
La collecte de données de mobilité est donc aussi possible à partir de dispositifs dédiés comme des *open beacon* (basés sur le RFID) [Cat+10] et [Ise+11], des *pocket trace recorders* (basés sur un module radio AUREL) [GPR09a] et [GPR09b], et des *sociometric badges* (basés sur le standard Zigbee) [Wu+08] et [Don+12].

1.2.2 Études sur les données

Il existe aujourd'hui plusieurs projets de recueil de jeux de données dédiés à leurs centralisation, classification et publication. Deux de ces projets ont attiré notre attention, notamment le projet CROWDAD [Cro] (Dartmouth College) et le Stanford Network Analysis Platform (SNAP) [LK16]. Nous allons détailler ici différents travaux basés sur des données de mobilité qui sont publiés dans les portails susmentionnés.

Reality Mining Le premier de ces travaux est une référence incontournable dans les études qui considèrent la mobilité humaine comme facteur déterminant dans l'analyse du réseau social (les contacts sociaux) des utilisateurs [Too+15], du développement de protocoles de communication pour des réseaux opportunistes [Cha+07a] aux études sur les inférences possibles à partir des données de mobilité [PM09]. On le retrouve donc transversalement par rapport à notre approche. Le projet était nommé *Reality Mining* [EP06] et mené entre Septembre 2004 et Juin 2005 au MIT Media Laboratory ; il a permis de collecter des données de mobilité de 109 utilisateurs à partir de téléphones portables dédiés capables d'enregistrer toute leur activité (l'état du téléphone, les comptes rendus des appels et des messages textuels, etc.). Parmi les données collectées, il y a des données concernant la mobilité des utilisateurs. Comme precisé dans le paragraphe précedent, les données de mobilité contenues dans cette collection concernent la localisation (*i.e.* les identifiants uniques des antennes GSM) et la proximité (*i.e.* les listes des dispositifs à portée Bluetooth) des utilisateurs. En complément, les données contiennnent aussi les résultats d'une enquête menée parmi les utilisateurs sur leurs comportements vis-à-vis de leurs amitiés (déclaration des listes d'amis, estimation spatiale et temporelle de ces amitiés). Dans [EPL09a], les premiers résultats liés à ce projet sont publiés. Notamment, les auteurs démontrent comment il est possible d'inférer 95% des amitiés entre les utilisateurs participant à l'expérience.

Infocom Une autre collection de données bien étudiée dans la communauté scientifique est issue du projet Haggle et disponible dans [Sco+09]. Elle contient les contacts Bluetooth des utilisateurs utilisant des dispositifs iMote (Intel) dans différents contextes. Ces dispositifs compacts ($\sim 45\,\text{mm} \times 35\,\text{mm}$) étaient capables d'enregistrer périodiquement les adresses MAC des autres dispositifs à portée. L'expérimentation pendant la conférence Infocom 2005 est décrite dans [Hui+05]. Dans cette expérimentation 54 dispositifs (dont finalement seulement 41 ont donnés des résultats exploitables) étaient distribués aux participants de la conférence. Les auteurs ont ainsi enregistré $\sim$ 28000 contacts (dont 80% entre deux dispositifs iMote et les 20% restant entre un iMote et un autre dispositif Bluetooth). À partir des ce données de mobilité, ils ont pu analyser les propriétés temporelles du réseau créé par les contacts entre les utilisateurs (c'est dans ce travail que les auteurs définissent les *Pocket Switched Networks*) pour pouvoir étudier l'impact que ces propriétés ont dans la conception des algorithmes de routage dédiés à ce type de réseau. Concernant la caractérisation des réseaux opportunistes, on retrouve ces données dans [Cha+07a] où les auteurs étudient l'influence que la mobilité humaine a dans la définition et l'évaluation des algorithmes de routage.

Roller Net Dans un contexte plus dynamique et inhabituel, on trouve le projet Roller Net [BL09]. Comme dans le cas de Haggle, la mobilité des utilisateurs est capturée en utilisant des modules Intel iMote qui utilisent la technologie Bluetooth. La particularité est que le contexte de collecte est une balade en patin en ligne au cœur de Paris. En conséquence la densité et la mobilité des utilisateurs est supérieure à celle des autres scénarios présentés précédemment. L'expérience de collecte et une première analyse fine des traces de mobilité collectées sont contenues dans [Tou+11]. Les auteurs montrent comment l'effet accordéon relevé dans les traces de mobilité impacte les performances des algorithmes de diffusion épidémique (*epidemic routing*).

Réseaux sociaux géo-localisés Un autre exemple de données de mobilité sont les données issues des *location-based social networks*. Le réseautage social géolocalisé permet aux utilisateurs de partager leur position géographique dans un réseau social. Dans [CML11], deux de ces jeux de données sont utilisés pour la proposition d'un nouveau modèle de mobilité. En particulier, les informations de mobilité couplées avec les informations sur les réseaux sociaux des utilisateurs permettent une analyse de l'influence que la structure du réseau social a sur la mobilité humaine. À partir des résultats de cette analyse, les auteurs proposent un modèle de mobilité prédictif capable d'améliorer les performances des trois modèles de comparaison (notamment un basé sur le *Random Walk* et un basé sur le *Gaussian Model*).

1.3 Modèles de mobilité

Dans le développement et l'évaluation par simulation des algorithmes de communication il est nécessaire de reproduire les mouvements des utilisateurs de façon réaliste. Plusieurs modèles de mobilité sont utilisés pour cela. Dans cette section nous allons présenter sept des

modèles les plus utilisés. En suivant la classification faite dans [CBD02], nous allons distinguer les modèles de mobilité individuelle, où chaque nœud se déplace indépendamment des autres, des modèles de mobilité de groupe où les nœuds sont organisés en groupes et où les mobilités des nœuds d'un même groupe ne sont pas indépendantes (le déplacement des nœuds dépend de la dynamique de déplacement de leur groupe d'appartenance). Chaque modèle sera présenté, en partant de sa référence bibliographique originale, par une description exhaustive de son comportement et l'image de la trajectoire d'un agent mobile pour 150 unités de temps (*step*).

1.3.1 Mobilité individuelle

Les modèles de mobilité individuelle sont ceux qui considèrent les déplacements de chaque nœud indépendamment des autres nœuds présents dans le système. Par nœud nous entendons des agents mobiles, c'est-à-dire les entités mobiles qui composent notre système. Dans notre cas d'étude les nœuds mobiles seront considérés comme des êtres humains. Nous allons décrire synthétiquement les caractéristiques de cinq de ces modèles.

Random Walk Vu son importance dans la communauté scientifique, il est bien de rappeler ici que le nom *Random Walk* apparaît pour la première fois dans [Pea05] grâce à Karl Pearson. Il est la représentation en terme de mobilité sur deux dimensions du mouvement d'une particule suspendue dans un liquide (mouvement Brownien, formalisé par Albert Einstein dans [Ein05]). C'est le modèle le plus simple pour décrire un mouvement aléatoire, il est donc utilisé pour décrire les mouvements de divers agents dans des domaines scientifiques variés. De l'économie [Coc88] à la biologie [BB88], en passant pour la psychologie [NP97], on retrouve l'utilisation de ce modèle dans la représentation d'un phénomène aléatoire évoluant par étapes (*steps*). Dans son implémentation pour la simulation d'agents mobiles, chaque nœud choisit une vitesse et une direction aléatoirement, respectivement dans les intervalles $[v_{min}, v_{max}]$ et $[0, 2\pi]$. Un nouveau choix est fait chaque fois qu'un temps t ou une distance d constante est atteint.

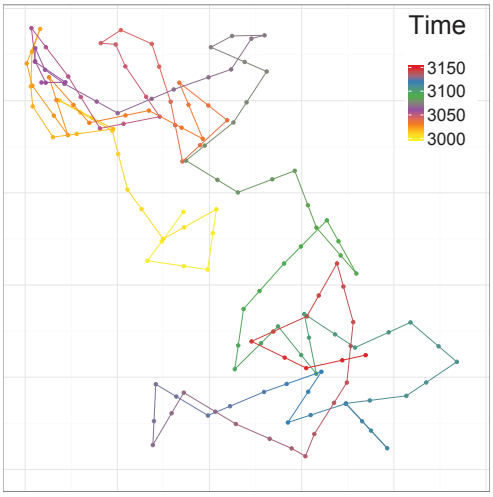


FIGURE 1.1 – Une trace issue du modèle de mobilité Random Walk

Random Waypoint [JM96] La formalisation du modèle est contenue dans [Bro+98]. Il est très utilisé dans la caractérisation des réseaux Ad Hoc, souvent avec des variations par rapport à la définition initiale. Un travail remarquable pour la mise au point dans l'utilisation de ce modèle pour la simulation des protocoles de routage dans des réseaux Ad Hoc est [YLN03]. La description du modèle est simple : chaque nœud est à l'arrêt pour un certain temps wt (*waiting time*) et, après avoir choisi un point de destination dans l'espace de simulation, il va se déplacer avec une vitesse uniformément distribuée dans l'intervalle $[v_{min}, v_{max}]$ (dans certaines implémentations v_{min} est fixé à 0). Après une nouvelle pause, une nouvelle destination et une nouvelle vitesse seront sélectionnées pour chaque nœud.

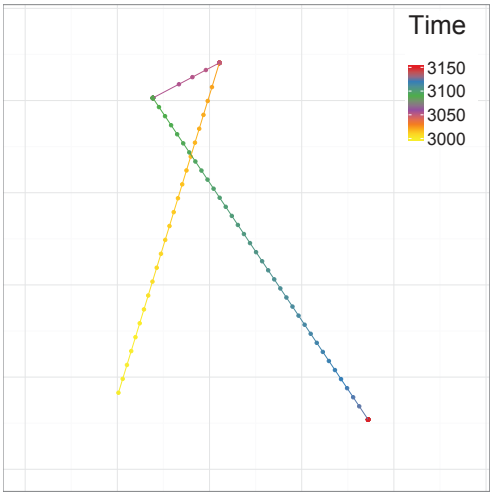


FIGURE 1.2 – Une trace issue du modèle de mobilité Random Waypoint

Random Direction Ce modèle [RMSM01] est défini pour résoudre le problème de la variabilité de la taille du voisinage émergeant pendant les simulations qui utilisaient le Random Waypoint. L'idée est de sélectionner une direction plutôt qu'un point de destination, c'est-à-dire qu'après le temps wt de pause initiale, une direction entre $[0, 2\pi]$ est choisie et une vitesse comprise entre $[v_{min}, v_{max}]$ sélectionnée. Quand le nœud arrivera à la limite de la zone de simulation (bord) et après une nouvelle pause, il choisira une nouvelle direction entre $[0, \pi]$ et il atteindra à nouveau le bord (dans la définition initiale il n'est pas envisageable de sortir d'un bord pour "réapparaître" dans le bord opposé, au contraire de certaines implémentations où il est possible de choisir comment gérer la collision avec le bord).

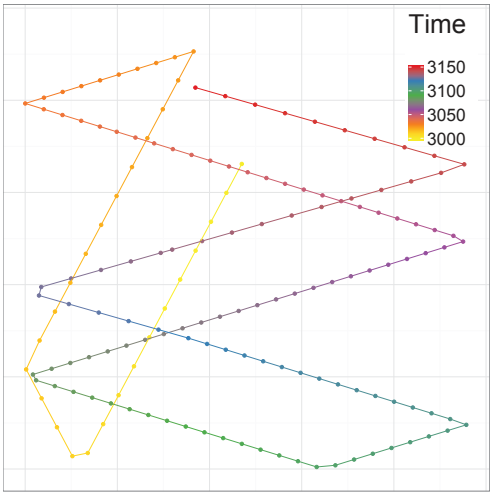


FIGURE 1.3 – Une trace issue du modèle de mobilité Random Direction

Truncated Lévy Walk Ce modèle [Rhe+11] est défini comme un modèle Random Walk qui utilise une distribution tronquée de Pareto ([Arn15] et [New05]) pour représenter le temps d'attente wt (*waiting time*) et la distance franchissable fl (*flight length*). Ce modèle est défini pour prendre en considération la distribution à queue lourde qui caractérise le rayon d'action (*flight length*, fl) dans la mobilité humaine. En particulier, les auteurs affirment qu'en utilisant ce modèle ils ont été capables de retrouver la même distribution concernant le temps d'inter-contact (le temps entre deux contacts pour un couple de nœuds) observés dans diverses études précédemment menées sur la mobilité humaine, études basées sur des données de mobilité réelles ([Cha+07a] et [MV05]).

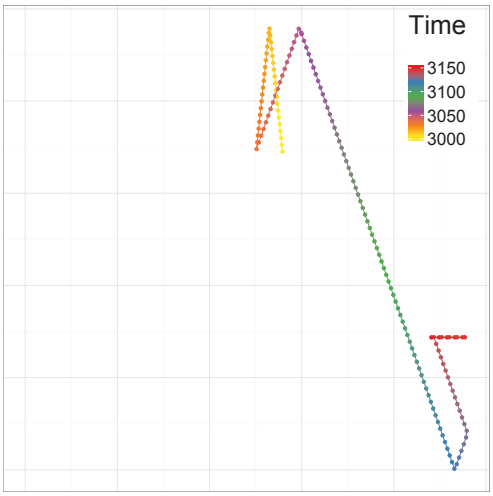


FIGURE 1.4 – Une trace issue du modèle de mobilité Truncated Lévy Walk

Gauss-Markov Le modèle Gauss-Markov [LH99] est conçu avec l'objectif de diminuer l'effet sans mémoire des modèles basés sur le Random Walk. Par "effet sans mémoire" on entend la complète indépendance du choix de mouvement au temps t par rapport au mouvement au temps $t-1$ qui est comprise dans la définition du Random Walk et des modèles dérivés. Le modèle Gauss-Markov impose qu'à chaque temps t et pour chaque agent mobile, la mise à jour de la position et de la vitesse va dépendre de la position et de la vitesse au temps $t-1$. Pour pouvoir contrôler l'entropie dans la mise à jour des paramètres de mobilité, un coefficient $0 \leq \alpha \leq 1$ est introduit dans le modèle. Pour $\alpha = 0$ la mobilité sera Brownienne (complètement aléatoire), tandis que pour $\alpha = 1$ la mobilité sera complètement linéaire.

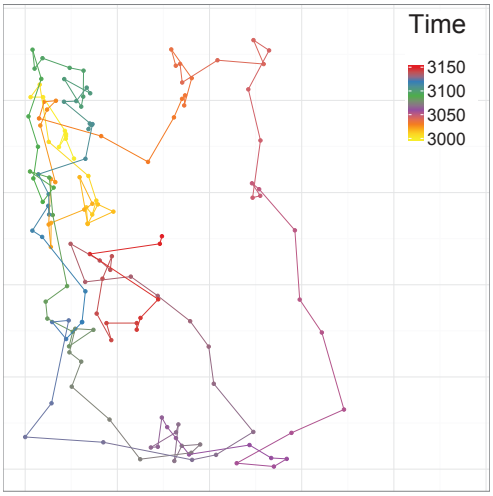


FIGURE 1.5 – Une trace issue du modèle de mobilité Gauss-Markov

1.3.2 Mobilité de groupe

Les modèles de mobilité de groupe considèrent le déplacement d'un groupe d'agents, *i.e.* le mouvement d'un agent dépendra du mouvement des autres agents faisant partie de son groupe. Nous allons décrire synthétiquement les définitions de deux de ces modèles.

Time-Variant Community Ce modèle [Hsu+07] est le premier à prendre en considération la non homogénéité des caractéristiques temporelles et spatiales des agents mobiles. Chaque agent est assigné à une communauté, chaque communauté a une position de référence (une aire spatiale à l'intérieur de l'espace de simulation). Le temps de simulation du modèle est partagé alternativement entre des périodes de mouvement normal (*normal mouvement periods*, NMP) et des périodes de mouvement de concentration (*concentration mouvement periods*, CMP). Pendant le temps de mouvement normal, les agents peuvent se déplacer librement dans l'espace de simulation tandis que pendant le temps de mouvement de concentration, les agents peuvent rejoindre leur communauté avec une certaine probabilité. Dans chaque période de temps (*epoch*), les agents peuvent avoir deux comportements, un comportement local (*local epoch*) et un comportement itinérant (*roaming epoch*). Le comportement local oblige les agents à une mobilité autour de leur communauté alors que le comportement itinérant les obligera à une mobilité globale dans la surface d'expérimentation. Au changement de comportement les agents choisissent uniformément une vitesse entre $[v_{min}, v_{max}]$ et une direction entre $[0, 2\pi]$, *i.e.* ils utilisent une mobilité de type Random Direction locale.

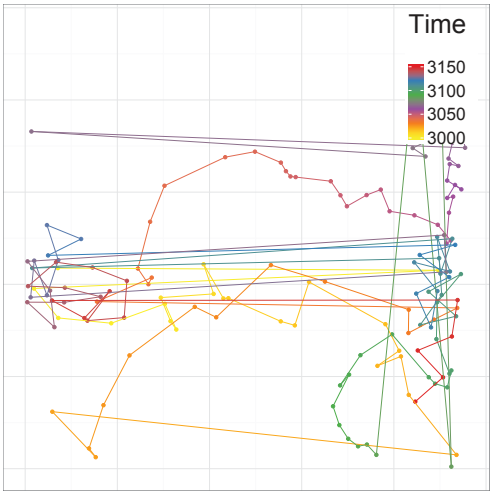


FIGURE 1.6 – Une trace issue du modèle de mobilité Time-Variant Community

Reference Point Group Dans ce modèle [Hon+99] les agents dans le système sont divisés en groupes. Chaque groupe a un point de référence (centre du groupe) et les agents appartenant au groupe sont uniformément distribués autour du point de référence. Le point de référence du groupe utilise un modèle aléatoire de déplacement (Random Walk) et les agents utilisent le même modèle pour se déplacer autour du point de référence. Les trajectoires de déplacement des points de référence des groupes conditionneront les trajectoires des agents dans le groupe.

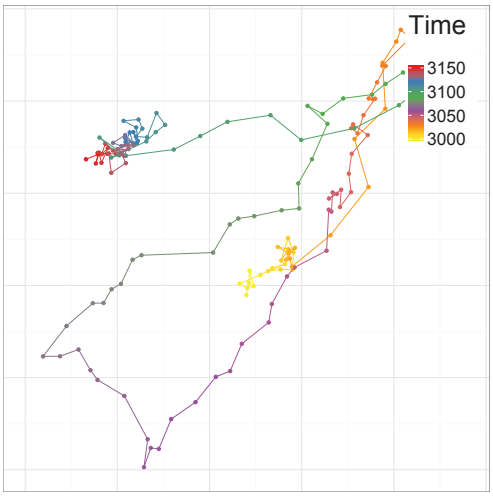


FIGURE 1.7 – Une trace issue du modèle de mobilité Reference Point Group

1.4 Impact des données sur la vie privée

Les données de mobilité (les traces de mobilité) ont un impact important sur la vie privée des utilisateurs. À partir des données de localisation et de proximité, il est possible d'inférer des informations de plus haut niveau que celles concernant la mobilité des utilisateurs. Dans l'étude des services basés sur la géo-localisation (*location-based services* (LBS)) et des mecanismes de protection envisageables, plusieurs résultats marquants ont été produits dans les dernieres lustres [BS03], [Sho+11], [BR15].
Nous allons nous intéresser aux informations de co-localisation, c'est-à-dire la possibilité de localiser simultanément au même endroit deux ou plusieurs utilisateurs. Les informations de co-localisation peuvent être *explicites* (déterministes) où *implicites* (non déterministes, probabilistes). Ces informations peuvent être déduites à partir des données de localisation ou de proximité, on étudiera leurs impacts sur la protection de la vie privée par des attaques d'inférence dans les chapitres 3 et 4.
L'impact des informations de co-localisation est l'objet de [Cra+10] où les auteurs utilisent des données provenant de photographies géo-localisées pour inférer les liens sociaux entre les utilisateurs. Les métadonnées des photographies, publiées sur un réseau social dédié, peuvent en effet contenir les coordonnées géographiques déclarées par les utilisateurs ou provenant directement de la caméra équipée d'un capteur GPS. À partir des localisations des photographies, ils discrétisent le globe terrestre en carrés de 80 km de coté et génèrent les événement de co-localisation avec une granularité temporelle de 24 heures (un jour). Les résultats montrent que deux utilisateurs ont environ 60% de possibilité d'avoir un lien social si ils ont 5 événement de co-localisation dans des lieux différents. La vérité de terrain pour pouvoir évaluer leur inférence est issue des informations personnelles (liste des "amis") contenues dans les profils des utilisateurs sur le même réseau social.

Plusieurs aspects intéressants du concept de la co-localisation peuvent être mis en avant :

1. l'impact que les données de co-localisation ont sur la protection de la vie privée des utilisateurs de certains services web qui requièrent des informations concernant la mobilité des utilisateurs,
2. la nécessité, jusqu'à présent, de connaitre la localisation exacte des utilisateurs ou des dispositifs qui reportent l'information de proximité pour pouvoir mener une attaque de co-localisation,
3. les informations de co-localisation peuvent être (1) explicites, *i.e.* révélées directement par l'utilisateur ou nécessaires au service géo-localisé utilisé, ou (2) implicites, *i.e.* générées à partir des informations sur l'environnement ou le comportement de l'utilisateur.

La quantification de l'impact des informations de co-localisation sur la perte de protection de la vie privée fait l'objet de [Olt+14]. Les auteurs quantifient la perte de protection de la localisation des utilisateurs (*location privacy*) par une inférence utilisant les données de co-localisation. Les résultats montrent que la protection des informations de localisation est réduite de 75% pour les utilisateurs dont on connait les informations de co-localisation.
Dans [Nou+09], les données de co-localisation (assimilées aux interactions sociales selon les autheurs), issues de différents jeux de données de mobilité, sont utilisées pour opérer une attaque par inférence basée sur des techniques d'apprentissage automatique (*machine learning*).
Les données de co-localisation issues des données de mobilité peuvent aussi aider à comprendre comment les interactions sociales évoluent à l'intérieur d'un groupe d'individus [Bro+14].
Les interactions sociales, représentables par les données de co-localisation, peuvent être prises en compte dans la définition des nouveaux modèles de mobilité utilisés, par exemple, dans la caractérisation des réseaux Ad Hoc ou dans la définition des nouveaux services géo-localisés.

1.5 Conclusion

Les jeux de données de mobilité les plus utilisés, dans la caractérisation des réseaux opportunistes et dans l'étude de l'impact de ces données sur la vie privées des utilisateurs, ne permettent pas une analyse fine des interactions sociales à l'intérieur d'un groupe dense. Pour combler ce manque, nous allons mettre en place une plateforme de collecte des données de mobilité à haute precision dans un contexte de regroupement (foule) dans un espace restreint.
À partir de données ainsi récoltées, nous allons proposer une approche et des modèles visant ò analyser le comportement social des utilisateurs, en plus de l'analyse de leur mobilité spatiale.
La precision des données de mobilité réelle collectées nous permettra aussi une évaluation quantitative de la fiabilité des modèles de mobilité présentés dans ce chapitre. Cette évaluation nous permettra de vérifier la fidélité avec laquelle ces mêmes modèles sont utilisé dans la simulation des algorithmes dédiés aux réseaux opportunistes.
L'exploitation des informations de co-localisation contenues dans différentes traces de mobilité réelles nous permettra une étude des risques que la manipulation de ces données peut

engendrer sur la protection de la vie privée des utilisateurs. Cette étude portera sur l'exploration des deux nouveaux scénario d'attaque basés sur le concept de co-localisation : dans le premier cas d'étude nous allons séparer la notion de localisation de celle de co-localisation, tandis que dans le deuxième cas d'étude nous allons mesurer l'impact des informations de co-localisation probabiliste sur la perte de protection des informations de localisation.

Chapitre 2

Macro mesures

Sommaire

Comprendre les dynamiques qui règlent la mobilité humaine est fondamental pour pouvoir paramétrer efficacement les modèles de mobilité utilisés dans le développement des réseaux de communication sans fils. En particulier, les informations empiriques issues de l'analyse des traces de mobilité réelle peuvent être utilisées directement dans la définition d'un modèle de mobilité [KKK06].
Dans ce chapitre nous allons présenter un ensemble de données de mobilité. Pour ce faire, nous allons présenter la plate-forme expérimentale utilisée pour la capture des données, ainsi que les déploiements effectués. Successivement, nous allons exposer l'analyse des propriétés spatiales et sociales des traces de mobilité ainsi collectées. Plus précisément, nous allons analyser :

- le *comportement spatial*, c'est-à-dire une étude des grandeurs qui sont significatives dans la mobilité humaine (vitesse, temps d'attente, distance franchissable ou rayon d'action) et de leur impact temporel sur les trajectoires elles-mêmes (juxtaposition entre temps d'attente et temps de marche) ;
- le *comportement social*, à travers l'étude des propriétés du graphe social qui modélise les interactions entre les individus présents et qui est généré à partir de différents modèles de contact.

Le comportement spatial est caractéristique de chaque individu en opposition au comportement social qui, par définition (du latin *socialis*, de *socius* qui veut dire associé), caractérise le comportement d'au moins deux individus.
Les résultats obtenus permettront de paramétrer les modèles de mobilité présentés dans le Chapitre 1. Les traces ainsi générées nous permettront d'effectuer une comparaison entre la réalité et les modèles. La comparaison portera sur les propriétés susmentionnées et sur l'analyse des performances d'un algorithme simple de diffusion (*broadcast*).

2.1 Expériences SOUK

Nous allons présenter trois expériences de collecte de traces de mobilité utilisant le même scénario. Ces expériences ont été réalisées entre 2012 et 2015 pendant des événements sociaux où un groupe d'individus était réuni autour d'un buffet. Nous avons collecté les positions de chaque individu tout le long de l'événement avec une grande precision grâce à une plate-forme nommée *SOUK*, acronyme de *Spatial Observation of hUman Kinetics*.

2.1.1 La plate-forme

La palte-forme *SOUK* est présenté dans [Kil+16a] et elle est composée de :
– un système de captation capable de collecter la position et l'orientation de chaque individu à partir des balises (*tags*) sans fils communicant avec des bornes fixes entourant l'espace d'expérimentation.
– une collection logicielle qui permet l'exploitation des données de captation, en temps réel ou en différé, avec une analyse des traces de mobilité sous différents points de vue.

La figure 2.1 montre schématiquement la composition de la plate-forme. Nous allons détailler ensuite les composants principaux de la plateforme.

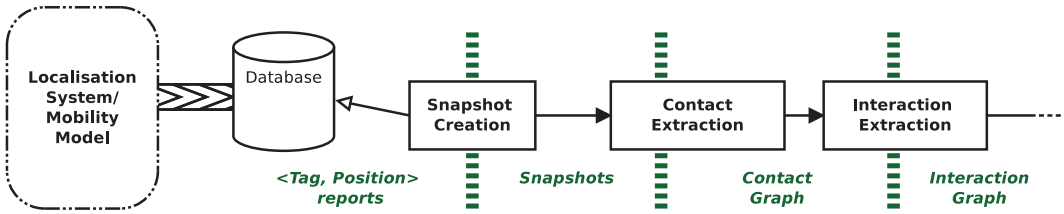


FIGURE 2.1 – Schéma illustratif de la plate-forme *SOUK*.

2.1.1.1 Système de captation

Le système physique de captation est basé sur la technologie Ubisense [Ubi]. Dans la configuration choisie, des balises (*tags*) de petite dimension ($(4 \times 4 \times 1.5\,\text{cm})$) et faible poids (25 g) sont utilisées conjointement avec des bornes de captation fixes, avec une portée conique de 90° et 25 m. La figure 2.2 montre le deux composants du système Ubisense. La communication

entre les balises et les bornes fixes est assurée en utilisant la technologie de transmission radio *Ultra Wide Band* (UWB) et le traditionnel signal Wi-Fi pour assurer la synchronisation des dispositifs.

(a) Balise (b) Borne

FIGURE 2.2 – Composants du système physique de captation.

Les balises sont placées sur chaque épaule de l'utilisateur pour pouvoir en déduire la localisation à partir des deux positions capturées. La portabilité des balises est simple au vu de leur dimensions. Les bornes de captation sont suspendues tout au tour de l'espace de l'expérience. En raison des contraintes de construction des bornes, pour une fréquence de captation de 1 Hz, le nombre maximal de participants équipés avec 2 balises est de 64. En d'autres termes, le partage du canal (TDMA) permet au système de collecter 128 positions par seconde, soit les positions des épaules de 64 individus avec une fréquence de 1 Hz.
Le système de localisation est relié à une base de données permettant le stockage horodaté des positions de chaque balise. La base de données peut être interrogée en temps réel ou a posteriori.

2.1.1.2 Collection logicielle

La plate-forme *SOUK* contient un ensemble de briques logicielles permettant le traitement des données brutes issues de la captation. A partir des n-uplet (tagID, x, y, z, timestamp), où x, y, z sont les coordonnées cartésiennes des balises, il est possible de générer des instantanés des positions des utilisateurs (*Snapshot*). À partir de *snapshots*, il est possible d'inférer les interactions sociales, les graphes de contact modélisant les potentiels contacts entre les utilisateurs étant donné un modèle de contact (*Contact Graph*) et les graphes d'interactions qui, à partir d'un modèle d'interaction (d'un niveau d'abstraction plus élevé qu'un modèle de contact), modélise les interactions entre les utilisateurs (*Interaction Graph*). Les graphes de contact et les graphes d'interaction sont des graphes dynamiques définis à chaque *snapshot*. Les graphes de contact modélisent la communication entre des dispositifs sans fils portés par les participants, tandis que, les graphes d'interaction peuvent représenter les échanges entre les utilisateurs, *e.g.* dans des configurations face-à-face ou en groupe.

2.1.2 Les données

Nous utiliserons des traces de mobilité provenant de trois événements sociaux où un certain nombre de participants prennent part à un buffet dans un espace clos. L'intérêt de ce scénario est celui de pouvoir observer la mobilité des participants dans un contexte de forte socialisation (une suite d'interactions entre les participants). Le tableau 2.1 résume les trois déploiements du système de collecte, *i.e.* les trois jeux de données.

Date	Nom	Surface d'expérimentation (m^2)	Nombre maximum de participants	Durées (minutes)
05-07-2012	souk	110	45	44
03-05-2015	milano	225	64	117
18-09-2015	cap2	100	56	101

Tableau 2.1 – Informations sur les expériences.

Souk Le jeu de données nommé *souk* (en mémoire du premier déploiement de la plateforme) contient les traces de mobilité de 63 participants au buffet d'inauguration d'un nouveau bâtiment au LAAS-CNRS, le laboratoires scientifique dans lequel nous travaillons à Toulouse. Sur les 63 traces totales, nous allons en retenir les 45 les plus consistantes (en terme de présence) sur environ trois quarts d'heure. La figure 2.3 montre le plan du hall principal du laboratoire où l'événement a eu lieu, avec la configuration spatiale des différents éléments présents dans la salle, notamment les deux buffets et le point de visualisation où une démonstration en temps réel de l'analyse des données d'interaction était montrée au public. Les participants à l'expérience étaient des scientifiques faisant partie du laboratoire d'accueil, des personnalités politiques locales et des journalistes.

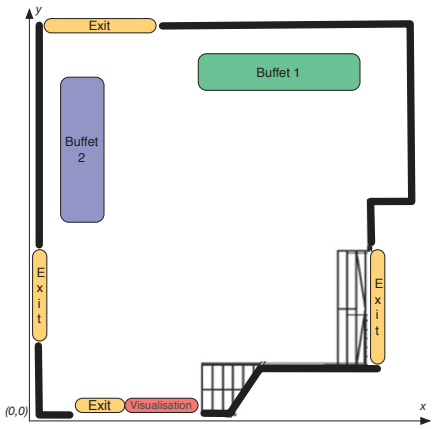


FIGURE 2.3 – Plan de l'espace de l'expérience souk.

Milano Le jeu de données nommé *milano* contient les traces de mobilité de 64 participants à l'événement de clôture d'un série d'expériences mené par plusieurs laboratoires de recherche français en collaboration avec une compagnie de théâtre expérimental italienne. Le projet de recherche, concernant la mesure de l'harmonie produite par une foule d'individus en mouvement, est décrit dans [Sto]. L'expérience qui a permis la collecte des traces de mobilité a eu lieu à l'intérieur du *Museo della scienza e della tecnologia Leonardo da Vinci* à Milan. La figure 2.4 montre le plan de l'espace de captation avec, en évidence, le placement des deux buffets prévu pour l'événement, une zone de visualisation qui contenait aussi les équipements technique de la plate-forme et une zone interdite à la marche entre les deux buffets. Les participants à l'expérience étaient surtout les personnes qui avait pris part à la performance théâtrale, les scientifiques concernés par les expériences au cœur de l'événement et certains visiteurs du musée invités à participer après la manifestation de leur curiosité pour l'événement.

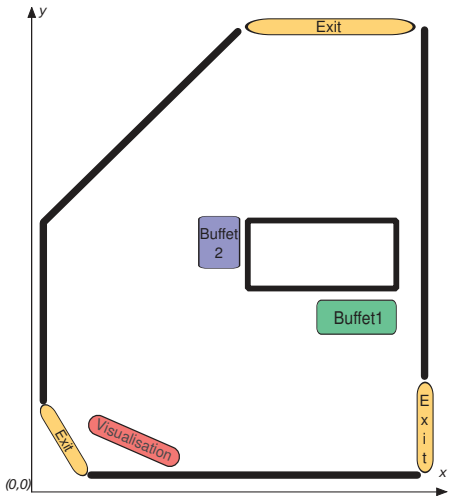


FIGURE 2.4 – Plan de l'espace de l'expérience milano.

Cap2 Le jeu de données nommé *cap2* contient les traces de mobilité de 56 participants au buffet de clôture d'une semaine d'expériences concernant la mobilité humaine, menée par plusieurs laboratoires de recherche français. Le buffet a eu lieu à l'intérieur d'une salle de spectacle dont le plan est reporté en figure 2.5. La position des buffets était symétrique et une seule voie de sortie de l'espace d'expérimentation était possible. À remarquer, la présence de la scène derrière le buffet numéro 4 a créé un banc improvisé qui a, en conséquence, généré une zone de rassemblement. Les participants à l'expérience étaient surtout des volontaires recrutés pour la semaine ainsi que des scientifiques membres des laboratoires concernés par l'organisation de l'expérience.

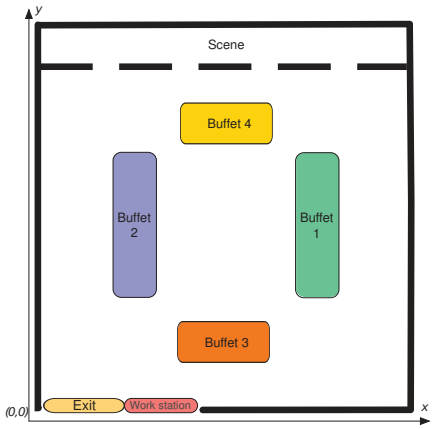


FIGURE 2.5 – Plan de l'espace de l'expérience cap2.

Les trois expériences de collecte décrites auparavant nous ont fourni des traces de mobilité dans des conditions spatiales, environnementales, sociales et temporelles diverses tout en suivant le même scénario, *i.e.* la participation à un événement social dans un espace clos, afin de capturer ainsi les interactions entre les individus présents. L'objectif est d'avoir un

échantillon suffisamment large pour pouvoir définir une collection de données réelles qui sera à la base de notre étude.

2.2 Analyse des traces réelles

À partir des traces réelles composant notre collection, nous allons présenter une analyse approfondie des propriétés spatiales et temporelles des comportements de mobilité capturés (contenus) dans les traces. Cette analyse sera suivie par une étude des caractéristiques déterminantes des contacts et des interactions issus de la mobilité. Ce lien de cause à effet entre les contacts et les interactions d'un côté et la mobilité de l'autre, n'est pas indubitable. On se déplace donc on interagit ou on interagit donc on se déplace ? L'interaction, et donc le contact, peuvent être à la fois la cause et la conséquence de la mobilité humaine. Cela implique que nous avons besoin d'un couplage entre l'analyse du comportement spatial et du comportement social des utilisateurs faisant partie d'un système mobile.
L'analyse portera donc sur le *comportement géographique* et sur le *comportement social* des individus pour pouvoir nous permettre de mieux comprendre les dynamiques propres à des phénomènes de déplacement dans des foules humaines. Simultanément, on pourra utiliser les résultats de cette analyse pour la paramétrisation des modèles de mobilité présentés dans le chapitre 1 et la génération de traces de mobilité synthétiques représentatives des données de mobilité réelle collectées.

2.2.1 Analyse préliminaire

Comme illustré dans le paragraphe 2.1.2, les espaces physiques d'expérimentation ont une forme polygonale et comprennent des sorties qui peuvent être utilisées librement par les individus impliqués dans l'expérience. La sortie peut être temporaire ou définitive.
Pour pouvoir analyser la présence tout le long des expériences, nous allons définir un polygone représentant la surface de la salle d'expérimentation. À partir des positions (x, y) collectées pour chaque balise, à chaque instant discret $t \in [1, T]$ avec $\Delta t = 1\,\mathrm{s}$, il est possible de connaitre le nombre total d'utilisateurs à l'intérieur du polygone. La figure 2.6 montre le nombre total d'utilisateurs présents pour chaque jeu de données. À ce stade on considère les positions de chaque balise, c'est-à-dire que si une des deux est à l'intérieur du polygone, alors l'utilisateur associé est considéré à l'intérieur de l'espace d'expérimentation.

La variabilité instantanée de cette mesure est due au bruit induit par la rapidité du signal de captation, tandis que, macroscopiquement, il est possible d'observer le flux d'entrée et de sortie des participants. À cette échelle, les différences concernant l'origine des participants aux expériences est bien en évidence : dans *milano*, la participation des visiteurs du musée introduit une plus grande variabilité par rapport aux deux autres expériences pendant lesquelles, après une certaine stabilité initiale, le nombre d'individus présents décroit définitivement.

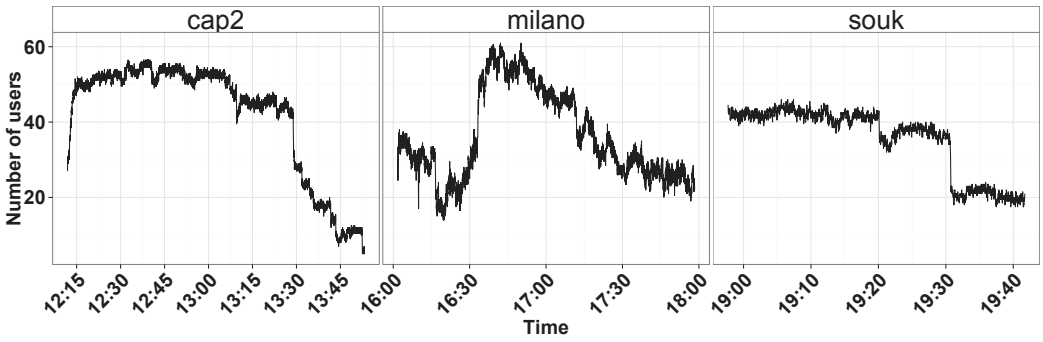


FIGURE 2.6 – Analyse des présences dans chaque expérience.

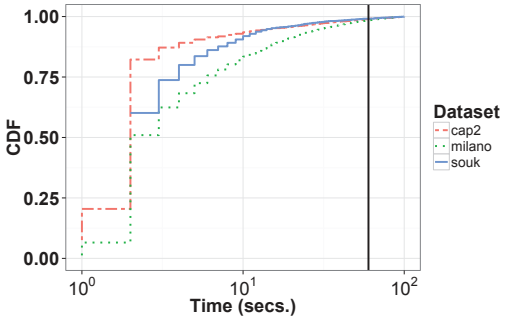


FIGURE 2.7 – Différence temporelle entre deux positions successives.

L'absence d'une position à un certain instant peut être due à une erreur de mesure (*e.g.* obstacle entre la balise et les bornes de captation) ou à la sortie de l'utilisateur de l'espace de captation. Dans le premier cas, il est possible d'avoir une partie de l'information de position (*i.e.* une seule des localisations des balises), pendant que, dans le deuxième cas il n'y aura aucune information de localisation. Pour pouvoir distinguer les deux cas, l'étude de la présence est complétée par la définition d'un *time-out*, c'est-à-dire l'identification d'une valeur de temps après laquelle on considèrera un utilisateur en dehors de l'espace de captation (la sortie peut être temporaire ou définitive). Pour définir la valeur de *time-out*, nous avons utilisé la fonction de répartition (*Cumulative Density Function*, CDF) de la variable aléatoire définie comme la différence temporelle entre deux positions successives dans les traces de chaque balise. En d'autres termes, nous avons mesuré la robustesse de la plate-forme vis-à-vis des erreurs de système (*i.e.* absences de positions). La figure 2.7 montre la function de répartition pour les trois jeux de données et, en vertical (ligne noire) la valeur choisie comme valeur de *time-out*. Nous pouvons remarquer la bonne précision du système de captation car, pour chaque jeu de données, 75% des positions collectées avec une granularité temporelle majeure de Δt ont une distance de moins de 10 s. On voit comment, dans le cas de *milano*, l'augmentation des dimensions et la non régularité de la forme de l'espace de captation réduisent la précision. En conséquence, nous définissons la valeur de *time-out* à 1 minute (ligne noire verticale sur la figure 2.7), une valeur comprise dans l'intervalle mesuré et suffisante pour justifier une sortie temporaire de l'espace de captation.

Pré-traitement des traces Pour avoir des trajectoires complètes sur l'intervalle de temps choisi pour chaque balise, une fois défini le *time-out*, nous allons reconstruire les trajectoires de chaque balise en ajoutant les points manquants à l'intérieur de l'espace de captation par interpolation linéaire. Nous allons utiliser une méthode basée sur une fonction linéaire simple.

Le résultat de l'interpolation linéaire nous permet aussi de recaler les positions de chaque balise aux mêmes instants de temps discret $t \in [1, T]$ avec $\Delta t = 1\,\text{s}$.
À partir des trajectoires complètes des deux balises (x_l, y_l, x_r, y_r) de chaque utilisateur $i \in [1, N]$, nous générons les traces de mobilité de chaque individu en considérant que sa position (x, y) est telle que

$$x = \frac{1}{2}(x_l + x_r), \quad y = \frac{1}{2}(y_l + y_r).$$

Ainsi pour chaque $t \in [1, T]$, nous avons soit la position (x, y) (l'utilisateur est présent) soit la valeur NaN (*Not a Number*) (l'utilisateur est à l'extérieur de l'espace d'expérience).

2.2.2 Propriétés spatiales

À partir des trajectoires complètes de mobilité dans les trois jeux de données réels, nous allons analyser les propriétés spatiales issues du comportement géographique des utilisateurs. En détail, nous allons analyser le profil des vitesses (la function des distribution des vitesses) qui nous permettra de définir une vitesse de marche et, consécutivement, le seuil d'immobilité, c'est-à-dire la valeur de seuil au dessous de laquelle le déplacement n'est pas identifiable comme de la marche. On rappelle que la plate-forme de captation physique est capable de récupérer la position des balises avec une haute fréquence et que les balises, en étant positionnées sur les épaules des utilisateurs, vont générer des vitesses non nulles au moindre mouvement de la partie supérieure du corps, même quand l'utilisateur est à l'arrêt. Une fois identifiés les valeurs significatives pour les vitesses, en utilisant la définition de l'état de marche et l'état d'immobilité, nous allons mesurer le rayon d'action (*flight length*, c'est-à-dire la longueur typique de marche entre deux arrêts) et le temps d'attente (où temps de pause, *waiting time*) qui caractérisent la mobilité à l'intérieur des foules.

2.2.2.1 Profils des vitesses

Pour visualiser la distribution des vitesses, nous allons calculer la vitesse avec trois temps d'échantillonnage différents (Δt) afin d'estimer la valeur de seuil indépendamment de la fréquence de captation. À partir des positions (x, y) des utilisateurs, *i.e.* $p_i(t)$ pour $i \in [1, N]$ et $t \in [1, T]$, la vitesse sera donc

$$V_{\Delta t}(i) = \frac{|p_i(t) - p_i(t - \Delta t)|}{\Delta t}, \qquad \text{où } \Delta t \in [1, 2, 3].$$

La figure 2.8 montre comment la variation de Δt ne modifie pas significativement les valeurs autour desquelles se concentre la majorité des vitesses. En autre termes, nous avons une forte densité des vitesses autour des $0.5\,\text{m/s}$, que nous allons identifier comme la valeur de référence pour la vitesse de marche (de déplacement) et $0.02\,\text{m/s}$, qui sera la valeur de

référence pour identifier des instant d'arrêt (d'immobilité).

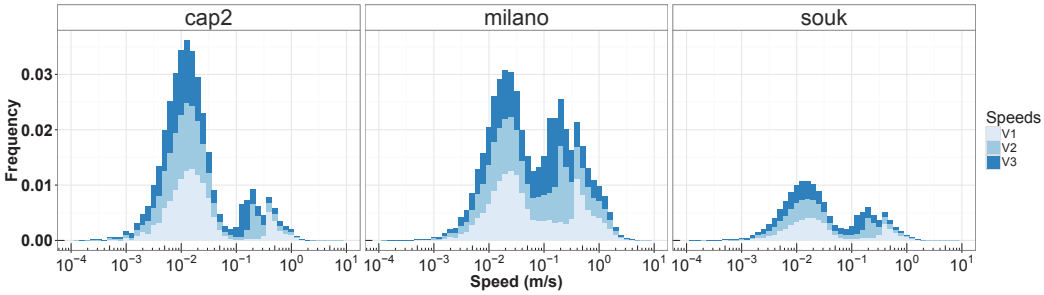


FIGURE 2.8 – Histogramme avec les densités de fréquences des vitesses dans les données réelles avec $\Delta t \in [1, 2, 3]$.

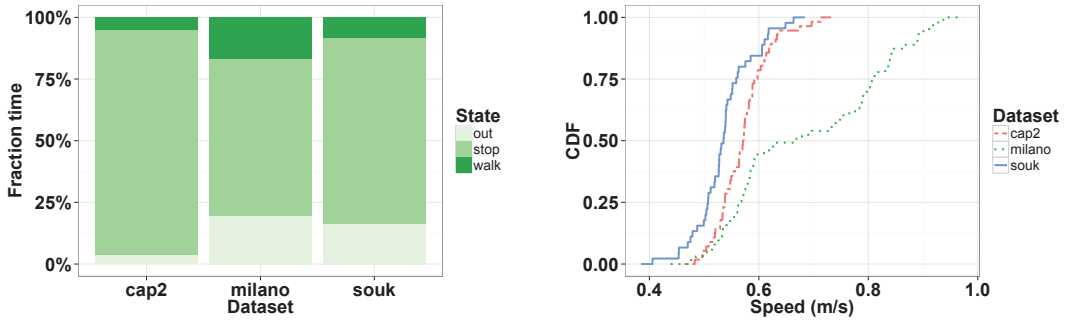
2.2.3 Caractérisation des états d'immobilité et de marche

Nous allons caractériser la mobilité des utilisateurs en définissant la mobilité au sein d'une foule dense comme la succession des états de marche et des états d'arrêt. Cette caractérisation est utilisée par la majorité des modèles de mobilité existants. Une fois les deux états définis, nous allons mesurer la vitesse de marche pour pouvoir valider notre caractérisation en comparant la valeur mesurée avec la valeur de seuil observée auparavant (voir le paragraphe 2.2.2.1). À ce propos, nous allons utiliser un algorithme basé sur le barycentre qui permet de définir l'état de l'utilisateur i à l'instant t à partir de sa trajectoire p_i. Pour pouvoir filtrer le bruit de mesure, notre algorithme utilisera un mécanisme de fenêtre glissant.

Définition 2.1 (Immobilité)
Étant donnée une fenêtre temporelle glissante de taille w centrée en $p_i(t)$ (on considère $\frac{w-1}{2}$ points avant et $\frac{w-1}{2}$ après), l'utilisateur i est immobile à l'instant t si au moins $m < w$ points appartenant à la fenêtre glissante sont géographiquement à l'intérieur d'un cercle de rayon r centré sur le barycentre généré par les n points. Autrement, l'utilisateur est considéré en déplacement (en mouvement).

La figure 2.9 montre les résultats de l'application de l'algorithme basé sur le barycentre aux données issues des trois expériences de collecte. Les tests nous ont conduit à choisir la paramétrisation suivante : $w = 5$, $m = 4$ et $r = 0.4\,\text{m}$.

La figure 2.9 (a) montre la répartition du temps total, pour chaque jeu de données, entre le temps de marche (*walk*), le temps d'arrêt (ou immobilité, *stop*) et le temps à l'extérieur de la zone d'expérience (*out*, défini dans le paragraphe 2.2.1). Comme attendu, dans un contexte de réception autour d'un buffet la majorité du temps est passée à l'arrêt avec un temps de marche inversement proportionnel à la densité spatiale, *e.g.* dans l'expérience avec la plus grande densité spatiale (*cap2*) on mesure le plus petit temps total de marche. La variation du temps passé à l'extérieur de la zone d'expérience (*out*) entre les trois différentes expériences dépend principalement de la configuration spatiale de l'espace d'expérience : dans



(a) (b)

FIGURE 2.9 – Résultats de la caractérisation de la mobilité. Les paramètres utilisés pour l'algorithme sont $w = 5$, $m = 4$ et $r = 0.4\,\mathrm{m}$.

cap2 il est beaucoup plus difficile de se retrouver "involontairement" dehors par rapport au deux autres expériences. Dans *souk* et *milano* la libre circulation et, en conséquence, le stationnement des individus à l'extérieur du polygone de captation sont plus fréquents grâce à la présence de plusieurs passages de sortie. La figure 2.9 (b) montre la fonction de répartition des vitesses de marche dans le cas des données réelles. Ici on remarque comment la vitesse de marche est inversement proportionnelle à la densité spatiale, *e.g.* dans l'expérience avec la plus petite densité spatiale (*milano*) on mesure la plus haute vitesse de marche. Cela correspond à nouveau à l'intuition : il est plus difficile de se mouvoir dans un environnement dense. En outre, dans un environnement moins dense et suffisamment large (*milano*) il est possible d'atteindre des vitesses de marche supérieures $0.8\,\mathrm{m/s}$. Simultanément, on peut vérifier que la paramétrisation choisie nous permet de mesurer des vitesses de marche comparables à celles observées dans l'analyse des profiles des vitesses (dans *souk* et *cap2* plus de 75% des vitesses de marche sont comprises entre $0.4\,\mathrm{m/s}$ et $0.6\,\mathrm{m/s}$, c'est-à-dire autour de $0.5\,\mathrm{m/s}$ qu'on avait identifié comme valeur de référence).

2.2.4 Propriétés de la mobilité

Nous allons analyser deux propriétés de la mobilité humaine, notamment la distance franchissable (*flight length*) et le temps d'attente (*waiting time*).

Distance franchissable Pour évaluer la distance franchissable nous allons mesurer la distance parcourue entre deux états d'arrêt, c'est-à-dire qu'après la segmentation des trajectoires en temps de marche et temps de pause nous allons étudier la répétition de ces états. La figure 2.10 montre la fonction de répartition de la variable aléatoire représentant le temps de marche. On retrouve des valeurs comparables dans les trois cas et la ligne noire verticale sur la figure 2.10 reporte la moyenne arithmétique sur l'ensemble des mesures (sur les trois expériences). Spécifiquement, on observe une valeur moyenne de 1.73 m, c'est-à-dire qu'un déplacement est en moyenne inférieur à deux mètres.

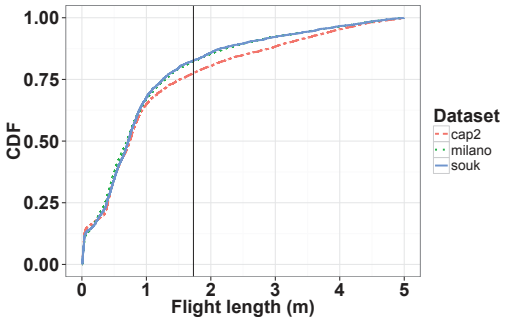


FIGURE 2.10 – Fonction de répartition de la distance franchissable dans le cas des données réelles.

Temps d'attente Symétriquement, pour évaluer le temps d'attente nous allons mesuré les temps entre deux états de marche. La figure 2.11 montre la fonction de répartition de la variable aléatoire représentant le temps d'attente. On retrouve des valeurs comparables dans les trois cas et la ligne noire verticale sur la figure 2.11 reporte la moyenne arithmétique sur l'ensemble des mesures (sur les trois expériences). Spécifiquement, on observe une valeur moyenne de 35 sec., c'est-à-dire qu'une attente est en moyenne inférieure à une minute.

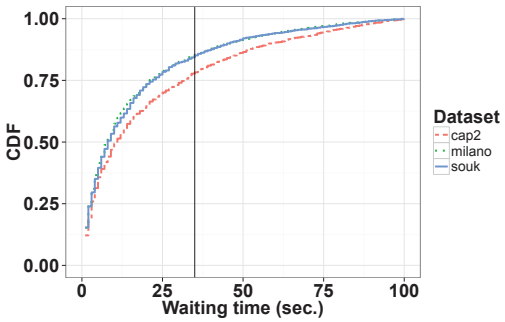


FIGURE 2.11 – Fonction de répartition du temps d'attente dans le cas des données réelles.

2.2.5 Propriétés sociales

Nous allons maintenant analyser les propriétés sociales issues des trajectoires. En détail, nous allons analyser, à l'aide des différents modèles de lien (nécessaires pour la définition des contacts et des interactions), les propriétés topologiques et dynamiques du graphe social existant à l'intérieur de la foule. Ce graphe sera défini à chaque instant t à partir de la matrice d'adjacence résultante de l'application des modèles de génération présentés plus loin.

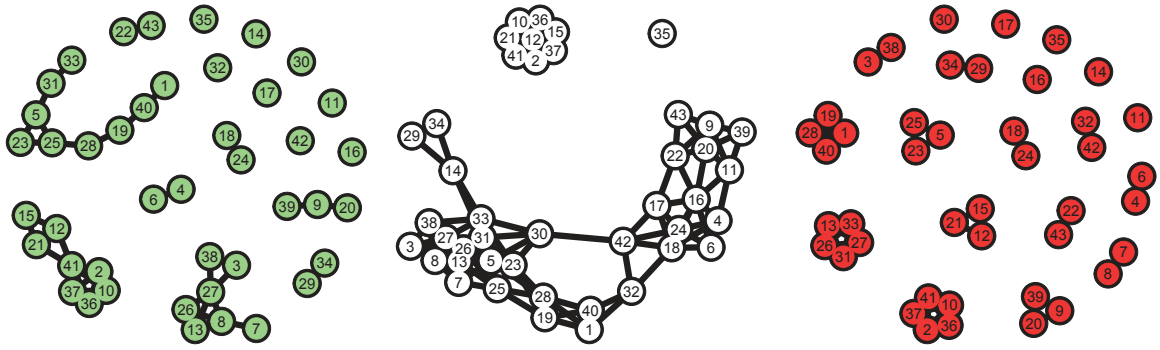
Si on considère le graphe à chaque instant, on pourra étudier le propriétés topologiques du graphe même, notamment le diamètre, la taille de la plus grande composante connexe et la modularité. Différemment, dans le cas où on considère le graphe dynamique comme l'évolution temporelle du graphe social tout au long de l'expérience, on pourra étudier les propriétés temporelles des liens présents entre les utilisateurs par le biais de l'analyse du temps de contact et du temps d'inter-contact.

2.2.5.1 Modèles de lien

L'objectif est de définir des modèles de lien pour des graphes sociaux G_t à chaque instant de temps discret $t \in [1, \ldots, T]$. Le graphe social à chaque instant sera un graphe simple (graphe non orienté, ne contenant pas des transitions multiples et des boucles) défini comme $G_t = (V_t, E_t)$ avec $V_t \subset V$, où V est l'ensemble des utilisateurs équipés dans l'expérience et E_t les liens existants entre les V_t utilisateurs présents à l'instant t. Les modèles de génération nous permettent de définir l'ensemble E_t, c'est-à-dire nous permettent de définir les contacts et les interactions entre les utilisateurs. La différentiation entre contacts et interactions n'est pas seulement lexicale mais aussi conceptuelle.
Le contact dénote une action de toucher physique, matérielle, de deux corps, alors nous allons considérer qu'un contact existe entre deux utilisateurs quand ils sont à portée de communication (toucher physique de leur champ de communication), communication supposée possible par le biais des dispositifs sans fil mobiles portés par les utilisateurs mêmes. Avec la même approche, vu qu'une interaction peut être définie comme une action réciproque de deux ou plusieurs phénomènes, nous allons considérer qu'une interaction existe entre deux ou plusieurs utilisateurs quand ils sont dans une configuration spatiale assimilable à un groupe.
Plus formellement, nous allons définir deux modèles de lien :

- modèle de *contact*, basé sur la distance euclidienne entre les positions instantanées des utilisateurs, deux utilisateurs sont en contact si et seulement si leur distance est inférieure à une valeur r de seuil, $E_t(u, v) = 1 \Leftrightarrow d(u, v) < r$. Le modèle simule donc la présence d'une connexion physique entre des dispositifs (idéalement transportés par les utilisateurs) sans fil et sera utilisé avec deux configurations différentes, c'est-à-dire avec une distance de seuil r de 1 m et 2 m.
- modèle d'*interaction*, basé sur un algorithme de *clustering* qui permet la caractérisation quantitative sur les dynamiques de formation des groupes à l'intérieur d'une foule d'utilisateurs. Deux où plusieurs utilisateurs sont en interaction si et seulement si ils font partie du même groupe. Chaque groupe sera un sous-graphe complet à l'intérieur du graphe social. Dans la suite du chapitre on fera référence à ce modèle par l'acronyme *ftf* en contraction de *face to face*.

La figure 2.12 montre les trois différents graphes sociaux générés à partir des modèles décrits auparavant à partir des positions des utilisateurs à un certain instant de temps. Les graphes en figure 2.12(a) et 2.12(b) sont issus du modèle de contact en utilisant les deux configurations choisies, tandis que la figure 2.12(c) contient le graphe dérivé du modèle d'interaction.



(a) $r = 1\,m$ (b) $r = 2\,m$ (c) ftf

FIGURE 2.12 – Graphes sociaux issus des modèles de lien à partir des positions à un instant donné.

Nous allons implémenter les deux modèles de génération et les appliquer sur l'ensemble des traces de mobilité réelles afin d'avoir, pour chaque jeu de données, une liste des matrices d'adjacence représentant les graphes sociaux à chaque instant et pour les différents modèles.

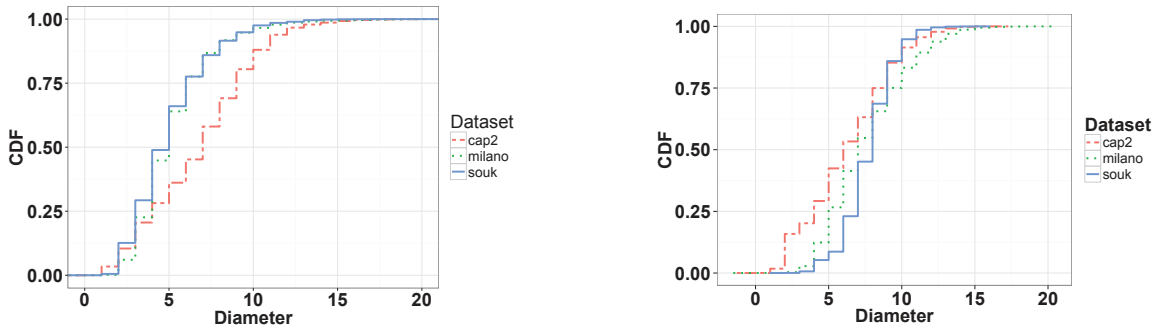
2.2.5.2 Analyse statique

À partir des matrices d'adjacence représentant les graphes sociaux à chaque instant de temps t défini dans l'expérience, nous allons pouvoir effectuer une analyse statique, c'est-à-dire une analyse des propriétés topologiques de chaque graphe.

Diamètre En suivant la définition presentée dans [New10], le diamètre d'un graphe est la longueur du plus long chemin géodésique qui existe entre tous les couples de nœuds reliés entre eux. En d'autres termes, le diamètre représente la distance (en théorie des graphes, le plus court chemin entre deux nœuds) la plus grande entre deux nœuds du graphe. Dans le cas d'un graphe non connexe, le diamètre sera le plus grand parmi les diamètres des composantes connexes.
La figure 2.13 montre la fonction de répartition de la mesure du diamètre dans le cas du modèle de contact dans les deux configurations de distance. Dans les deux cas, dans 75% des graphes, le diamètre est inférieur à 10.

Dans le cas du modèle de génération des interactions, le diamètre a une valeur par construction toujours égale à un car les graphes seront toujours divisés en sous-graphes complets par l'algorithme de *clustering*.

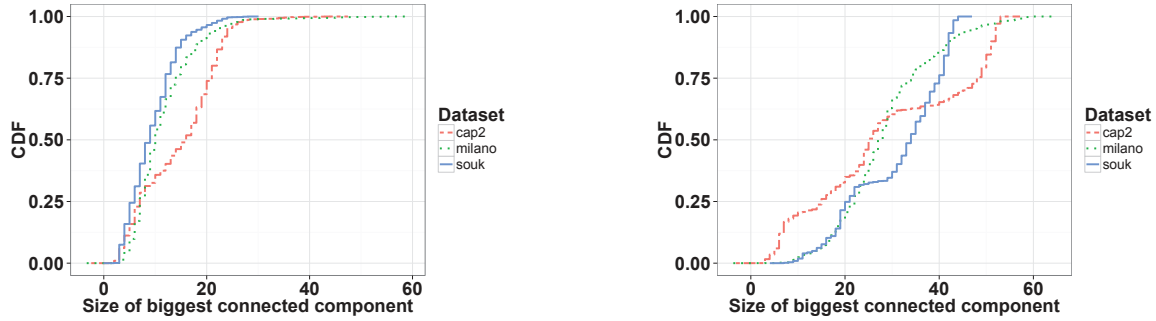
Composante connexe Dans la majorité des cas, tous les utilisateurs ne sont pas reliés par un chemin, *i.e.* le graphe social est non connexe. Une composante connexe est un sous-



(a) $r = 1\,m$ (b) $r = 2\,m$

FIGURE 2.13 – Fonction de répartition de la mesure du diamètre dans le cas du modèle de génération des contacts dans les deux configurations considérées.

graphe où il existe un chemin entre chaque couple de nœuds. La mesure de la taille de la plus grande de ces composantes nous permet de caractériser la connectivité du réseau. La figure 2.14 montre la fonction de répartition de la taille de la plus grande composante connexe dans le cas du modèle de contact dans les deux configurations de distance.



(a) $r = 1\,m$ (b) $r = 2\,m$

FIGURE 2.14 – Fonction de répartition de la taille de la plus grande composante connexe par des contacts à $1m$ (a) et $2m$ (b).

Dans la figure 2.14 (a) nous pouvons remarquer que pour *milano* et *souk*, dans 50% des cas (valeur médiane) la taille de la composante connexe est au plus égale à 10, c'est-à-dire qu'il y aura respectivement au plus 16% ($\frac{10}{64}$) et 22% ($\frac{10}{45}$) du total des utilisateurs connectés. Dans le cas de *cap2*, l'expérience avec la plus importante densité spatiale, dans 50% des cas il y aura environ 29% ($\frac{16}{56}$) du total des utilisateurs connectés. Dans la figure 2.14 (b) nous pouvons remarquer une augmentation généralisée attendue à cause du doublement du rayon de contact. En outre, l'oscillation des distributions dans *souk* et *cap2* montre comment, avec le modèle de contact de rayon $r = 2m$, la taille de la plus grande composante connexe est sensible aux déplacements des individus dans un environnement dense. Dans un espace moins dense, comme il est le cas dans *milano*, il faudra un rayon plus important pour introduire la même variabilité dans la taille des composantes connexes.

La figure 2.15 montre la fonction de répartition de la taille de la plus grande composante connexe dans le cas du modèle d'interaction. En ligne avec nos attentes, nous avons une diminution de la taille par rapport au résultat issu de l'application du modèle de contact. Dans ce cas, l'impact de la densité spatiale est moins évident, c'est-à-dire que le niveau d'interaction (définie par la formation des sous-groupes par *clustering*) à l'intérieur d'un groupe d'individus est indépendant de la densité spatiale. Pour 50% des cas, et dans les trois expériences, la taille de la plus grande composante connexe est au plus égale à 5.

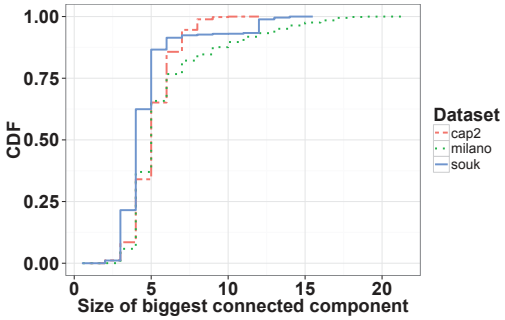


FIGURE 2.15 – Fonction de répartition de la taille de la plus grande composante connexe dans le cas du modèle de génération des interactions.

Modularité Selon la définition donnée dans [New06], l'indice de modularité donne la mesure de la possibilité de partager le graphe en modules (ou communautés). Un graphe avec une haute valeur de modularité a des fortes connexions (nombre significatif de liens) entre les nœuds faisant partie du même module (communauté) et des faibles connexions (nombre faible de liens) avec les nœuds à l'extérieur du module (communauté). L'indice de modularité est défini dans l'intervalle $[-\frac{1}{2}, 1]$. Le tableau 2.2 reporte la mesure de l'indice de modularité dans les graphes générés à partir des données réelles en utilisant les modèles de génération présentés auparavant.

Modèle	cap2	milano	souk
Contact 1m	0.6402295	0.673432	0.7049831
Contact 2m	0.3380523	0.3528862	0.1487984
Interaction (ftf)	0.8276722	0.8086421	0.8247639

Tableau 2.2 – Indice de modularité dans les jeux de données réels.

Les valeurs mesurées pour l'indice de modularité sont cohérentes avec les différentes topologies des graphes sociaux. Les graphes sociaux les plus "modulaires" sont ceux dérivant du modèle de génération des interactions, graphes composés des sous-graphes complets, tandis que les moins "modulaires" sont ceux dérivant du modèle de génération des contacts où l'indice de modularité est inversement proportionnel à la distance de seuil considéré.

2.2.5.3 Analyse dynamique

À partir de l'ensemble des graphes sociaux issus des trois expériences, nous allons pouvoir effectuer une analyse des propriétés dynamiques du système de mobilité, c'est-à-dire une mesure des propriétés de communication à l'intérieur du groupe d'individus représenté par

l'ensemble des graphes. En d'autres termes, nous allons caractériser les contacts et les interactions par la mesure du temps de contact (*contact time*) et d'inter-contact (*inter-contact time*). Les distributions des temps seront mesurées pour chaque couple d'utilisateurs (*i.e.* chaque lien $(i, j) \in E(G_t) \forall t \in [1, \ldots, T]$) et elles seront agrégées pour rendre la fonction de répartition complémentaire (*complementary cumulative distribution function*) de l'ensemble des temps de contact et d'inter-contact.

Temps de contact Le temps de contact est la durée d'un lien dans l'ensemble des graphes sociaux dynamiques. La figure 2.16 montre le résultat de la mesure du temps de contact agrégé sur l'ensemble des liens (des couples d'utilisateurs) sous la forme de fonction de distribution complémentaire (sur l'axe des ordonnés la $\mathbb{P}$[temps de contact $> t$], temps t reportés sur l'axe des abscisses) pour les trois jeux de données.

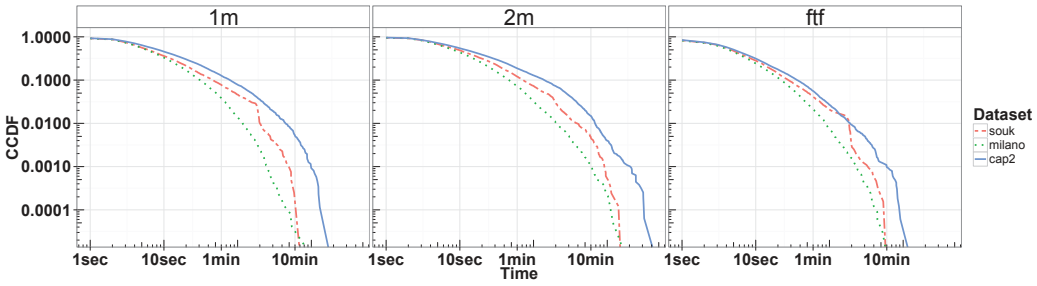


FIGURE 2.16 – Fonction de distribution complémentaire des temps de contact dans les données réelles dans un repère log-log.

Le temps de contact varie de façon proportionnelle à la densité spatiale dans les trois cas étudiés. Comme on s'attendait, dans le cas du modèle de génération basé sur le *clustering* (*ftf*), le temps de contact est inférieur à celui mesuré dans le cas du modèle de génération basé sur la distance.

Temps d'inter-contact En suivant la definition du temps d'inter-contact donnée dans [Cha+07a], nous allons mesurer le temps écoulé entre des contacts, c'est-à-dire entre la fin d'un contact et le début du suivant sans considérer le temps écoulé entre le début de l'expérience et le premier contact et, symétriquement, la fin du dernier contact et la fin de l'expérience. Le temps d'inter-contact est un paramètre central dans la caractérisation des algorithmes pour les réseaux opportunistes entre dispositifs portables sans fils [PC11]. La figure 2.17 montre le résultat de la mesure du temps d'inter-contact dans le cas des données réelles.

Souvent le temps d'inter-contact dans un système mobile est modélisé en utilisant une loi de puissance (*power law*) dont le coefficient est inféré statistiquement sur l'agrégation total des mesures. La figure 2.17 montre comment la fonction de distribution du temps d'inter-contact est assimilable à une loi de puissance tronquée naturellement à un temps comparable à la durée de l'expérience.

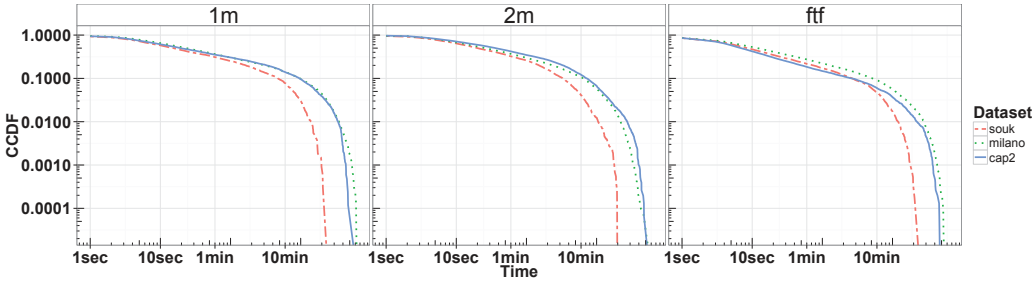


FIGURE 2.17 – Fonction de distribution complémentaire des temps d'inter-contact dans les données réelles.

2.3 Paramétrisation des modèles

Suivant l'implementation du générateur `pymobility` [Pan16] des modèles de mobilité présentés dans la Section 1.3, nous allons choisir les paramètres nécessaires à la simulation des traces de mobilité à partir des mesures effectuées sur les données de mobilité réelles. Dans le choix des paramètres, nous allons exposer les différences que l'implémentation dans `pymobility` comporte par rapport aux définitions originales des modèles.
Tous les modèles ont besoin de deux paramètres :

1. le nombre de nœuds mobiles à l'intérieur du système,
2. la dimension de l'espace de simulation.

Le reste des paramètres dépendra de la complexité du modèle et de son implémentation. Nous allons présenter les modèles en suivant le schéma choisi par l'auteur de l'implémentation, c'est-à-dire que les modèles de mobilité individuelle seront présentés avant les modèles de mobilité de groupe.

***Random Waypoint* (rwp)** C'est le modèle le plus simple à paramétrer car il ne nécessite que deux valeurs de seuil décrivant la vitesse des nœuds. Les deux valeurs de vitesse mesurées dans la partie 2.2.2.1 sont utilisées : vitesse maximale v_{max} et minimale v_{min}, respectivement $1\,\mathrm{m/s}$ et $0.02\,\mathrm{m/s}$. La vitesse de marche sera donc autour de $0.5\,\mathrm{m/s}$.

Modèles dérivés du *Stochastic Walk* L'auteur de *pymobility* choisit une implémentation de base pour les modèles qui définissent le choix uniforme d'une direction dans un intervalle décrit en radiant. Les modèles concernés sont donc *Random Walk* (rw), *Random Direction* (rd) et *Truncated Lévy Walk* (tlw). À signaler ici la première différence d'implémentation par rapport aux définitions originales : l'intervalle où sera choisi uniformément la direction est toujours fixé à $[0, \pi]$. L'impact de ce choix est minimisé grâce à l'introduction d'un paramètre de gestion des collisions contre les bornes de l'espace de simulation, paramètre à deux valeurs possibles (*reflect*, *wrap*) pour permettre aux nœuds, après une collision avec les bornes, de

pouvoir repartir en sens inverse (on inverse la direction, *i.e.* $[0, -\pi]$) ou de réapparaître à l'opposé de l'espace. Les autres paramètres nécessaires et les différentes fonctions qui les implémentent sont résumés dans le tableau 2.3.

Modèles	Distance franchissable	Temps d'attente	Vitesse
Random Walk	Constante	*None*	Constante
Random Direction	Uniforme	Uniforme	Uniforme
Truncated Lévy Walk	Loi de puissance tronquée	Loi de puissance tronquée	Loi de puissance tronquée

Tableau 2.3 – Paramétrisation des modèles dérivés du *Stochastic Walk*.

Dans le cas du *Random Walk*, la distance franchissable (*Flight length*) et la vitesse des nœuds (*Node velocity*) sont constantes, en conséquence on utilisera les valeurs mesurées auparavant sur les données réelles, respectivement la valeur moyenne de 1.65 m et la vitesse de marche 0.5 m/s.
Pour le modèle *Random Direction*, les fonctions uniformes seront définies entre 0 et une valeur maximale pour la distance franchissable (*Flight length*) et le temps d'attente (*Waiting time*), valeurs respectivement fixes à 1.65 m et 35 sec. Pour ce qui concerne la vitesse des nœuds, la fonction d'implémentation est choisie uniformément dans l'intervalle $[v_{min}, v_{max}]$. Pour faire en sorte que la vitesse de marche soit autour de 0.5 m/s, nous utilisons comme valeur de borne 0.02 m/s et 1 m/s.
L'implémentation du modèle *Truncated Lévy Walk* requiert la paramétrisation des deux fonction de distribution suivant des lois de puissance (*Power law*) $p(x) \propto x^{-\alpha}$. L'annexe A décrit la paramétrisation des deux fonctions de distribution pour la distance franchissable (*Flight length*) et le temps d'attente (*Waiting time*) qui nous donne respectivement $\alpha_{fl} = 2.1$ et $\alpha_{wt} = 2$. La fonction pour la vitesse des nœuds est fonction de la distance franchissable, c'est-à-dire une loi de puissance avec le même paramètre α_{fl}.

Gauss-Markov (gm) Le paramètre principal de ce modèle est α, $0 \leq \alpha \leq 1$ qui indique le caractère aléatoire (l'entropie, *randomness*) de la mise à jour, à chaque instant t, de la vitesse et de la direction de chaque nœud. Si $\alpha = 0$ la mobilité des nœuds sera complètement aléatoire, tandis que pour $\alpha = 1$ la mobilité des nœuds sera linéaire. Les deux autres paramètres caractéristiques pour l'implémentation dans *pymobility* de ce modèle sont la variance de α et la vitesse moyenne des nœuds. Nous avons fixé $\alpha = 0.5$, $Var(\alpha) = 1$ et $\bar{v} = 0.5$ m/s.

***Reference Point Group* (rpgm)** Le paramètre essentiel pour l'implémentation de ce modèle de mobilité de groupe est le paramètre d'agrégation. Ce paramètre va permettre de moduler la distribution spatiale des nœuds autour du centre du groupe, *i.e.* la distance entre les nœuds et le point de référence pour chaque groupe. Ce paramètre est compris entre 0 et 1, respectivement dans le cas où les nœuds seront aléatoirement distribués dans la surface d'expérimentation et le cas où les nœuds seront très proches du centre du groupe. Pour le choix de la valeur à utiliser pendant la simulation, nous allons utiliser la mesure de la modularité faite dans l'analyse statique des graphes sociaux. En conséquence, nous allons fixer ce paramètre à 0.73 pour représenter au mieux la valeur de la modularité dans les trois

jeux de données réelles. Pour compléter la paramétrisation, le simulateur aura besoin des bornes pour les vitesses qui seront fixées à $v_{min} = 0.02\,\mathrm{m/s}$ et $v_{max} = 1\,\mathrm{m/s}$.

***Time-Variant Community* (trw)** Dans l'implémentation de ce modèle on trouve une différence majeure par rapport à sa définition originale, à savoir chaque groupe a comme référence un point mobile (selon le modèle *Random Direction*) dans l'espace et non pas une région spatiale fixe. Pour ce qui concerne le comportement dans les différentes périodes de mouvement, il est possible de définir dans le simulateur une liste de durées des différentes périodes de comportement de mobilité (voir la section 1.3.2). Pendant chacun des ces temps un coefficient d'agrégation sera défini pour permettre aux nœuds de se déplacer plus où moins loin du point de référence. En particulier, si le coefficient d'agrégation est 0 les nœuds n'ont aucun point d'attraction (de référence) et se déplacent en suivant un modèle de mobilité *Random Walk*, tandis que si le coefficient d'agrégation est 1 ils se déplacent autour du point de référence. Pour rester le plus proche de sa définition originale, nous allons définir deux temps d'égale durée avec deux valeurs pour le coefficient d'agrégation, 0 et 0.73 (d'après la mesure de la modularité), de manière à simuler respectivement le comportement local (*local epoch*) et le comportement itinérant (*roaming epoch*). Pour compléter la paramétrisation, le simulateur requiert des bornes pour des vitesses qui seront fixées à $v_{min} = 0.02\,\mathrm{m/s}$ et $v_{max} = 1\,\mathrm{m/s}$.

Avec les paramètres décrits dans cette Section, nous allons pouvoir générer des traces de mobilité synthétiques pour 55 nœuds mobiles dans une surface de $144\,m^2$ et pour une durée de 90 minutes.

2.4 Comparaison

Après la paramétrisation issue des mesures effectuées sur les traces réelles, nous allons étudier les propriétés spatiales et sociales des traces synthétiques générés à partir des modèles au centre de notre étude. L'analyse des résultats ainsi obtenus permet d'évaluer la fiabilité qu'ont les différents modèles pour la représentation d'un contexte particulier, ici des individus dans un espace dense. En outre, la connaissance des propriétés spatiales et sociales des modèles permet la sélection du modèle adapté aux besoins de l'étude.
Pour ce qui concerne l'utilisation des traces synthétiques dans la simulation de systèmes mobiles et, en particulier, dans le développement d'algorithmes de communication dans des systèmes opportunistes, nous allons comparer les performances d'un simple algorithme de diffusion (*broadcast*) sur les traces réelles et sur les traces synthétiques. Cette comparaison permettra une évaluation du réalisme des traces produites par les modèles et de leur applicabilité à la simulation de systèmes répartis.

2.4.1 Propriétés spatiales

Nous allons voir en quoi les propriétés spatiales des traces synthétiques sont en accord avec les mesures menées auparavant sur les traces réelles.

Vitesses La figure 2.18 montre la répartition des vitesses mesurées sur les sept modèles de mobilité.

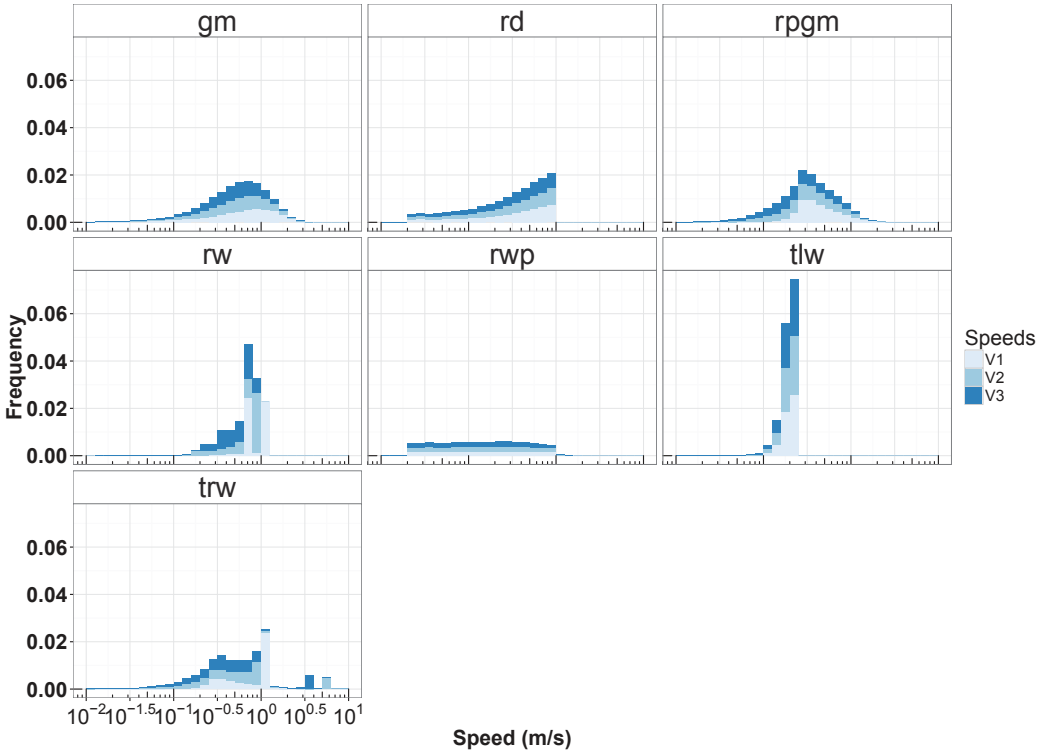


FIGURE 2.18 – Histogramme avec les densités de fréquences des vitesses dans les traces synthétiques.

Dans les cas des modèles Gauss-Markov (gm) et *Reference Point Group* (rpgm) on retrouve des répartitions des vitesses similaires à celles mesurées dans les données réelles (voir figure 2.8). Dans ce deux cas, les vitesses de marche sont bien réparties autour de 0.5 m/s, qui est la valeur de référence pour la vitesse de marche mesuré dans les cas réels.

Caractérisation des états d'immobilité et de marche La figure 2.19 montre les proportions du temps total de simulation entre les états de marche et les états d'immobilité pour chaque modèle.

Dans la définition des modèles il n'y a pas la possibilité d'envisager des instants de temps passés à l'extérieur de la surface de simulation. En conséquence, l'état nommé *out* n'a pas de signification dans ce cas. En modifiant la stratégie de prise en considération des collisions

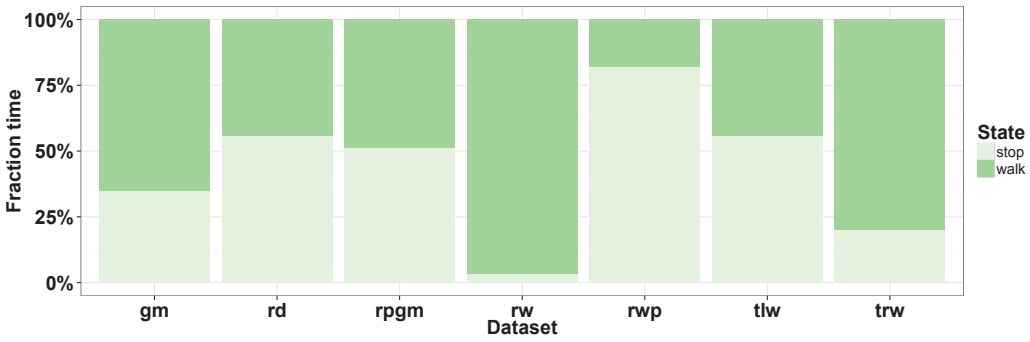


FIGURE 2.19 – Répartition du temps totale de la simulation.

des nœuds mobiles contre le bord de l'espace de simulation on pourra, dans certains modèles, introduire la génération des états d'absence. Étant donné que nous nous limitons à une analyse des modèles existants, on gardera la possibilité de cette modification pour une éventuelle proposition future d'un nouveau modèle de mobilité.
Le modèle *Random Waypoint* (rwp) est le seul qui s'approche bien au cas réel (moins du 25% du temps total en état de marche). En ligne avec nos attentes, le plus inadapté à notre réalité est le *Random Walk* (rw) où la quasi totalité du temps de simulation est en état de marche.

La figure 2.20 montre la fonction de répartition des vitesses de marche mesurées après la caractérisation des états de marche et d'arrêt.

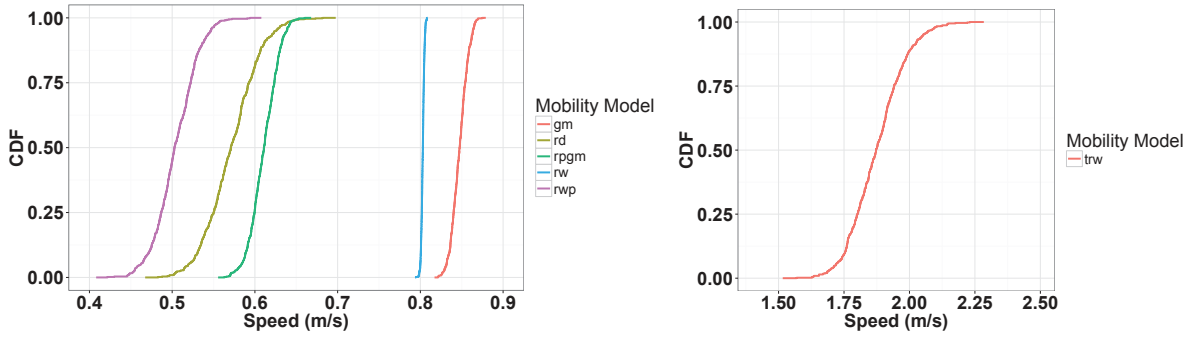


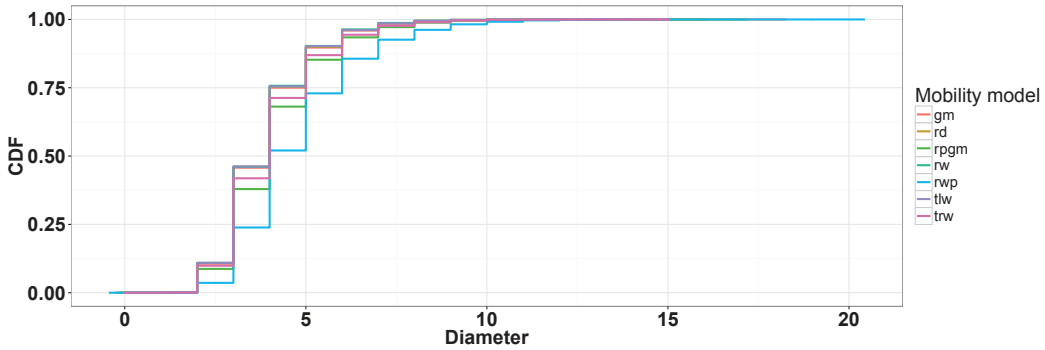
FIGURE 2.20 – Fonction de répartition des vitesses de marche.

Les vitesses mesurées sont globalement supérieures à celles mesurées sur les traces réelles. Le *Random Waypoint* (rwp) est le plus proche au cas réel avec une densité spatiale comparable, c'est-à-dire *souk* (on rappelle que dans la génération des traces synthétiques une seule valeur de densité spatiale est utilisée).

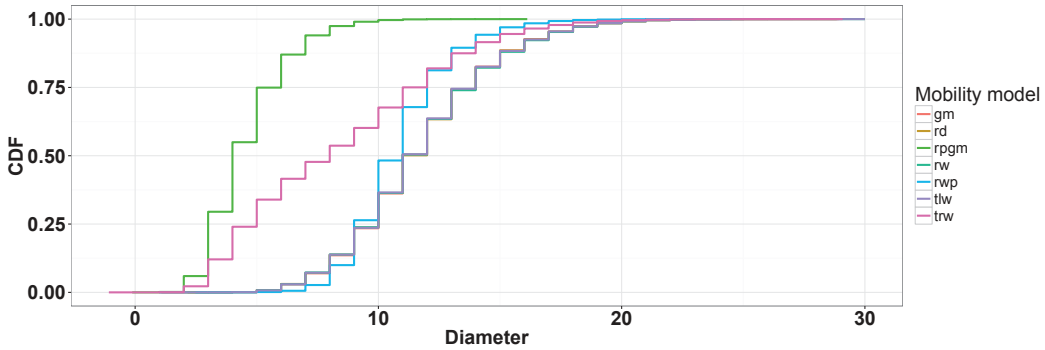
2.4.2 Propriétés sociales

Nous avons vu que les modèles de mobilité génèrent des traces aux propriétés spatiales parfois bien différentes des traces capturées. Ce défaut n'est pas forcément rédhibitoire si la connectivité entre les nœuds est réaliste. En utilisant les deux modèles de lien, nous étudions les propriétés des graphes sociaux pour chaque modèle de mobilité et à chaque instant de simulation.

Diamètre La figure 2.21 montre la fonction de répartition de la variable aléatoire décrivant le diamètre des graphes à chaque instant de simulation.



(a) $r = 1\,m$

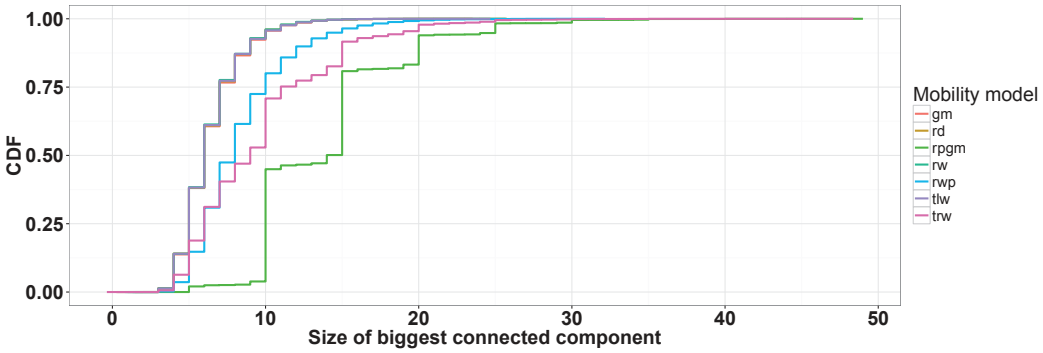


(b) $r = 2\,m$

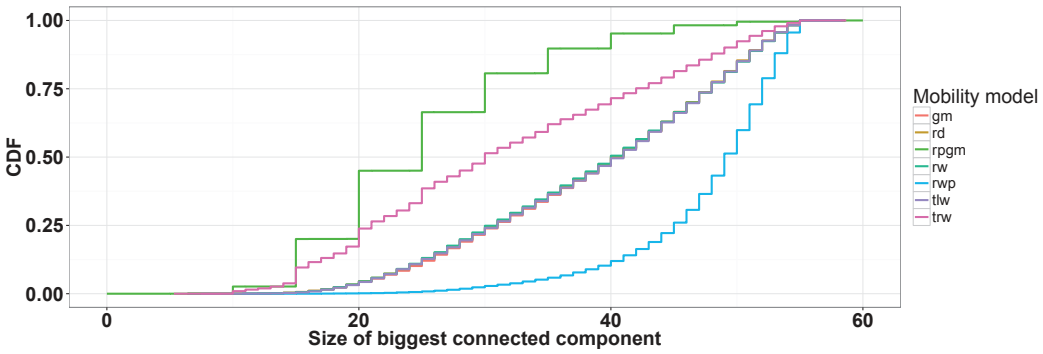
FIGURE 2.21 – Fonction de répartition de la mesure du diamètre dans le cas des traces synthétiques.

Dans le cas du modèle de contact avec une distance de référence de 1 m, figure 2.21 (a) 75% des graphes ont un diamètre inférieur ou égal à 5, la moitié par rapport à la valeur médiane mesurée pour le traces réelles. Dans le cas du modèle de contact avec une distance de référence de 2 m, figure 2.21 (b) seul le modèle *Reference Point Group* (rpgm) offre une approximation correcte du comportement relevé sur les traces réelles.

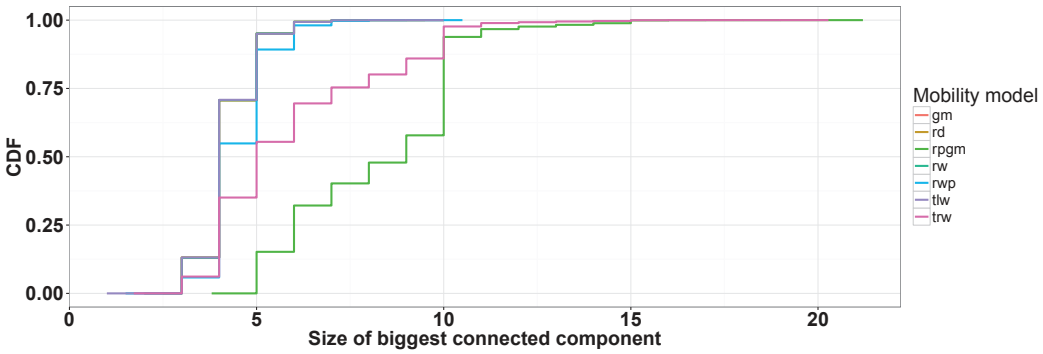
Composante connexe La figure 2.22 montre la fonction de répartition de la variable aléatoire décrivant la taille de la plus grande composante connexe à chaque instant de simulation.



(a) $r = 1\,m$



(b) $r = 2\,m$



(c) ftf

FIGURE 2.22 – Fonction de répartition de la mesure de la taille de la plus grande composante connexe dans le cas des traces synthétiques.

Dans le cas du modèle de contact basé sur une distance de 1 m, figure 2.22 (a) la valeur médiane de la taille de la plus grande composante connexe oscille entre 6 et 14. Cela implique que, dans la moitié du temps de simulation, entre 11% et 25% des utilisateurs font partie de la plus grande composante connexe. Ces valeurs sont légèrement au dessous des celles observées sur les traces réelles. Quand la distance de référence pour le modèle de contact augmente,

figure 2.22 (b) la taille de la plus grande composante connexe augmente de façon différente entre les modèles de mobilité individuelle, faisant exception du *Random Waypoint* (rwp), et le modèle de mobilité de groupe. Ces derniers approximent mieux les traces réelles.
Dans le cas du modèle d'interaction basé sur le *clustering* en figure 2.22 (c), en suivant nos attentes, la taille de la plus grande composante connexe est la plus petite parmi celles mesurées auparavant. Dans ce cas, ce sont les modèles de mobilité individuelle qui réussissent à mieux approximer les cas réels par rapport aux modèles de mobilité de groupe.

Temps de contact et d'inter-contact La figure 2.23 montre la fonction de distribution complémentaire des temps de contact.

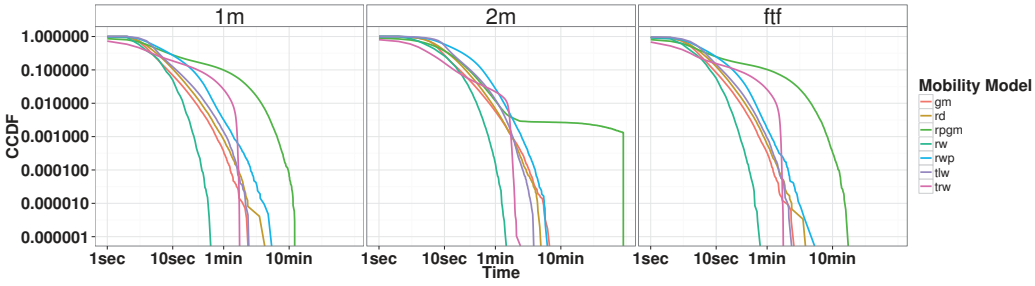


FIGURE 2.23 – Fonction de distribution complémentaire des temps de contact.

Le temps de contact est globalement plus court que dans les cas réels. Le modèle *Reference Point Group* (rpgm) est celui qui approxime au mieux les temps de contact mesurés dans les traces réelles.

La figure 2.24 montre la fonction de distribution complémentaire des temps d'inter-contact.

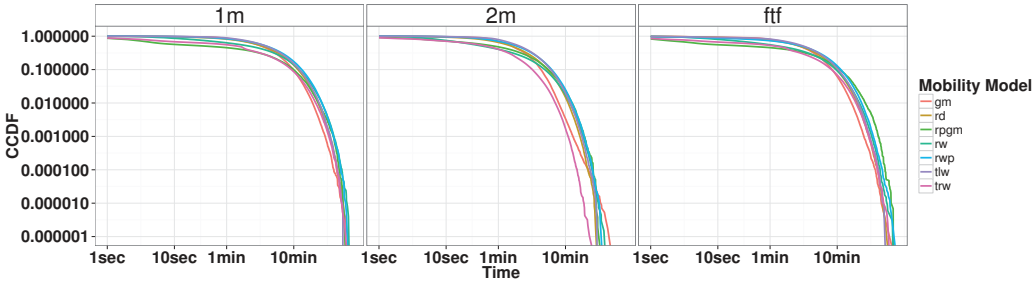


FIGURE 2.24 – Fonction de distribution complémentaire des temps d'inter-contact.

Le temps d'inter-contact est, au contraire du temps de contact, globalement plus long que dans les cas réels. Le comportement des modèles de mobilité individuelle et ceux de mobilité de groupe est comparable.
Cette sous estimation pour le temps de contact et sur estimation pour le temps d'inter-contact sont cohérentes avec la mesure du temps de marche et du temps d'immobilité. Dans les traces de mobilité réelles, les utilisateurs passent globalement plus de temps immobiles par rapport aux traces synthétiques. L'effet de ne pas considérer dans les modèles l'aspect social de la

mobilité des utilisateurs, à un impact sur la simulation des contacts (ou interactions) entre les nœuds synthétiques.

2.4.3 Algorithme de diffusion

Une des primitives de base des systèmes distribués est la diffusion (*broadcast*). Nous allons mesurer les performances d'un algorithme de diffusion dans le cas des traces réelles et des traces synthétiques. L'analyse sera globale et locale en présentant respectivement les temps de diffusion observés sur l'ensemble des jeux de données et les temps de diffusion utilisateur par utilisateur. Le temps de diffusion que nous allons mesurer est celui nécessaire à joindre au moins la moitié du nombre total d'utilisateurs présents dans chaque situation. La mesure du temps de diffusion est faite pour chaque nœud à différents instants.
La figure 2.25 montre l'évolution de la médiane des temps de diffusion sur l'ensemble du système pour les traces de mobilité réelles avec les trois modèles de génération des graphes sociaux.

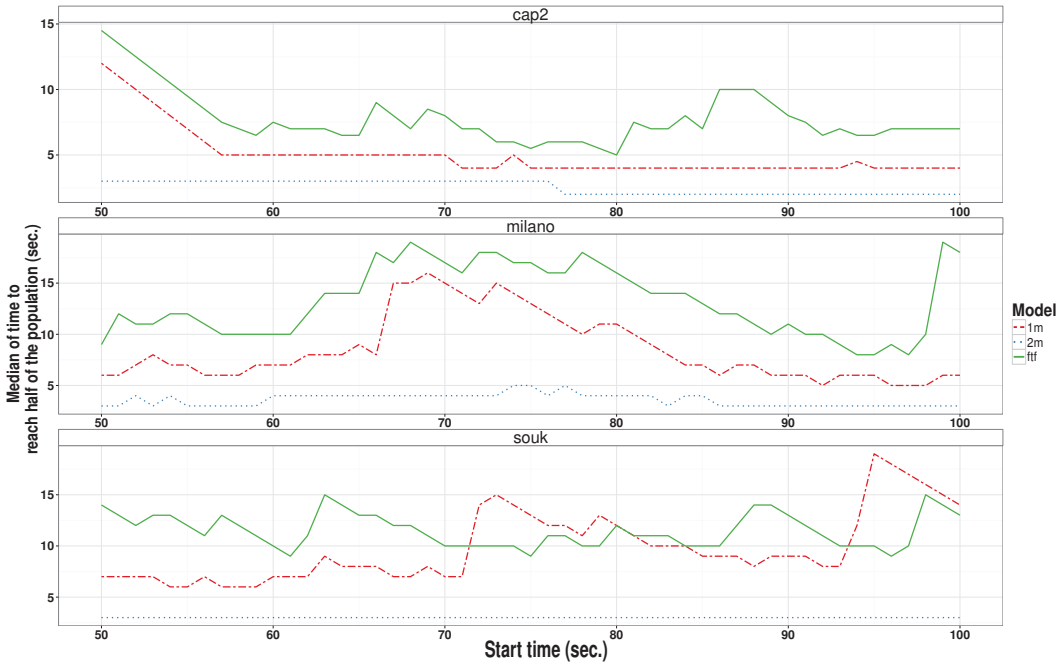


FIGURE 2.25 – Médiane du temps de diffusion dans le cas des traces de mobilité réelles.

La figure 2.26 montre l'évolution de la médiane des temps de diffusion sur l'ensemble du système pour les traces de mobilité synthétiques avec les trois modèles de génération des graphes sociaux.
On peut constater une sous estimation du temps de diffusion dans le cas des traces synthétiques par rapport aux temps de diffusion mesurés sur les traces réelles. En outre, la faible variabilité de la médiane dans le cas des traces synthétiques nous fait penser à une invariabilité par rapport au nœud source et à l'instant de début de la diffusion. Nous allons vérifier ces deux comportements dans la suite de ce paragraphe.

FIGURE 2.26 – Médiane du temps de diffusion dans le cas des traces de mobilité synthétiques.

Comportement global La figure 2.27 montre les boîtes à moustaches des temps de diffusion dans le cas des traces réelles (en rouge) et synthétiques (en bleu) après l'application du modèle de génération des contacts (dans les deux configurations, 1 m et 2 m) et du modèle de génération des interactions (ftf).

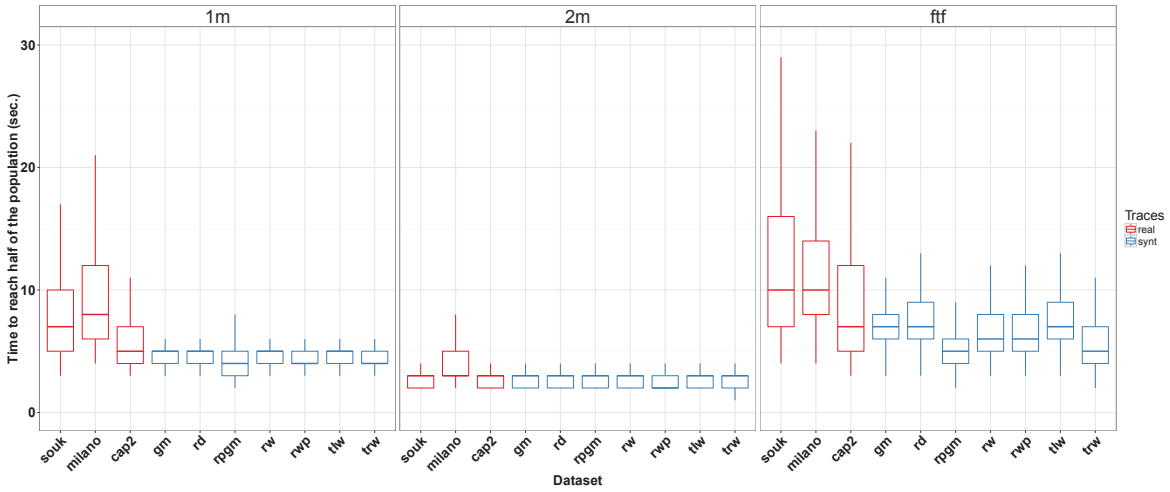
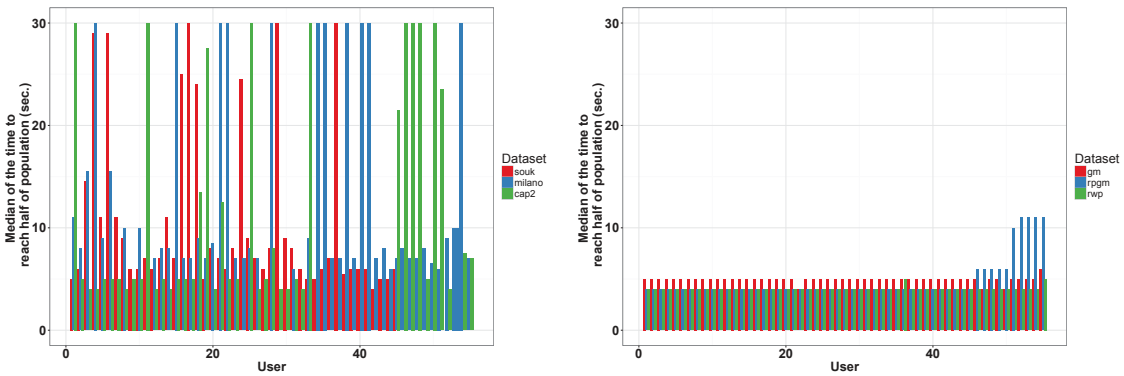


FIGURE 2.27 – Boîtes à moustaches des temps de diffusion sur la moitié des nœuds mobiles.

On remarque "l'optimisme" des modèles de mobilité par rapport aux données de mobilité réelles surtout dans le cas où les graphes sociaux sont générés à partir du modèle de génération basé sur le *clustering* (ftf). Dans les cas où le contacts sont générés à partir du modèle basé sur la distance, les performances de l'algorithme de diffusion sur l'ensemble des traces synthétiques sont similaires tandis que dans le cas des traces réelles les performances sont proportionnelles à la densité spatiale. Cette sous estimation du temps de diffusion peut avoir un gros impact sur la caractérisation du système.

Comportement local La figure 2.28 montre la médiane du temps de diffusion nécessaire à chaque nœud pour atteindre la moitié des nœuds présents dans le système. Dans la figure 2.28(a) on reporte les médianes des temps dans le cas des traces de mobilité réelles tandis que dans la figure 2.28(b) on reporte les médianes des temps pour les traces synthétiques générés à partir des modèles de mobilité Gauss-Markov (gm), *Reference Point Group* (rpgm) et *Random Waypoint* (rwp).

On remarque que dans le cas des traces synthétiques le temps de diffusion est homogène pour tous les nœuds du système avec une exception pour les temps de diffusion dans le cas du modèle de déplacement de groupe *Reference Point Group*. Dans ce cas, il y a des temps uniformes pour des sous-ensembles composés de 5 nœuds. Ce résultat met en évidence la faiblesse des modèles dans la distinction des comportements singuliers à l'intérieur du système que l'on observe dans les traces réelles.



(a) (b)

FIGURE 2.28 – Histogrammes des médianes des temps de diffusion pour les traces réelles (a) et pour trois des ensembles des traces synthétiques (b).

2.5 Conclusion

Dans ce long chapitre nous avons montré comment des mesures concernant les propriétés spatiales et sociales sur des traces de mobilité réelles, issues d'un système de captation à haute precision, nous permettent une paramétrisation réaliste dans l'utilisation de sept différents modèles de mobilité.
La comparaison entre les traces synthétiques ainsi générées et les traces réelles nous permet d'estimer l'éloignement entre les modèles et la réalité qu'ils sont censés représenter. Les résultats sur l'analyse des propriétés sociales et sur les performances de l'algorithme de diffusion, montrent bien l'effet de la négligence de l'aspect de la mobilité sociale de la parte des modèles de mobilité.
Dans le chapitre suivant nous utiliserons les traces synthétiques pour la comparaison des résultats d'inférence obtenus sur les traces réelles par le biais d'une attaque par co-localisation.

Chapitre 3

Algorithme LOCA

Sommaire

L'utilisation des informations issues des traces de mobilité est un vaste sujet. Dans ce chapitre nous allons nous intéresser à l'impact que ces informations ont sur la protection de la vie privée des utilisateurs, en particulier l'exploitation de données de co-localisation. Récemment, les attaques par co-localisation ont fait l'objet de plusieurs études visant à démontrer l'impact que les données de localisation ont sur la protection de la vie privée des utilisateurs d'applications mobiles. Dans ce scénario le lien entre les données de localisation et celles de co-localisation est intrinsèque à l'approche.
Nous allons présenter un scénario d'attaque différent, où il est possible d'utiliser les données de co-localisation sans avoir besoin de localiser les utilisateurs. Nous montrons qu'il est possible de mettre en place une attaque utilisant la co-localisation sans connaître les positions géographiques des utilisateurs concernés par l'attaque. Dans ce but, nous allons utiliser des données de proximité qui peuvent être collectées, par exemple, à partir d'un object de captation capable de reconnaître la présence des dispositifs portables autour de lui. Nous avons choisi de nommer notre attaque Loca , *Location-Oblivious Co-location Attack*. Nous allons démontrer l'efficacité de cette attaque en utilisant des traces de mobilité réelles et synthétiques.

3.1 Contexte

La relation intrinsèque entre les données de localisation et l'inférence par co-localisation des liens sociaux entre les utilisateurs est un point de référence dans les travaux sur ce sujet. En d'autres termes, à partir de la localisation géographique des utilisateurs il est possible de générer des données de proximité physique qui vont être utilisés pour inférer la proximité sociale des même utilisateurs. En supposant que proximité physique et proximité sociale sont fortement corrélées, il est possible de générer des données de proximité en connaissant la localisation soit des utilisateurs eux mêmes, soit des dispositifs qui localisent les utilisateurs.
Dans [Bil+13], les auteurs considèrent deux sources possibles de données de proximité : (1) un attaquant malveillant est capable d'intercepter des paquets (date, RSSI et identifiant du dispositif) à partir de 37 points d'accès Wi-Fi à l'intérieur d'une région de $130\,\mathrm{m} \times 250\,\mathrm{m}$, (2) chaque dispositif est capable de collecter la liste des adresses MAC, la puissance du signal (RSSI) et la date des paquets reçus de dispositifs voisins. Dans le premier cas, il faut connaitre la position exacte des 37 points d'accès pour pouvoir en déduire la localisation de chaque utilisateur par triangulation et, en conséquence, les données de proximité entre les utilisateurs. Dans le deuxième cas, les données de proximité sont générées à partir de l'hypothèse que la distance réelle entre deux dispositifs et la puissance du signal (RSSI) sont corrélées. À partir de ces données de proximité, ils construisent le graphe social en définissant la rencontre entre deux utilisateurs comme une "interaction significative", c'est-à-dire une proximité physique pour au moins t_{min} minutes à une distance inférieure à d_{max} mètres. Ils est évident que dans le premier cas, la localisation des points d'accès, et donc celle des utilisateurs, est nécessaire pour générer les données de proximité (ou co-localisation) et implémenter une attaque par inférence des relations sociales existantes au sein du groupe d'utilisateurs. A contrario, dans le deuxième cas, les informations de proximité sont indépendantes des localisations des utilisateurs et, vu la similarité de ce travail avec notre approche, nous allons le considérer comme algorithme de comparaison (dans la suite du texte sera identifié comme *state of art* SoA) pour comparer les performances des deux algorithmes.

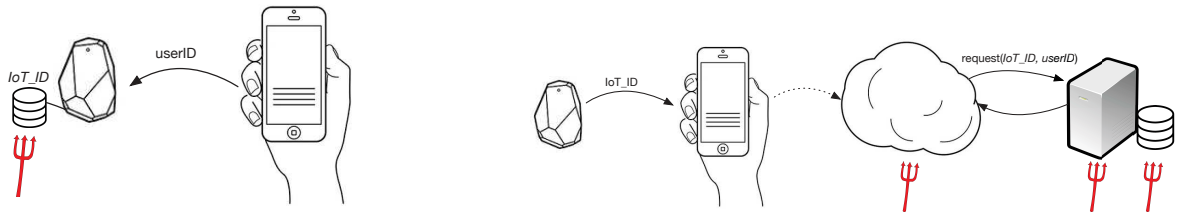
3.2 Scénario d'attaque

La diffusion massive d'objets connectés ouvre de nouveaux scénarios d'attaque sur les données manipulées par ces objets. Ces objets ne sont pas souvent localisés précisément mais ils peuvent révéler des informations de co-localisation (ou de proximité) à l'insu des utilisateurs qui les utilisent. Considérons deux exemples réels de ce scénario :

1. Les lecteurs utilisant la technologie par radio-identificatoin (*radio frequency identification*, RFID) permettent de lire des tags passifs dans un rayon d'1 m [NR06]. Les micro-contrôleurs supportant le protocole *Bluetooth Low Energy* (BLE), *e.g.* RFduino, peuvent détecter des dispositifs utilisant le même protocole jusqu'à plusieurs dizaines de metres [GOP12]. Ces dispositifs coûtent quelques dizaines d'euro. Un attaquant pourrait

donc utiliser ces dispositifs, en association avec une horloge temps réel (*real-time clock*, RTC) et une mémoire de stockage (*e.g.* une carte SD), pour un déploiement rapide (sans besoin de mesures particulières, en particulier la localisation) afin de mémoriser toutes les adresses MAC des dispositifs présents dans la zone d'attaque (voir figure 3.1 (a)).

2. Un attaquant pourrait écouter l'activité réseau générée par des utilisateurs des dispositif mobiles en relation avec des objets connectés style iBeacon (Apple) ou Eddystone (Google). Dans le cas des balises radio, il existe des produits déjà prêts à l'utilisation qui offrent une interaction avec les objets et l'environnement autour des utilisateurs [Est]. Chaque balise peut être associée à un contenu web visualisable sur les dispositifs des utilisateurs. Ces interactions entre les dispositifs portables des utilisateurs et des balises connectées (adresse IP du dispositif, url associé à la balise, date) pourraient être interceptées par un attaquant à différent niveaux du réseau (voir figure 3.1 (b)).



(a) (b)

FIGURE 3.1 – Deux scénarios d'attaque réels. (a) Objet sous le contrôle de l'attaquant, (b) Compromission de l'infrastructure de communication.

Ces deux exemples montrent comment il pourrait être possible à un attaquant de collecter des données de co-localisation sur un groupe d'utilisateurs sans avoir d'informations de localisation des utilisateurs visés par l'attaque.
Dans le cas de notre étude, nous allons implémenter un algorithme d'inférence des interactions sociales entre les utilisateurs à partir de données de co-localisation collectées sans connaître les données de localisation des utilisateurs. Autrement dit, nous allons simuler le déploiement d'un ensemble de capteurs dont on ne connaît aucune information de position, mais capables d'enregistrer les identifiants (l'identité) des dispositifs présents à leur portée. Ces capteurs nous permettront de collecter des informations de proximité que nous utiliserons pour construire une cartographie virtuelle de l'espace d'expérimentation, carte qui nous permettra de définir des trajectoires virtuelles à partir desquelles l'attaque sera menée.

3.3 Modèle et algorithme

Dans cette Section nous allons décrire la modélisation du système d'inférence, à travers la formalisation des éléments constituant le système, ainsi que l'algorithme utilisé pour la génération des trajectoires virtuelles des utilisateurs et, par la suite, du graphe social issu des données de co-localisation.

3.3.1 Modèle du système

Nous considérons qu'un attaquant a accès à une série de données de proximité issue d'un ensemble K d'objets fixes (dont on ne connait pas la localisation), sans fil et avec des capacités de captation (*e.g.* des capteurs sans fils). Les données de proximité doivent être telles que : (1) chaque utilisateur est identifiable de manière univoque à l'intérieur de la série des données de captation (*e.g* chaque utilisateur a un identifiant unique) et (2) chaque enregistrement issu de la captation contient la date en temps réel de l'événement. Comme précisé dans la section 3.2, les données de proximité peuvent provenir des différentes sources auxquelles une entité malveillante peut avoir accès.
À partir des données de proximité, l'attaquant pourra isoler les informations concernant un ensemble d'utilisateurs N pour un temps T, qui sera le temps de l'attaque et sera discrétisé pour la définition des instants $t \in \{1, \ldots, T\}$. En conséquence, la totalité des données disponibles est un ensemble de K enregistrements $(\mathsf{hist}_k)_{k=1..K}$ issus des capteurs, chacun fournissant un ensemble d'utilisateurs présents à sa proximité à chaque instant t ($\mathsf{hist}_k : [1, T] \to 2^{[1,N]}$).

3.3.2 Algorithme de génération

Pour pouvoir utiliser les données brutes hist_k, l'algorithme de transformation des données de co-localisation procède en trois étapes :

1. agrégation des données de proximité provenant de la captation pour pouvoir définir, pour chaque utilisateur, la suite temporelle des capteurs rencontrés, c'est-à-dire un vecteur de proximité avec les capteurs,
2. construction d'une cartographie virtuelle des positions des capteurs (inconnues par hypothèse) par le biais d'une matrice de transition qui mesure les flux de déplacement des utilisateurs d'un capteur à un autre,
3. définition d'une matrice d'interaction basée sur la co-localisation de chaque couple d'utilisateurs et dépendante (proportionnelle) de la matrice de transition définie auparavant.

Nous allons définir ci-dessus les trois structures des données nécessaires à l'implémentation de l'algorithme susmentionné.

Définition 3.1 (Vecteurs de proximité)
Étant donné l'ensemble d'enregistrements des capteurs $(\mathsf{hist}_k)_{k=1..K}$*, le vecteur de proximité* p_i *pour chaque utilisateur* i ($i \in 1 \ldots N$) *est défini par :*

$$p_i : \quad [1, T] \to [1, K]$$

$$t \mapsto \quad \begin{cases} k & si\ i \in \mathsf{hist}_k(t) \\ 0 & sinon. \end{cases}$$

Dans cette définition, nous considérons, sans perte de généralité, que les capteurs n'ont pas de superposition dans leur surface de détection (*i.e.* un et seulement un capteur peut enregistrer à la fois la présence d'un utilisateur). En vérité, dans la simulation de l'attaque nous prendrons en considération cette éventualité (qui a un intérêt pratique) en générant des identifiants virtuels uniques pour les surfaces de superposition de deux ou plusieurs capteurs. En autres termes, si un utilisateur $i \in \mathsf{hist}_k(t) \wedge i \in \mathsf{hist}_{k'}(t)$, alors $p_i(t) = kk^{'}$. On peut considérer le vecteur de proximité comme la trajectoire virtuelle d'un utilisateur dans un espace où les repères spatiaux sont définis par les capteurs présents.

À partir des vecteurs de proximité, il est possible de définir la matrice de transition qui capture les transitions des utilisateurs entre un capteur et un autre. La matrice sera identifiée comme matrice R dans la suite du chapitre. Un élément de la matrice $r_{k,k'}$ mesure le flux de déplacement entre le capteur k et k'.

Définition 3.2 (Matrice de transition)
Pour chaque couple $(k, k') \in [1, K]^2$*, l'élément* $r_{k,k'}$ *de la matrice de transition* R *est*

$$r_{k,k'} = \frac{1}{\sigma_k} \sum_{i\in[1,N]} |\{(i,t,t') : p_i(t) = k \wedge p_i(t') = k' \wedge p_i(t'') = 0, \forall t'' \in]t, t'[\}|$$

$$\textit{avec} \qquad \sigma_k = \sum_{i\in[1,N],t\in[1,T]} \delta_{k,p_i(t)},$$

où δ *est le delta de Kronecker.*

En raison des contraintes physiques du système (la dispersion spatiale des capteurs), la matrice de transition est une matrice creuse et, évidemment, non symétrique. Le flux de déplacement entre un capteur et un autre peut indirectement donner une estimation de la distance réelle entre les capteurs en question. On appellera cette estimation distance virtuelle. Pour quantifier les interactions entre les utilisateurs et évaluer leur profil de co-localisation, nous allons mesurer la similarité entre leurs vecteurs de proximité (un vecteur de proximité contient la suite temporelle des capteurs rencontrés par un utilisateur dans un espace temporel discret d'instants $t \in [1, \ldots, T]$), c'est-à-dire la similarité entre les trajectoires virtuelles des utilisateurs. Pour cet objectif, nous définissons un indice de similarité (*score*) :

$$\forall (i,j) \in [1,N]^2, \quad score\,(i,j) = \sum_{t\in[1,T]} r_{p_i(t),p_j(t)}. \tag{3.1}$$

La définition de l'indice de similarité est donc basée sur les valeurs de la matrice de transition précédemment définie. En d'autres termes, la similarité entre les trajectoires virtuelles des utilisateurs est proportionnelle à la distance virtuelle qu'il y a entre leurs positions (dans un espace où les seuls repères spatiaux sont les capteurs).
Les interactions (les événements de co-localisation) entre les utilisateurs seront résumées dans la matrice d'interaction, nommée matrice M. Par construction, elle est symétrique et contient tous les événements de co-localisation résultant de l'inférence.

Définition 3.3 (Matrice d'interaction)
L'élément $m_{i,j}$ de la matrice d'interaction M est défini par :

$$m_{i,j} = m_{j,i} = \begin{cases} score\,(i,j) & si\ i \neq j. \\ 0 & si\ i = j. \end{cases} \tag{3.2}$$

D'après la définition de l'algorithme d'inférence et, plus particulièrement, des structures de données nécessaires pour mener l'attaque, on entrevoit que LOCA utilise les données de proximité brutes issues des capteurs de deux façons complémentaires. Premièrement, elles vont servir pour la définition d'une cartographie virtuelle de l'espace de captation (d'expérimentation, d'attaque). Deuxièmement, elles vont nous permettre de quantifier les interactions (co-localisation) entre les utilisateurs. L'algorithme complet, avec les détails d'implémentation des différentes structures de données, est décrit dans l'algorithme 1.

3.4 Résultats expérimentaux

Dans cette section nous allons présenter les résultats expérimentaux en utilisant des traces de mobilité réelles et des traces de mobilité synthétiques. Après avoir présenté les traces de mobilité, nous allons décrire la méthode de l'attaque et son évaluation, avant d'illustrer les résultats numériques.

3.4.1 Traces de mobilité

Les traces de mobilité que nous allons utiliser dans l'évaluation de notre algorithme d'inférence ont été partiellement présentées dans le chapitre 2.
En particulier, les traces synthétiques sont générées à partir du simulateur `pymobility`, en utilisant les sept modèles de mobilité présentés dans la section 1.3, avec deux configurations spatiales, *i.e.* deux densités différentes. Les configurations sont décrites dans le tableau 3.1.

	Dimensions $[m]$		Nœuds			Durée
Nom	Largeur	Longueur	N	v_{min}	V_{max}	T
HD	10	10	40	$0.01\,\mathrm{m/s}$	$0.5\,\mathrm{m/s}$	1000
LD	30	30	40	$0.01\,\mathrm{m/s}$	$0.5\,\mathrm{m/s}$	1000

Tableau 3.1 – Paramètres de génération pour le simulateur `pymobility`.

La configuration nommée *HD* (*high density*) indique une haute densité spatiale, c'est-à-dire, 40 agents mobiles dans $100\,\mathrm{m}^2$, tandis que *LD* (*low density*) fait référence à un scénario avec 40 agents mobiles dans $900\,\mathrm{m}^2$. Les deux configurations nous donneront la possibilité d'évaluer la robustesse de notre algorithme par rapport à la variation de la densité spatiale.

Algorithm 1 LOCA algorithm to compute interaction matrix from sensors logs

Require: K sensor logs hist_k ▷ For brevity, the case of overlapping sensors/virtual sensors creation is omitted

Ensure: return the interaction matrix M

function COMPUTESIMILARITYMATRIX(hist_k)
  `int` $P[1 \dots N, 1 \dots T] = \{\{0\}\}$ ▷ Proximity logs for users, initially null
  `int` $R[1 \dots K, 1 \dots K] = \{\{0\}\}$ ▷ Transition matrix on sensors, initially null
  `int` $M[1 \dots N, 1 \dots N] = \{\{0\}\}$ ▷ Interaction matrix between users, initially null

  for $t = 1 \dots T$ **do** ▷ *First step* : compute proximity logs
    for $k = 1 \dots K$ **do**
      for all $i \in \mathsf{hist}_k(t)$ **do**
        $P[i, t] \leftarrow k$
      end for
    end for
  end for

  for $i = 1 \dots N$ **do** ▷ *Second step* : compute transitions
    for $t = 1 \dots T - 1$ **do**
      if $P[i, t] \neq 0$ **then** ▷ User i is detected at time t :
        if $\{t' > t, P[i, t'] \neq 0\} \neq \emptyset$ **then** ▷ If user i is detected after t,
          let $t_i = \min\{t' > t, P[i, t'] \neq 0\}$
          $R[P[i, t], P[i, t_i]] \leftarrow R[P[i, t], P[i, t_i]] + 1$ ▷ then update transition matrix accordingly.
        end if
      end if
    end for
  end for
  for $k, k' = 1 \dots K$ **do**
    $R[k, k'] \leftarrow R[k, k'] / \sum_{t \in T} |\mathsf{hist}_k(t)|$ ▷ Normalize transitions
  end for

  for $i = 1 \dots N$ **do**
    for $j = i + 1 \dots N$ **do**
      $score \leftarrow 0$
      for $t = 1 \dots T$ **do**
        $score \leftarrow score + R[P[i, t], P[j, t]]$ ▷ *Third step* : compute interactions based on similarity
      end for
      $M[i, j] \leftarrow score$
      $M[j, i] \leftarrow score$
    end for
  end for
  return M ▷ Output interaction matrix
end function

En ce qui concerne les traces réelles, nous allons utiliser trois jeux de données différents, deux issus des expérimentations présentées dans la section 2.1, notamment SOUK et MILANO, et un issu d'un projet mené par le MIT MediaLab [Don+12]. Ce dernier jeu de données contient les traces de mobilité de 39 employés équipés avec des badges sociométriques (*sociometric badges*) à l'intérieur d'un bureau *open space* pendant un mois. Le jeu de données contient aussi les informations sur les réelles interactions sociales entre les employés, notamment toutes les données de proximité *face-to-face* détectées par les badges dont les employés sont équipés. Vu que le nombre d'employés présents est très variable pendant tout le long de l'expérience, nous allons considérer pour notre étude seulement les données concernant les jours où il y a au moins 20 employés présents.

3.4.1.1 Vérité de terrain

Pour pouvoir utiliser un jeu de données dans une attaque par inférence, il y a nécessairement besoin qu'il contienne des informations par rapport à la vérité de terrain (*ground truth*) à partir de laquelle il sera possible de vérifier la précision de l'inférence. Dans les cas d'une attaque par co-localisation, nous avons besoin de connaitre le nombre réel d'interactions sociales entre les utilisateurs.
Dans le cas des traces synthétiques, nous allons utiliser le modèle de contact basé sur la distance présenté dans la section 2.2.5.1, tandis que pour les traces réelles issues de nos expérimentations, le modèle de contact utilisé sera celui basé sur le *clustering* et présenté dans la même section. La vérité de terrain, c'est-à-dire les interactions sociales réelles des utilisateurs dans le système, sera représentée par une matrice symétrique, à diagonale nulle, G, où chaque élément $g_{i,j}$ est l'estimation (la somme) des interactions sociales entre l'utilisateur i et l'utilisateur j.

3.4.2 Stratégie d'attaque et évaluation

Pour chaque jeu de données, nous allons simuler le déploiement de K capteurs dans l'espace expérimental. Les données de proximité issues des capteurs sont les seules informations en entrée de notre algorithme d'inférence. Aucune information concernant la position des capteurs n'est disponible pour l'attaquant. Pour évaluer les performances de l'attaque, nous allons comparer la matrice d'interaction M, résultat de l'algorithme Loca , avec la matrice G dérivée des informations sur les contacts (interactions) sociaux des utilisateurs. Pour évaluer la précision de notre algorithme nous utiliserons deux métriques :

1. une inférence globale sur le réseau social concernant tous les utilisateurs,
2. une inférence locale sur les meilleurs contacts pour chaque utilisateur.

Inférence globale L'objectif est de mesurer l'inférence sur le réseau social global dans le sens où nous allons déduire les plus importants liens sociaux. On considérera seulement

les K_{in} couples (i, j) qui ont le plus grand nombre d'interactions contenues dans la matrice G. Nous utiliserons l'analyse par la fonction d'efficacité du récepteur (*Receiver Operating Characteristic*, courbe ROC) afin de comparer les différentes configurations de notre attaque, en représentant l'évolution du taux des vrais positifs (dans notre cas, la prédiction des K_{in} couples avec le plus d'interactions) par rapport à l'évolution du taux des faux positifs (dans notre cas, la prédiction des couples qui ne sont pas dans le K_{in} réel avec le plus d'interactions). La *Area Under the Receiver Operating Characteristic* (AUC) résumera les résultats obtenus avec la ROC. Les valeurs de la AUC peuvent être interprétées comme la probabilité que LOCA assigne une grande valeur (*high score*, c'est-à-dire que LOCA prédit comme positif) à un couple choisi aléatoirement parmi ceux contenus dans le K_{in} les plus significatifs (plus d'interactions sociales), par rapport à un autre choisi aléatoirement parmi ceux qui ne sont pas parmi les K_{in} les plus significatifs. Cette méthode est souvent utilisée dans l'analyse par classification afin d'identifier lequel des modèles étudiés a les meilleures capacités prédictives. Nous allons choisir $K_{in} = 3N$, où N est le nombre d'utilisateurs dans le système, pour obtenir un réseau social avec un degré moyen de 3. Malgré tous, on observera une grande différence parmi les degrés des différents nœuds dans les cas de traces réelles à cause des différentes activités sociales (différents rôles, participants contre serveurs, bavards contre timides).

Inférence locale Pour chaque nœud $i \in [1, N]$, nous allons mesurer la taille de l'intersection entre les 10 meilleurs contacts dans G et dans M. L'algorithme LOCA doit être très robuste par rapport à la variation de sociabilité des utilisateurs réels pour réussir à donner un bon résultat à cette mesure. Le test ne considère pas la distorsion existante entre les individus les plus sociables (qui auront surement 10 contacts avec des valeurs significatives) et les individus le plus solitaires (qui probablement auront moins de 10 contacts significatifs).

La précision de l'attaque dépendra aussi du nombre de capteurs K déployés et de leur portée. Ces deux paramètres vont définir la surface totale couverte par la captation.
Intuitivement, plus la surface d'expérimentation sera couverte et plus l'algorithme d'inférence sera précis (plus la trajectoire virtuelle, issue des données de proximité, sera proche de la trajectoire réelle). Un autre paramètre important, à tenir en considération par rapport à la précision de l'attaque, est la densité spatiale des utilisateurs. Les capteurs devraient être placés dans les endroits les plus denses en population pour capturer le maximum d'informations de proximité. L'attaquant n'a pas forcement le contrôle sur le placement des capteurs. Pour évaluer la précision de l'attaque par rapport à différentes configurations de déploiement spatiale, nous allons simuler diverses stratégies de placement :

- *Grid* : les capteurs sont placés dans une grille régulière à partir du centre géométrique de l'espace d'expérimentation.
- *Density* : les capteurs sont placés en ayant une information a priori concernant les sous-zones à haute densité spatiale (*e.g.* des zones de rassemblement ou de passage, des zones où il est fortement probable que des interactions sociales surviennent). Nous verrons que cette stratégie de déploiement sera la plus efficace, c'est-à-dire nous donnera une borne supérieure pour l'évaluation de la précision de l'inférence. Les informations

concernant la densité spatiale des utilisateurs dans un certain contexte peuvent provenir de plusieurs sources, comme, par exemple, des observations précédentes au déploiement ou l'identification des points névralgiques d'un certain environnement (*e.g.* un restaurant à l'intérieur d'un campus plutôt qu'un parking dans le même campus).
- *Spiral* : les capteurs sont placés à partir du centre géométrique de l'espace d'expérimentation en suivant une spirale, dans le sens inverse aux aiguilles d'une montre, qui s'étend jusqu'à la périphérie de l'espace de simulation.
- *Random* : les capteurs sont placés aléatoirement dans la surface d'expérimentation.

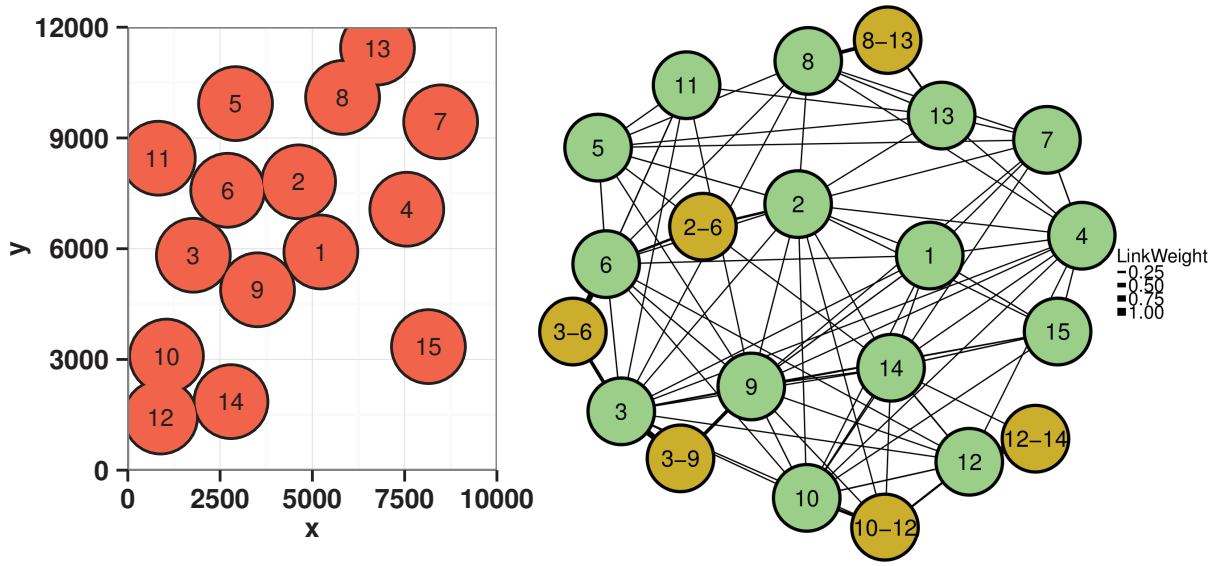
À noter que l'impact d'un mauvais placement des capteurs est double : (1) les vecteurs de proximité pourraient être très creux et donc donner lieu à des valeurs de similarité entre les trajectoires virtuelles (matrice d'interaction) non significatives, (2) la relation entre la distance virtuelle et la distance réelle pourrait être biaisée à cause d'une faible détection des flux de déplacement entre un capteur et un autre.
Le dernier paramètre significatif pour l'évaluation de notre attaque est le rayon de portée des capteurs. Un rayon de 50 cm est comparable à la portée d'un dispositif type iBeacon ou RFID, tandis que 100 cm peut être la portée représentative d'un dispositif utilisant le protocole BLE (*Bluetooth Low Energy*). La variation de la portée de captation sera examinée par la suite pour identifier la relation que ce paramètre a vis à vis des autres paramètres impactant le système d'attaque.

3.4.3 Cartographie virtuelle

Un point central de notre approche est l'estimation de la distance réelle entre les différents capteurs présents. La matrice de transition R est le résultat de cette estimation à partir des flux d'utilisateurs qui se déplacent d'un capteur à un autre. Implicitement, nous considérons que le nombre d'utilisateurs en déplacement entre un capteur et un autre est directement proportionnel à leur distance. La figure 3.2 montre la corrélation entre la position réelle des capteurs (figure 3.2 (a)) et la cartographie virtuelle (figure 3.2 (b)), dérivée de la visualisation du graphe issu de la matrice de transition R en utilisant l'algorithme de Fruchterman-Reingold.

D'après l'analyse des images dans la figure 3.2, nous pouvons remarquer que :
- les capteurs qui sont physiquement proches sont bien représentés par des valeurs élevées dans la matrice de transition, valeurs représentées par l'épaisseur des liens concernés. L'intuition selon laquelle la distance physique entre deux capteurs peut être mise en relation proportionnelle aux flux d'utilisateurs se déplaçant entre les deux capteurs est vérifiée.
- la création des capteurs virtuels, due à l'intersection de leur portée (surface de captation) est bien implémentée dans la génération de la matrice de transition.

La bonne construction de cette cartographie virtuelle (*i.e.* de la matrice de transition R) sera à la base de la réussite de notre attaque par inférence.



(a) Déploiement réel en configuration *Density*

(b) Visualisation de la carte virtuelle

FIGURE 3.2 – (a) Déploiement en suivant la stratégie *Density* de 15 capteurs dans l'espace d'expérimentation de l'expérience SOUK. (b) Représentation sous forme de graphe pondéré de la matrice de transition R par le biais de l'algorithme basé sur les forces de Frunchterman-Reingold. Les capteurs portant un identifiant de type $(x - y)$ sont les capteurs virtuels créés par l'intersection de deux capteurs réels.

3.4.4 Résultats

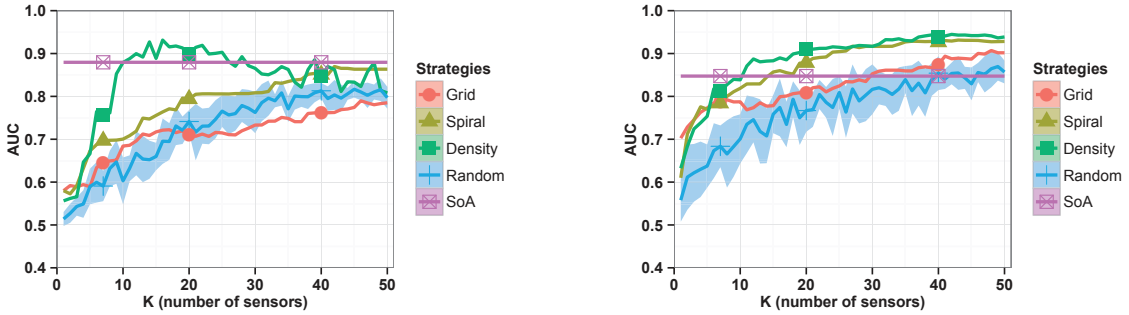
Dans cette section nous allons présenter les résultats numériques de notre attaque, dans le cas de l'inférence globale sur le réseau social composé par l'ensemble des utilisateurs, et dans le cas de l'inférence locale en essayant de déduire pour chaque utilisateur la liste de ses meilleurs contacts. Les résultats des deux inférences liées à l'implémentation de l'algorithme LOCA seront présentés dans le cas des traces de mobilité réelles et dans le cas des traces de mobilité synthétiques. Nous allons simuler un déploiement de K capteurs de 1 m de portée.

3.4.4.1 Inférence globale

L'objectif est d'inférer les K_{in} couples d'utilisateurs qui ont le plus grand nombre d'interactions dans le réseau social global. La valeur de référence sera $K_{in} = 3N$, avec N nombre total d'utilisateurs dans chaque jeu de données.

Traces réelles La figure 3.3 présente les résultats de l'inférence globale dans le cas des traces de mobilité réelles. Pour les deux jeux de données issus des expérimentations présentées dans le chapitre 2 et nommés SOUK et MILANO, nous avons calculé la AUC en considérant

les différentes stratégies de déploiement et un nombre K variable de capteurs déployés. En conséquence, si on considère $K = 15$ capteurs dans les données SOUK, la valeur de la AUC résultant de l'inférence est de 66% pour le déploiement *Random*, 72% pour la stratégie *Grid*, 76% pour la stratégie *Spiral* et 89% pour le déploiement basée sur la densité spatiale *Density*.



(a) SOUK (b) MILANO

FIGURE 3.3 – Visualisation de la AUC pour (a) SOUK et (b) MILANO dans les différents configuration de déploiement.

Pour ce qui concerne la méthode de déploiement *Random*, la valeur reportée est la moyenne arithmétique sur 10 tirages. La surface colorée autour de la ligne moyenne indique la déviation standard sur les 10 tirages. Dans la figure 3.3, il est possible d'observer aussi le résultat de l'algorithme de comparaison [Bil+13] (SoA) qui ne dépend pas du nombre de capteurs déployés et dont la valeur de la AUC vaut 88% et 84% respectivement dans SOUK et MILANO.

L'impact du nombre de capteurs K est évident : l'augmentation du nombre de capteurs déployés (plus de surface couverte) améliore la precision de l'inférence. La seule exception est pour les données SOUK dans le cas d'un déploiement basé sur la densité spatiale (*Density*). 16 capteurs suffisent pour atteindre la valeur maximale de AUC (93,14%), c'est-à-dire que il suffit de couvrir 11,4% de 110 m^2 pour inférer la quasi totalité des 150 meilleurs liens sociaux dans le réseau d'interaction globale. Ce résultat est supérieur à celui obtenu avec le déploiement de 110 capteurs sans superposition (valeur de la AUC de 89,2% pour 85% de surface couverte, point non visible sur la figure 3.3 (a)). On peut donc en déduire que la captation des zones à faible densité introduit du bruit dans le processus d'inférence.
La perte de précision avec un nombre trop élevé de capteurs est due aussi au bruit de captation introduit par la création des capteurs virtuels à l'intersection des surfaces de captation d'au moins deux ou plus capteurs. Ce bruit est dû au fait que la création des capteurs virtuels introduit dans la matrice de transition R des éléments (les capteurs virtuels) qui ont des faibles valeurs, c'est-à-dire petit flux de transition entre un capteur et un autre. Cette dispersion des informations de transition dans la matrice R va biaiser le calcul des similarités entre les trajectoires virtuelles et donc le calcul de la matrice d'interaction M. La précision de l'inférence globale sera l'objet du paragraphe 3.4.4.1.

Concernant les diverses stratégies de déploiement, la meilleure stratégie après celle basée sur la densité spatiale s'avère être celle où les capteurs sont placés suivant une spirale qui a son origine au centre de l'espace d'expérimentation (*Spiral*). Ce résultat montre que de nombreuses interactions sociales se produisent au centre de l'espace physique où les utilisateurs se déplacent (dans certaines condition donc, le centre de l'espace peut être classifié comme un point d'intérêt).
La figure 3.4 montre en détail la courbe ROC dans le cas des données SOUK en utilisant $K = 15$ capteurs de 1 m de portée. Il est intéressant de remarquer comment dans le cas d'un déploiement basé sur la densité spatiale (*Density*), 50% des $K_{in} = 150$ couples d'utilisateurs le plus en interaction visés par l'inférence sont tout de suite identifiés, comme dans le cas de l'algorithme de comparaison (SoA). Pour la même quantité de true positifs, les stratégies *Spiral* et *Grid* accumulent environ 12,5% de false positifs, prix à payer pour le manque d'information a priori (dans le cas spécifique, le manque d'information sur la densité spatiale). Dans le cas *Random*, il est facile d'obtenir environ 25% de vrai positifs (*i.e.* 25% des couples) avant que le taux de faux positifs ne se mette à grandir.

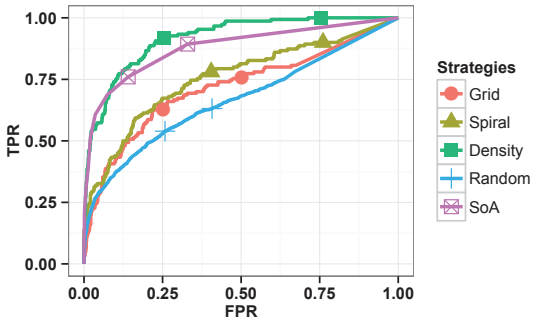


FIGURE 3.4 – Courbe ROC pour les données SOUK en utilisant 15 capteurs dans les différentes stratégies. En abscisse le taux de faux positifs (FPR) et en ordonnée le taux de vrais positifs (TPR).

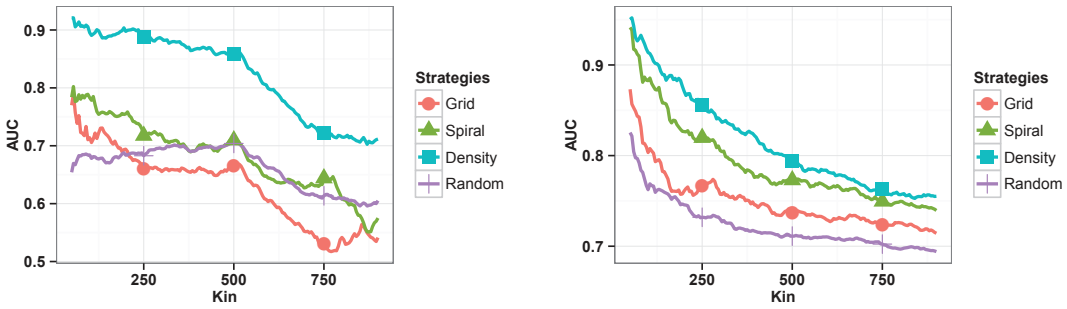
Dans le cas des données issues de l'expérience Badge [Don+12] la simulation d'inférence est une variante de celle présentée auparavant pour les données de SOUK et MILANO. Cette variation est due à la spécificité des données mises à disposition par les auteurs et au manque d'informations précises sur les conditions et les modalités de collecte des données. Nous avons choisi de simuler un déploiement des capteurs autour des points d'intérêt collectif (où la majorités des interactions sociales sont censées se produire) à l'intérieur du bureau. Les points identifiés sont les salles de réunion, la salle cafétéria et cuisine, les points d'impression et la salle de direction. Une autre différence est aussi par rapport à la portée et au nombre des capteurs, dans cet environnement nous allons simuler la présence de $K = 30$ capteurs de portée 50 cm, c'est-à-dire une couverture d'environ 7.8% de l'espace total. Le tableau 3.2 montre les valeurs de la AUC résultant de l'inférence.

Day :	4	5	6	7	8	11
Population :	24	28	31	32	29	23
AUC :	0.776	0.740	0.683	0.665	0.620	0.588

Tableau 3.2 – Valeurs de la AUC pour le données Badge.

Ces résultats montrent une grande variabilité de précision pour l'algorithme LOCA due à la mobilité réduite et au moindre nombre d'interactions sociales que peuvent avoir des employés dans un bureau par rapport à des participants à un événement social comme c'était le cas dans SOUK et MILANO. Ces changements peuvent justifier les faibles résultats d'inférence dans le jours 8 et 11. On rappelle au lecteur que les jours d'expérience choisis pour l'inférence sont ceux où il y a plus de 20 employés présents dans le bureau et que chaque jour est indépendant d'un autre.

Précision de l'inférence globale Pour évaluer la précision de l'algorithme LOCA sur l'inférence globale, nous allons étudier l'impact du nombre de couples visés par l'attaque, *i.e.* K_{in}, et la portée des capteurs utilisés pour la captation.
La figure 3.5 montre la variation de la valeur de la AUC en fonction de la valeur de K_{in}. On rappelle que le nombre total de participants N est de 45 et 64, respectivement pour SOUK et MILANO. En conséquence, il y a $\frac{N(N-1)}{2}$ couples, c'est-à-dire 990 pour SOUK et 2016 pour MILANO, et donc pour $K_{in} = 250$ l'inférence va viser respectivement 25% et 12% des meilleurs liens sociaux issus des interactions sociales tout le long de l'expérience.



(a) SOUK (b) MILANO

FIGURE 3.5 – L'impact de K_{in} dans les données SOUK (a) et MILANO (b), avec un déploiement de 15 capteurs de portée $1m$.

On remarque que pour K_{in} croissant, les performances de l'algorithme diminuent en gardant l'ordre des résultats pour les diverses stratégies, *i.e.* la meilleure stratégie de déploiement est celle basée sur la densité spatiale, suivie par le placement à spirale et ainsi de suite jusqu'au placement aléatoire qui, dans les données de l'expérience SOUK et pour K_{in} suffisamment élevé, a des performances comparables aux deux stratégies de placement "géométriques" (sans besoin d'information a priori).

La figure 3.6 montre la variation des performances de LOCA dans l'inférence globale (toujours en terme de AUC) par rapport à la variation de la portée de 15 capteurs fixes, uniformément distribués dans l'espace d'expérimentation. À cause de la différence de surface dans les deux expériences, la valeur maximale pour la AUC est atteinte à deux valeurs différentes de portée. Dans le cas des données SOUK, la AUC maximale est de 93% pour des capteurs de 2.6 m de rayon, tandis que pour les données de MILANO, la AUC est de 86% pour un rayon de 1.7 m. La perte de performance est liée à la création des capteurs virtuels. Dans SOUK la densité spatiale est importante (les interactions sociales sont plus "condensées" que dans MILANO) et donc LOCA est plus robuste par rapport à la "dispersion" d'informations due à la création des capteurs virtuels. La création des capteurs virtuels va introduire des faibles valeurs dans la matrice de transition qui est la composante principale pour le calcul de la matrice d'interaction et donc à la base des résultats de l'inférence.

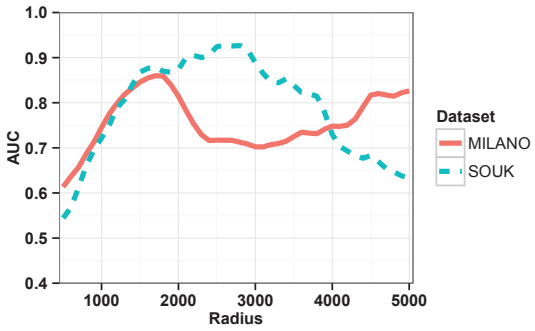


FIGURE 3.6 – Impacte de la porté de 15 capteurs uniformément distribué dans l'espace.

Traces synthétiques Le tableau 3.3 résume les résultats de l'inférence sur les traces synthétiques générées à partir du simulateur `pymobility`. Il n'est pas surprenant que dans la configuration plus dense (HD) on obtient les meilleurs résultats, l'espace de simulation est 9 fois plus petit que celui de la configuration à faible densité (LD) et donc, en conséquence, la surface couverte par la captation proportionnellement plus importante. Encore une fois, l'impact que la bonne construction de la matrice de transition a sur le succès de l'attaque est remarquable.

Nous pouvons remarquer que dans le cas des modèles de mobilité de groupe (*i.e.* trw et rpgm) qui considèrent les déplacements de chaque nœud par rapport aux déplacements des autres nœuds faisant partie du même groupe et qui considèrent donc un lien d'appartenance (pas vraiment d'interaction) entre différents nœuds, l'inférence est plus efficace, même avec une stratégie de déploiement aléatoire (*Random*).

3.4.4.2 Inférence locale

L'objectif est d'inférer les 10 meilleurs contacts sociaux pour chaque utilisateur présent dans le système. Cette inférence sera menée sur les deux jeux de données issus de nos expériences pratiques, SOUK et MILANO.

		Strategy			
Model	Config	Grid	Spiral	Density	Random
gm	*HD*	0.651	0.719	0.790	0.703
gm	*LD*	0.523	0.540	0.553	0.551
rd	*HD*	0.663	0.738	0.840	0.752
rd	*LD*	0.529	0.553	0.555	0.557
rpgm	*HD*	0.834	0.872	0.922	0.870
rpgm	*LD*	0.696	0.700	0.798	0.762
rw	*HD*	0.696	0.766	0.916	0.820
rw	*LD*	0.524	0.545	0.722	0.658
rwp	*HD*	0.685	0.735	0.785	0.698
rwp	*LD*	0.557	0.604	0.580	0.547
tlw	*HD*	0.665	0.723	0.850	0.753
tlw	*LD*	0.520	0.534	0.600	0.561
trw	*HD*	0.837	0.874	0.898	0.860
trw	*LD*	0.650	0.685	0.770	0.748

Tableau 3.3 – Résultats de l'inférence globale sur les traces synthétiques en utilisant des capteurs de porté 1 m.

La figure 3.7 montre la probabilité de l'inférence dans une configuration de déploiement basée sur la densité spatiale (*Density*) et sur la spirale (*Spiral*) de 15 capteurs de portée de 1 m. Contre-intuitivement (au vu du résultat de l'inférence globale), l'inférence locale donne des meilleurs résultats sur les données SOUK plutôt que MILANO. Naturellement, ce résultat ne contredit pas celui global vu la différence entre les deux objectifs de l'attaque : l'inférence des meilleurs contacts sociaux contenus dans le réseau social global n'est pas directement liée à l'inférence des meilleurs contacts pour chaque utilisateur (dans l'inférence globale la différence de sociabilité entre les différents utilisateurs n'est pas déterminante). Le meilleur résultat de SOUK par rapport à MILANO est aussi lié à la variabilité de présences au cours de l'expérience : dans SOUK le nombre de participant est plus stable que dans MILANO, donc les interactions sociales (les contacts) peuvent se consolider pour donner suffisamment d'information à l'algorithme pour l'inférence. Cependant, ces résultats montrent une bonne précision de l'algorithme Loca dans le cas des données éparpillées dans l'espace et dans le temps (*sparse*) : nous pouvons détecter correctement en moyenne 6.5 sur 10 des meilleurs contacts pour chaque utilisateur de l'expérience SOUK en utilisant 15 capteurs de 1 m de portée déployés selon la densité spatiale.

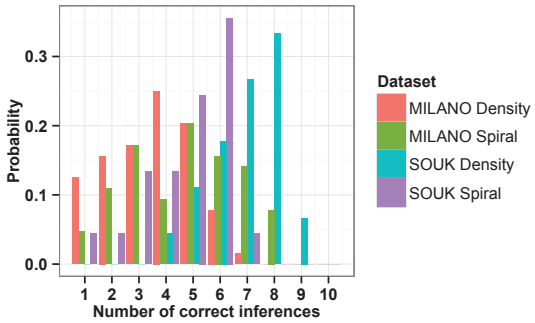


Figure 3.7 – Probabilité d'avoir exactement x amis sur 10 identifiés par Loca en utilisant 15 capteurs de 1 m de porté.

3.4.4.3 Conclusion

Dans cette section concernant les résultats de l'implémentation de l'algorithme d'inférence LOCA sur différents jeux de données, nous avons mis en évidence les facteurs qui peuvent conditionner la précision de notre attaque, notamment le nombre total de capteurs déployés et leur portée ainsi que la stratégie utilisée pour leur déploiement. Avoir une information a priori, comme une estimation de la densité spatiale, améliore nettement les résultats de l'inférence dans le cas global comme dans le cas local. En outre, sans informations a priori, un déploiement géométrique ordonné à partir du centre de l'espace (*e.g.* Spiral) est plus performant qu'un déploiement aléatoire.

3.5 Contre mesures

Nous allons brièvement illustrer ici des possibles stratégies pour réduire l'impact de ce type d'attaque sur l'inférence de la vie privée des utilisateurs. LOCA étant une attaque par co-localisation, toutes les stratégies déjà élaborées pour ce type d'attaque sont aussi valables. L'application d'une *mix-zone* [BS03] entre deux zones de captation pourra détériorer l'estimation de la distance entre deux capteurs et, en conséquence, la fiabilité de la matrice de transition. Dans le même objectif, l'utilisation d'un système des pseudonymes pourrait aider les utilisateurs à protéger leur vie privée, à condition que le service fourni par le système de captation n'ait pas impérativement besoin de l'identité des utilisateurs.
Une autre stratégie pourrait être celle de limiter la surface disponible à l'attaque, c'est-à-dire masquer au réseau les informations concernant l'identité. Une solution possible pourrait être celle de concevoir un protocole de communication "silencieux" capable de ne pas dévoiler l'identité du dispositif même en pleine zone de captation.

Chapitre 4

Co-localisation implicite

Sommaire

Dans ce chapitre, pour compléter l'exploration des nouveaux scénarios d'attaque par co-localisation, nous allons quantifier l'impact d'une attaque par co-localisation implicite, *i.e.* des informations probabilistes sur la co-localisation, dans un scénario de macro-mobilité. À ce propos, nous allons formaliser le problème en définissant un modèle du système ciblé et un algorithme d'inférence. L'évaluation de notre approche sera faite à partir d'un jeu de données de mobilité réelle issu d'une expérience de collecte de données à partir de téléphones portables.

4.1 Problématique

Les services basés sur la localisation (*location-based services*), qui utilisent les capacités de localisation embarquées dans les dispositifs mobiles (*e.g.* GPS), sont de plus en plus diffusés. Malgré l'utilité de ces services, l'impact que l'utilisation des données de localisation a sur la protection de la vie privée des utilisateurs est amplement démontré par des récentes études ([BSM10], [CML11]). Des mécanismes de protection ont été également proposés dans la communauté scientifique, notamment basés sur l'anonymisation et ou l'obscurissement des traces de mobilité [Sho+11].
Plus récemment, la perte de protection dans la vie privée due à l'utilisation des informations de co-localisation a été quantifiée dans [Olt+14]. Les informations de co-localisation sont en effet largement disponibles pour les fournisseurs de services en ligne, *e.g.* des *tags* dans les réseaux sociaux, la détection des visages à partir des photographies téléversées en ligne ou les utilisateurs des adresses IP générées par traduction d'adresse réseau (*network address translation*, NAT).

Nous nous intéressons à des informations de co-localisation probabiliste ou implicite dans le sens où l'information de co-localisation est une valeur de probabilité rapporté à la possibilité que deux utilisateurs sont co-localisés à une certaine heure de la journée. Par exemple, savoir que "Alice et Bob sont co-localisés à midi 70% des jours de la semaine" ou "des collègues sont suivant co-localisés pendant les heures de travail" nous donne une information implicite s'attachant à la possibilité que deux ou plusieurs utilisateurs partagent la même localisation au même moment. Ce genre d'informations va donc définir la co-localisation implicite en opposition à la co-localisation explicite où on va savoir précisément que, par exemple, "Le 24 février 1969 à minuit Alice et Bob étaient ensemble au Royal Albert Hall de Londres".
Les profils probabilistes de co-localisation sont générés à partir des informations diverses issues du comportement, de l'entourage et de "l'historique" des localisations des utilisateurs ciblés. Sans perte de généralités, les profiles peuvent résumer la probabilité de co-localisation entre un couple d'utilisateurs (*e.g.* Alice et Bob) où un ensemble d'utilisateurs (*e.g.* des collègues).
À partir des profils probabilistes de co-localisation, nous allons présenter un algorithme d'inférence afin de pouvoir quantifier l'impact de ce type d'information sur la perte de protection de la vie privée des utilisateurs.

4.2 Modèle et inférence

Dans cette section nous allons formaliser le modèle de l'attaque en décrivant le scénario réel ciblé et les informations auxquelles l'adversaire a accès. Ensuite, nous allons définir l'algorithme d'inférence qui nous permettra de vérifier l'impact des informations de co-localisation sur la perte de protection de la vie privée.

4.2.1 Formalisation

Nous allons considérer un système de N utilisateurs mobiles dans une zone particulière et un adversaire, typiquement un fournisseur de services en ligne qui essaye d'inférer la localisation des utilisateurs. Nous allons modéliser la mobilité des utilisateurs dans un temps et un espace discrets, c'est-à-dire que nous allons considérer T instants et, à chaque instant de temps, un utilisateur sera localisé dans une des M régions qui composent la zone considérée. L'adversaire observe sporadiquement la localisation des utilisateurs, plus précisément, à chaque instant $t \in [0, T]$ l'adversaire mémorise les localisations d'un sous-ensemble de dimension $n \leq N$ d'utilisateurs. L'adversaire a aussi accès à des profils de localisation et de co-localisation des utilisateurs sous la forme des deux distributions probabilistes :

1. la probabilité qu'un utilisateur soit localisé dans une des M régions à une certaine heure de la journée,
2. pour chaque couple d'utilisateurs, la probabilité de co-localisation à une certaine heure de la journée.

Ces distributions probabilistes seront à la base de la définition de l'algorithme d'inférence.

4.2.2 Inférence

Nous allons définir un algorithme heuristique pour quantifier l'impact des données de co-localisation implicite dans la perte de protection de la vie privée. En autre termes, l'algorithme implémentera une attaque par inférence avec l'objectif de dévoiler la localisation de l'utilisateur ciblé u à un certain instant t. Pour essayer de prédire la bonne localisation, l'algorithme prendra en entrée les profils probabilistes de co-localisation de l'utilisateur cible et la localisation des utilisateurs v co-localisés avec u à l'instant t. La localisation sélectionnée l sera celle qui maximise la somme suivante :

$$\sum_{v \,\text{localisé dans}\, l} \mathbb{P}(u, v \,\text{sont co-localisés}).$$

Si à l'instant t l'utilisateur u n'est pas co-localisé avec un des autres utilisateurs, *i.e.* $|v| = 0$, alors la localisation prédite l sera celle la plus probable pour u à la même heure du jour. Nous allons comparer la localisation prédite l avec la localisation réelle de u à l'instant t et nous allons mesurer la précision de l'attaque en calculant la proportion des localisations correctes.
Pour pouvoir évaluer notre algorithme, nous allons définir un point de comparaison (*baseline*) avec un simple algorithme d'inférence qui n'utilise pas les informations de co-localisation. L'algorithme de comparaison va prédire systématiquement la localisation l la plus probable pour l'utilisateur u à la même heure du jour par rapport à t.
Les comportements des deux algorithmes utilisés est résumé dans l'exemple suivant :

Example

- 5 utilisateurs, A, B, C, D, E, $u = A, v \in \{B, C, D, E\}$
- L'objectif est d'inférer la localisation de $u = A$ à l'instant t
- Les localisations des utilisateurs B, C, D, E à l'instant t sont respectivement x, y, y, x
- La vraie localisation de l'utilisateur u à l'instant t est y
- La plus probable localisation de l'utilisateur u à la même heure que t est x

v	$\mathbb{P}(u, v \,\text{sont co-localisés})$	Localisation de v (L)
B	0.2	x
C	0.19	y
D	0.15	y
E	0.1	x

- La somme pour la localisation x est 0.3

❍ La somme pour la localisation y est 0.34

➪ L'algorithme heuristique prédit y ✓

➪ L'algorithme de comparaison prédit x ✗

4.3 Évaluation expérimentale

Dans cette Section nous allons présenter le jeu de données réel utilisé pour l'évaluation et les résultats de cette évaluation.

4.3.1 Données

Les données utilisées dans l'evaluation de l'attaque par inférence proviennent du jeu de données nommé Reality Mining [EPL09b] et [EP06]. Comme décrit dans la section 1.2, cette collection de données contient des données de localisation, avec les identifiants (IDs) uniques des antennes GSM auxquelles les téléphones portables sont connectés, et les données de proximité avec la liste des dispositifs (adresses MAC) à portée Bluetooth. La présence conjointe de ces deux informations concernant la mobilité des utilisateurs nous a fait choisir cette collection. Les auteurs ont rendu publiques plus de 400.000 heures des données issues des dispositifs portables concernés par l'expérience. Nous allons nous concentrer sur les $\sim$ 10 millions d'identifiants GSM reportés (*i.e.* événements de localisation) et les $\sim$ 2 millions d'adresses MAC enregistrées (*i.e.* événements de proximité).

4.3.1.1 Manipulation

Nous allons manipuler les données pour en extrapoler explicitement les événements de localisation et les événements de proximité. Pour ce qui concerne les événements de localisation, nous supposons qu'à chaque instant t la localisation de chaque utilisateur est définie par l'identifiant de l'antenne GSM à laquelle il est connecté. Plus précisément, après avoir fixé une granularité temporelle $\Delta t = 1\,\text{heure}$ pour notre intervalle discret $[0, T]$, nous allons définir la localisation d'un utilisateur à l'instant $t \in [0, T]$ comme l'identifiant de l'antenne GSM à laquelle il s'est le plus connecté pendant l'heure considérée. Après cette définition, nous sommes obligé de diviser les utilisateurs dans l'expérience par rapport à leur opérateur téléphonique car seuls les utilisateurs qui partagent le même opérateur téléphonique reporteront le même ensemble d'identifiants d'antennes GSM (voir encadré). En conséquence, nous allons sélectionner le sous-ensemble plus grande en nombre d'utilisateurs qui déclarent (dans le questionnaire contenu dans le jeu de données) avoir le même opérateur téléphonique, c'est-à-dire *T-Mobile*. Parmi ces utilisateurs, nous allons éliminer ceux qui ont des informations manquantes (*e.g.* l'adresse MAC) pour en retenir finalement 42.

Une station de transmission de base GSM est identifiée par un code unique qui a la forme MCC :MNC LAC :CID.
- MCC est le *Mobile Country Code* et dépend du pays où la station est située,
- MNC est le *Mobile Network Code* et dépend du réseau et de l'opérateur téléphonique qui utilisent la station,
- LAC est le *Local Area Code* et définit la zone où se trouve la station,
- CID est le *Cell tower ID* et identifie l'antenne à l'intérieur de la zone définie par le LAC.

La participation des ces utilisateurs à l'expérience n'est pas homogène, la Figure 4.1 reporte les périodes de participation des utilisateurs sélectionnés.

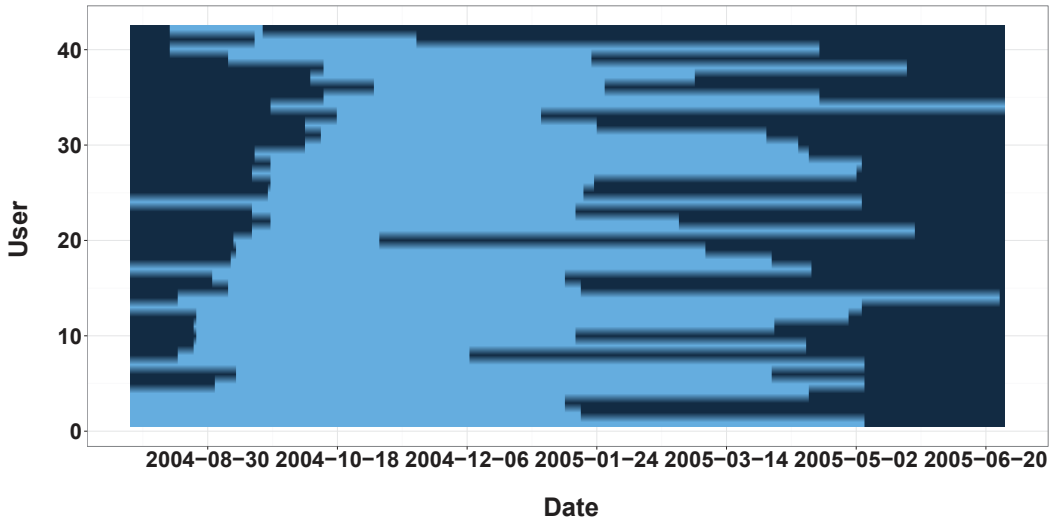


FIGURE 4.1 – Date de début et date de fin, les lignes en blue clair identifient les intervalles de participation.

Nous fixons notre intervalle $[0, T]$ comme l'intersection des tous les intervalles afin d'avoir 330 jours échantillonnés (discrétisés) avec $\Delta t = 1$heure.

4.3.1.2 Profils de co-localisation

Les profils probabilistes de co-localisation sont des motif (*patterns*) issus des comportements des utilisateurs ou générés à partir d'informations connues par rapport aux relations existantes entre les utilisateurs. Formellement, nous allons utiliser des profils probabilistes de co-localisation qui résument la probabilité de co-localisation d'un couple d'utilisateurs (u, v) à une certaine heure du jour, c'est-à-dire $\mathbb{P}(u, v \text{ sont co-localisé} \mid u \leftrightarrow_{r_k} v)$, où r_k modélise le type de relation qui existe entre u et v. Dans la suite du Chapitre nous allons générer des profils de co-localisation entre chaque couple d'utilisateurs sans définir un ensemble précis de relations (*e.g.* collègues, amis, etc.). Les événements de co-localisation qui vont nous permettre de générer les profils sont les données de proximité, *i.e.* l'ensemble des *scans* Bluetooth.

Définition 4.1 (Co-localisation)
Deux utilisateurs sont co-localisés à un certain instant de temps discret $t \in [0, T]$ si et seulement si au moins un des deux utilisateurs fait état d'au moins un événement de proximité avec l'autre utilisateur dans l'intervalle de temps considéré.

D'après cette définition, pour chaque couple d'utilisateurs nous avons schématiquement un vecteur binaire de longueur T qui résume tous les évènements de co-localisation. À partir de ces vecteurs, nous allons calculer la probabilité de co-localisation pour chaque couple d'utilisateurs à chaque heure de la journée $(00 - 23)$. La probabilité de co-localisation à une certaine heure est le rapport entre le nombre total d'événements de co-localisation à l'heure considérée et le nombre d'heures actives pour le couple dans l'intervalle total $[0, T]$. Il suffit qu'un des deux utilisateurs soit actif pour que l'heure soit considérée active. Un utilisateur est actif si et seulement si il fait état d'au moins un événement de proximité dans l'heure considérée.
La Figure 4.2 montre le profil probabiliste de co-localisation moyenné sur tous les couples d'utilisateurs considérés pour l'évaluation de l'attaque.

On remarque que, globalement, la probabilité de co-localisation croit entre 8 heures et 16 heures, puis que la tendance s'inverse pour le reste de la journée. Ce résultat met en évidence la nature des participants à l'expérience Reality Mining : tous les utilisateurs qui ont participé à la collecte sont soit des étudiants soit des travailleurs à l'intérieur du campus du MIT. La probabilité donc d'être co-localisés est concentrée pendant les heures de travail.
Pour pouvoir évaluer quantitativement l'impact des données de co-localisation implicite par l'inférence de la localisation des utilisateurs ciblés, nous allons générer des profils de co-localisation pour chaque couple d'utilisateurs.

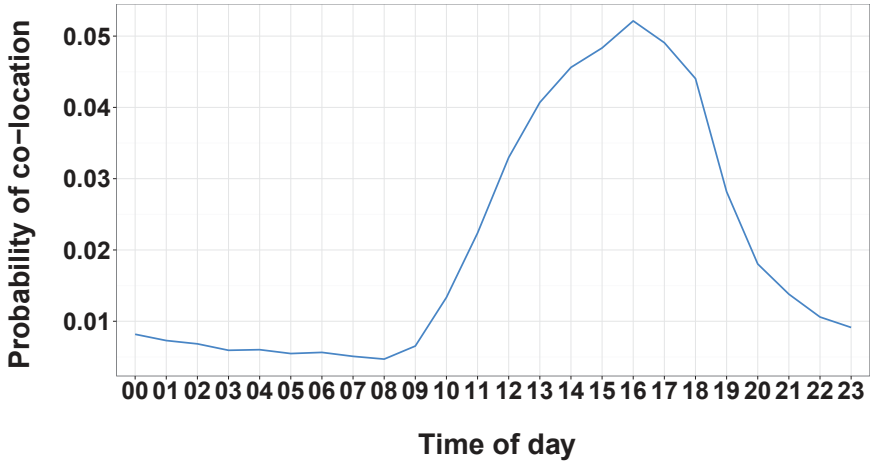


FIGURE 4.2 – Profil probabiliste de co-localisation moyenné sur l'ensemble des couples d'utilisateurs.

4.3.2 Résultats expérimentaux

En partant de l'ensemble des données de localisation des 42 utilisateurs sélectionnés, nous allons inférer toutes les localisations disponibles par le biais des deux algorithmes présentés dans la Section 4.2. Nous allons évaluer la probabilité d'inférence de chaque algorithme à chaque heure du jour comme le rapport entre le nombre des localisations inférées et le nombre de localisations reportées par tous les utilisateurs à l'heure considérée. La Figure 4.3 montre les résultats de l'inférence menée par l'algorithme heuristique et l'algorithme de comparaison (*baseline*).

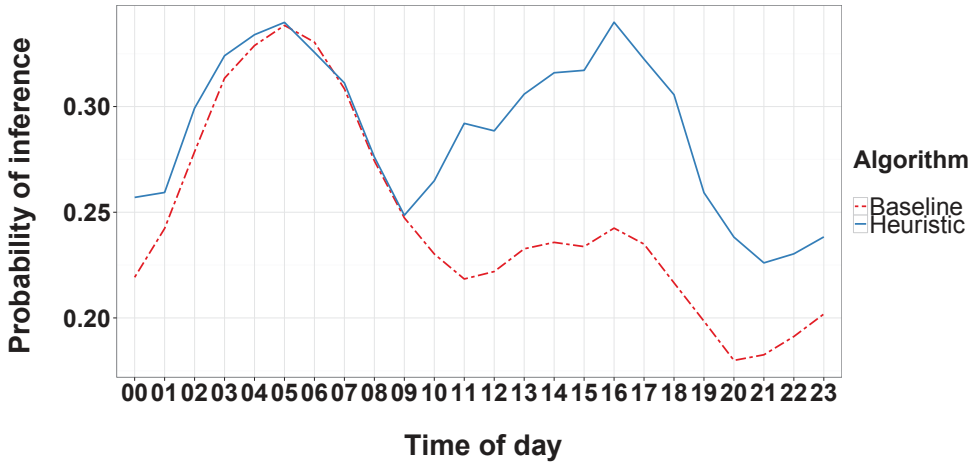


FIGURE 4.3 – Résultat de l'inférence en utilisant les deux algorithmes.

Les résultats nous confirment que l'impact des données de co-localisation est présent dans les heures de la journée où la majorité des événements de proximité sont reportés. On remarque que pendant la nuit les deux algorithmes ont des performances similaires car, en

absence d'événements de proximité pour l'utilisateur cible, les deux algorithmes ont le même comportement en prévoyant la localisation connue la plus fréquente à l'heure de l'attaque. Résultat prévisible aussi par un comportement partagé par la majorité des utilisateurs, celui de dormir la majorité du temps au même endroit : il est très probable qu'au moment de l'inférence la localisation de l'utilisateur ciblé est celle la plus fréquemment observée dans son "historique" si l'inférence a lieu entre 1 heure et 9 heures.
Pendant le reste du temps, on remarque comment la prise en considération des informations concernant la co-localisation augmente significativement le résultat de l'inférence. L'algorithme heuristique est capable d'améliorer le résultat de l'algorithme de comparaison jusqu'à 40 %.

4.4 Conclusion

Nous avons démontré l'impact des données de co-localisation implicite dans la perte de protection de la vie privée. L'algorithme heuristique présenté dans ce chapitre a mis en évidence comment l'information de co-localisation, même quand elle est considérée de manière non déterministe, peut avoir un effet significatif sur la perte de protection de la vie privée par rapport à une inférence de base menée en utilisant seulement les données "historique" de mobilité.

Conclusion et perspectives

Pour conclure, nous allons résumer les travaux présentés dans ce manuscrit, en énumérant nos contributions, avant de décrire les perspectives futures que nous envisageons pour la suite de nos travaux.

Résumé et contributions

L'évolution des systèmes mobiles en terme des capacités de communication et de captation est à la base des motivations des travaux contenus dans cette thèse. La possibilité de collecter et d'analyser des masses de données de mobilité toujours plus importantes et précises a entraîné l'étude de certains phénomènes jusqu'alors inobservables.
Comprendre les dynamiques qui régissent la mobilité humaine est un point crucial dans la caractérisation des réseaux de communication *Ad Hoc*, caractérisation qui sera efficace si les modèles de mobilité utilisés dans les simulations sont fidèles à la réalité qu'ils sont censés représenter.
La centralisation et l'analyse des masses de données de mobilité ont un fort impact sur la protection de la vie privée (numérique et non) des utilisateurs concernés. Dans un moment de l'histoire où la connectivité (les liens sociaux numériques) entre les êtres humains est au centre des activités quotidiennes de chacun de nous, nous ne pouvons plus ignorer les répercussions de la dispersion numérique de nos données personnelles vis-à-vis de notre vie privée. La massive utilisation des services en ligne ne doit pas nous faire perdre de vue le bilan service-qualité du service-bénéfice.
Nous avons étudié donc la relation existante entre la mobilité sociale et la proximité spatiale à partir des données de mobilité collectées dans différentes situations de la vie réelle.

Les contributions de cette thèse sont les suivantes :
- Une analyse fine des propriétés spatiales et sociales d'un ensemble de traces de mobilité réelles collectées expérimentalement à l'aide d'une plateforme à haute précision.
- Une comparaison exploratoire entre des traces de mobilité réelles et des traces de mobilité synthétiques générées à partir de sept modèles différents de mobilité.
- La conception et l'analyse d'un algorithme d'inférence par co-localisation décorrélée des informations sur la localisation des utilisateurs ciblés.
- La quantification du potentiel des données de co-localisation non-déterministes sur la perte de protection de la vie privée d'un ensemble d'utilisateurs.

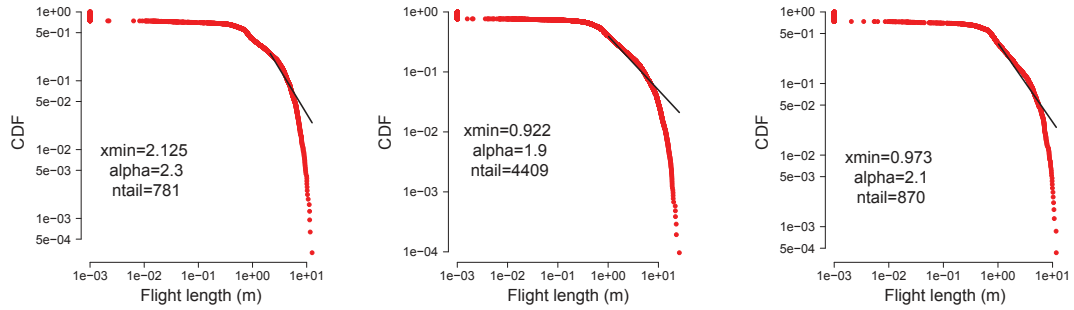
Perspectives futures

La prise en considération de l'aspect social dans la modélisation d'un système mobile basé sur un groupe d'utilisateurs dans un environnement spatialement dense pourrait permettre l'amélioration des modèles de mobilité existants. La définition de nouveaux paramètres caractérisant le comportement social, outre celui spatial, des agents mobiles dans un contexte de foule devrait aboutir à la formalisation de nouveaux modèles de mobilité qui pourront relever le défi d'une mise au point efficace dans le réseaux opportunistes. Dans ce cas spécifique, la formalisation des modèles de génération pour des graphes de connection entre les agents mobiles pourrait permettre de valoriser les contacts par rapport à la mobilité.
Le maniement des traces de mobilité humaine tout en respectant la protection de la vie privée des utilisateurs sera cruciale dans un avenir proche où la diffusion massive des objets connectés (dans une galaxie des systèmes cyber-physiques) se profile. L'augmentation du potentiel de captation de traces de mobilité contenant les informations de localisation et de proximité exigera la détermination de nouvelles stratégies de protection de la vie privée.
L'exploitation des données de co-localisation probabiliste, données basée sur l'analyse (l'extraction d'informations, *data mining*) des données hétérogènes collectées autour des utilisateurs, se place certainement dans l'univers de la science des données (*big data*). L'amélioration dans la modélisation de ces informations et la possibilité d'une analyse toujours plus efficace des données en ligne permettront une évolution dans les services délivrés par les applications mobiles.
Nous avons montré que les traces de mobilité d'un groupe d'individus contiennent des informations autres que celles liées à la simple localisation, notamment celles concernant le réseau social à l'intérieur du groupe. De nouvelles études sur les traces de mobilité pourraient mettre en évidence d'autres aspects du comportement humain contenus ou prédictibles à partir des données sur la mobilité.

Annexe A

Modélisation statistique d'une loi de puissance

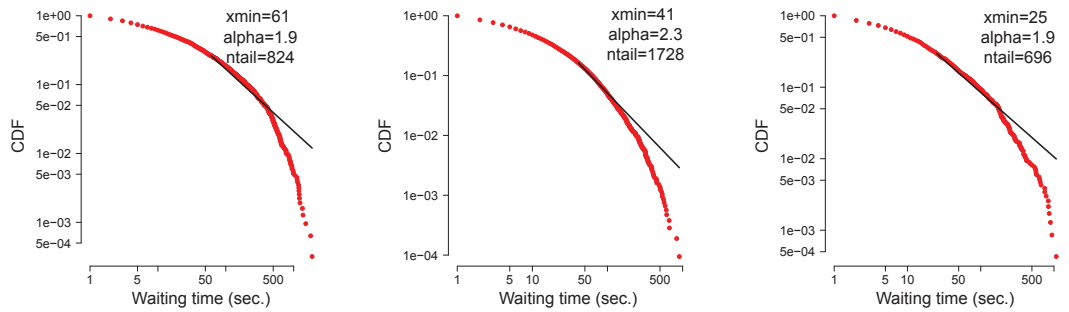
Pour la parametrisation des deux fonctions de distribution suivant des lois de puissance, nous allons utiliser la méthode presentée dans [CSN09]. La Figure A.1 montre les résultats sur les mesures de la distance franchissable (*Flight length*) dans les trois jeux de données réels. La paramétrisation est basée sur une distribution continue.



(a) (b) (c)

Figure A.1 – Paramétrisation de la loi de puissance suivie par la distance franchissable (*Flight length*) dans le cas des données (a) cap2, (b) milano, (c) souk.

La Figure A.2 montre les résultats de la paramétrisation sur les mesures du temps d'attente (*Waiting time*) dans les trois jeux de données réels. La paramétrisation est basée sur une distribution discrète.



(a) (b) (c)

FIGURE A.2 – Paramétrisation de la loi de puissance suivie par le temps d'attente (*Waiting time*) dans le cas des données (a) cap2, (b) milano, (c) souk.

Bibliographie

[Ant+12] Athanasios Antoniou, Evangelos Theodoridis, Ioannis Chatzigiannakis et Georgios Mylonas. « Human mobility trace acquisition and social interactions monitoring for business intelligence using smartphones ». Dans : *Informatics (PCI), 2012 16th Panhellenic Conference on*. IEEE. 2012, p. 1–6 (cf. p. 5).

[Arn15] Barry C Arnold. *Pareto distribution*. Wiley Online Library, 2015 (cf. p. 10).

[BB88] Pierre Bovet et Simon Benhamou. « Spatial analysis of animals' movements using a correlated random walk model ». Dans : *Journal of theoretical biology* 131.4 (1988), p. 419–433 (cf. p. 8).

[Bil+13] Igor Bilogrevic et al. « Inferring social ties in academic networks using short-range wireless communications ». Dans : *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*. ACM. 2013, p. 179–188 (cf. p. 6, 44, 54).

[BL09] Farid Benbadis et Jeremie Leguay. *CRAWDAD dataset upmc/rollernet (v. 2009-02-02)*. Downloaded from http ://crawdad.org/upmc/rollernet/20090202. Fév. 2009 (cf. p. 7).

[BR15] Claudio Bettini et Daniele Riboni. « Privacy protection in pervasive systems : state of the art and technical challenges ». Dans : *Pervasive and Mobile Computing* 17 (2015), p. 159–174 (cf. p. 12).

[Bro+14] Chloë Brown, Neal Lathia, Cecilia Mascolo, Anastasios Noulas et Vincent Blondel. « Group colocation behavior in technological social networks ». Dans : *PloS one* 9.8 (2014), e105816 (cf. p. 13).

[Bro+98] Josh Broch, David A Maltz, David B Johnson, Yih-Chun Hu et Jorjeta Jetcheva. « A performance comparison of multi-hop wireless ad hoc network routing protocols ». Dans : *Proceedings of the 4th annual ACM/IEEE international conference on Mobile computing and networking*. ACM. 1998, p. 85–97 (cf. p. 9).

[BS03] Alastair R Beresford et Frank Stajano. « Location privacy in pervasive computing ». Dans : *IEEE Pervasive computing* 1 (2003), p. 46–55 (cf. p. 12, 59).

[BSM10] Lars Backstrom, Eric Sun et Cameron Marlow. « Find me if you can : improving geographical prediction with social and spatial proximity ». Dans : *Proceedings of the 19th international conference on World wide web*. ACM. 2010, p. 61–70 (cf. p. 5, 61).

[Cat+10] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton et Alessandro Vespignani. « Dynamics of person-to-person interactions from distributed RFID sensor networks ». Dans : *PloS one* 5.7 (2010), e11596 (cf. p. 6).

[CBD02] Tracy Camp, Jeff Boleng et Vanessa Davies. « A survey of mobility models for ad hoc network research ». Dans : *Wireless communications and mobile computing* 2.5 (2002), p. 483–502 (cf. p. 8).

[Cha+07a] Augustin Chaintreau, Pan Hui, Jon Crowcroft, Christophe Diot, Richard Gass et James Scott. « Impact of human mobility on opportunistic forwarding algorithms ». Dans : *Mobile Computing, IEEE Transactions on* 6.6 (2007), p. 606–620 (cf. p. 6, 7, 10, 30).

[Cha+07b] Augustin Chaintreau, Abderrahmen Mtibaa, Laurent Massoulie et Christophe Diot. « The diameter of opportunistic mobile networks ». Dans : *Proceedings of the 2007 ACM CoNEXT conference*. ACM. 2007, p. 12 (cf. p. 5).

[CKB14] Mathieu Cunche, Mohamed-Ali Kaafar et Roksana Boreli. « Linking wireless devices using information contained in Wi-Fi probe requests ». Dans : *Pervasive and Mobile Computing* 11 (2014), p. 56–69 (cf. p. 6).

[CML11] Eunjoon Cho, Seth A Myers et Jure Leskovec. « Friendship and mobility : user movement in location-based social networks ». Dans : *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2011, p. 1082–1090 (cf. p. 7, 61).

[Coc88] John H Cochrane. « How big is the random walk in GNP ? » Dans : *The Journal of Political Economy* (1988), p. 893–920 (cf. p. 8).

[Cra+10] David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher et Jon Kleinberg. « Inferring social ties from geographic coincidences ». Dans : *Proceedings of the National Academy of Sciences* 107.52 (2010), p. 22436–22441 (cf. p. 12).

[Cro] *Crawdad : Dartmouth A Community Resource for Archiving Wireless Data At Dartmouth*. `/http://crawdad.org`. Juin 2016 (cf. p. 6).

[CSN09] Aaron Clauset, Cosma Rohilla Shalizi et Mark EJ Newman. « Power-law distributions in empirical data ». Dans : *SIAM review* 51.4 (2009), p. 661–703 (cf. p. 71).

[Dob14] Douglas Howard Dobyns. *Proximity Based Social Networking*. US Patent App. 14/523,780. 2014 (cf. p. 5).

[Don+12] Wen Dong, Daniel Olguin-Olguin, Benjamin Waber, Taemie Kim et Alex Sandy Pentland. « Mapping Organizational Dynamics with Body Sensor Networks ». Dans : *International Workshop on Wearable and Implantable Body Sensor Networks*. T. 0. Los Alamitos, CA, USA : IEEE Computer Society, 2012, p. 130–135 (cf. p. 6, 50, 55).

[Ein05] Albert Einstein. « Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen ». Dans : *Annalen der physik* 322.8 (1905), p. 549–560 (cf. p. 8).

[EP06] Nathan Eagle et Alex Pentland. « Reality mining : sensing complex social systems ». Dans : *Personal and ubiquitous computing* 10.4 (2006), p. 255–268 (cf. p. 6, 64).

[EPL09a] Nathan Eagle, Alex Sandy Pentland et David Lazer. « Inferring friendship network structure by using mobile phone data ». Dans : *Proceedings of the national academy of sciences* 106.36 (2009), p. 15274–15278 (cf. p. 6).

[EPL09b] Nathan Eagle, Alex Sandy Pentland et David Lazer. « Inferring friendship network structure by using mobile phone data ». Dans : *Proceedings of the National Academy of Sciences* 106.36 (2009), p. 15274–15278 (cf. p. 64).

[Est] *estimote, Real-world context for your apps*. `http://estimote.com/`. 2016 (cf. p. 45).

[Fal03] Kevin Fall. « A delay-tolerant network architecture for challenged internets ». Dans : *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*. ACM. 2003, p. 27–34 (cf. p. 4).

[GOP12] Carles Gomez, Joaquim Oller et Josep Paradells. « Overview and evaluation of bluetooth low energy : An emerging low-power wireless technology ». Dans : *Sensors* 12.9 (2012), p. 11734–11753 (cf. p. 44).

[GPR09a] Sabrina Gaito, Elena Pagani et Gian Paolo Rossi. « Fine-grained tracking of human mobility in dense scenarios ». Dans : *Sensor, Mesh and Ad Hoc Communications and Networks Workshops, 2009. SECON Workshops' 09. 6th Annual IEEE Communications Society Conference on*. IEEE. 2009, p. 1–3 (cf. p. 6).

[GPR09b] Sabrina Gaito, Elena Pagani et Gian Paolo Rossi. « Opportunistic forwarding in workplaces ». Dans : *Proceedings of the 2nd ACM workshop on Online social networks*. ACM. 2009, p. 55–60 (cf. p. 6).

[Hon+99] Xiaoyan Hong, Mario Gerla, Guangyu Pei et Ching-Chuan Chiang. « A group mobility model for ad hoc wireless networks ». Dans : *Proceedings of the 2nd ACM international workshop on Modeling, analysis and simulation of wireless and mobile systems*. ACM. 1999, p. 53–60 (cf. p. 12).

[Hsu+07] Wei-jen Hsu, Thrasyvoulos Spyropoulos, Konstantinos Psounis et Ahmed Helmy. « Modeling time-variant user mobility in wireless mobile networks ». Dans : *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*. IEEE. 2007, p. 758–766 (cf. p. 11).

[Hui+05] Pan Hui, Augustin Chaintreau, James Scott, Richard Gass, Jon Crowcroft et Christophe Diot. « Pocket switched networks and human mobility in conference environments ». Dans : *Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*. ACM. 2005, p. 244–251 (cf. p. 5–7).

[Ise+11] Lorenzo Isella, Juliette Stehlé, Alain Barrat, Ciro Cattuto, Jean-François Pinton et Wouter Van den Broeck. « What's in a crowd ? Analysis of face-to-face behavioral networks ». Dans : *Journal of theoretical biology* 271.1 (2011), p. 166–180 (cf. p. 6).

[JM96] David B Johnson et David A Maltz. « Dynamic source routing in ad hoc wireless networks ». Dans : *Mobile computing*. Springer, 1996, p. 153–181 (cf. p. 9).

[Kil+16a] Marc-Olivier Killijian, Roberto Pasqua, Matthieu Roy, Gilles Trédan et Christophe Zanon. « Souk : Spatial Observation of hUman Kinetics ». Dans : *Computer Networks* (2016) (cf. p. 16).

[KKK06] Minkyong Kim, David Kotz et Songkuk Kim. « Extracting a Mobility Model from Real User Traces. » Dans : *INFOCOM*. T. 6. 2006, p. 1–13 (cf. p. 15).

[LH99] Ben Liang et Zygmunt J Haas. « Predictive distance-based mobility management for PCS networks ». Dans : *INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*. T. 3. IEEE. 1999, p. 1377–1384 (cf. p. 10).

[LK16] Jure Leskovec et Andrej Krevl. *SNAP Datasets : Stanford Large Network Dataset Collection*. `http://snap.stanford.edu/data`. Juin 2016 (cf. p. 6).

[MC98] Joseph P Macker et M Scott Corson. « Mobile ad hoc networking and the IETF ». Dans : *ACM SIGMOBILE Mobile Computing and Communications Review* 2.1 (1998), p. 9–14 (cf. p. 4).

[MV05] Marvin McNett et Geoffrey M Voelker. « Access and mobility of wireless PDA users ». Dans : *ACM SIGMOBILE Mobile Computing and Communications Review* 9.2 (2005), p. 40–55 (cf. p. 6, 10).

[Nag+05] Siamäk Naghian, Tapio Lindstrom, Tero Kärkkäinen, Jarmo T Mäkinen, Keijo Lähetkangas et Kai Mustonen. *Mobile mesh Ad-Hoc networking*. US Patent 6,879,574. 2005 (cf. p. 4).

[New05] Mark EJ Newman. « Power laws, Pareto distributions and Zipf's law ». Dans : *Contemporary physics* 46.5 (2005), p. 323–351 (cf. p. 10).

[New06] Mark EJ Newman. « Modularity and community structure in networks ». Dans : *Proceedings of the national academy of sciences* 103.23 (2006), p. 8577–8582 (cf. p. 29).

[New10] Mark Newman. *Networks : an introduction*. Oxford university press, 2010 (cf. p. 27).

[Nou+09] Anastasios Noulas, Mirco Musolesi, Massimiliano Pontil et Cecilia Mascolo. « Inferring interests from mobility and social interactions ». Dans : *NIPS Workshop on Analyzing Networks and Learning with Graphs*. 2009, p. 2–88 (cf. p. 13).

[NP97] Robert M Nosofsky et Thomas J Palmeri. « An exemplar-based random walk model of speeded classification. » Dans : *Psychological review* 104.2 (1997), p. 266 (cf. p. 8).

[NR06] Pavel V Nikitin et KVS Rao. « Performance limitations of passive UHF RFID systems ». Dans : *IEEE Antennas and Propagation Society International Symposium*. T. 1011. 2006 (cf. p. 44).

[Olt+14] Alexandra-Mihaela Olteanu, Kévin Huguenin, Reza Shokri et Jean-Pierre Hubaux. « Quantifying the effect of co-location information on location privacy ». Dans : *Privacy Enhancing Technologies Symposium*. Springer. 2014, p. 184–203 (cf. p. 13, 61).

[Pan16] André Panisson. *pymobility*. 2016. URL : `https://github.com/panisson/pymobility` (visité le 01/04/2016) (cf. p. 31).

[PC11] Andrea Passarella et Marco Conti. « Characterising aggregate inter-contact times in heterogeneous opportunistic networks ». Dans : *International Conference on Research in Networking*. Springer. 2011, p. 301–313 (cf. p. 30).

[Pea05] Karl Pearson. « The problem of the random walk ». Dans : *Nature* 72 (1905), p. 342 (cf. p. 8).

[Per08] Charles E Perkins. *Ad hoc networking*. Addison-Wesley Professional, 2008 (cf. p. 4).

[PM09] Andrei Papliatseyeu et Oscar Mayora. « Mobile habits : Inferring and predicting user activities with a location-aware smartphone ». Dans : *3rd Symposium of Ubiquitous Computing and Ambient Intelligence 2008*. Springer. 2009, p. 343–352 (cf. p. 6).

[PPC06] Luciana Pelusi, Andrea Passarella et Marco Conti. « Opportunistic networking : data forwarding in disconnected mobile ad hoc networks ». Dans : *Communications Magazine, IEEE* 44.11 (2006), p. 134–141 (cf. p. 4).

[Rhe+09] Injong Rhee, Minsu Shin, Seongik Hong, Kyunghan Lee, Seongjoon Kim et Song Chong. *CRAWDAD dataset ncsu/mobilitymodels (v. 2009-07-23)*. Downloaded from http ://crawdad.org/ncsu/mobilitymodels/20090723. Juil. 2009 (cf. p. 6).

[Rhe+11] Injong Rhee, Minsu Shin, Seongik Hong, Kyunghan Lee, Seong Joon Kim et Song Chong. « On the levy-walk nature of human mobility ». Dans : *IEEE/ACM transactions on networking (TON)* 19.3 (2011), p. 630–643 (cf. p. 10).

[RMSM01] Elizabeth M Royer, P Michael Melliar-Smith et Louise E Moser. « An analysis of the optimum node density for ad hoc mobile networks ». Dans : *Communications, 2001. ICC 2001. IEEE International Conference on*. T. 3. IEEE. 2001, p. 857–861 (cf. p. 9).

[Sal+10] Marcel Salathé, Maria Kazandjieva, Jung Woo Lee, Philip Levis, Marcus W Feldman et James H Jones. « A high-resolution human contact network for infectious disease transmission ». Dans : *Proceedings of the National Academy of Sciences* 107.51 (2010), p. 22020–22025 (cf. p. 5).

[Sco+09] James Scott, Richard Gass, Jon Crowcroft, Pan Hui, Christophe Diot et Augustin Chaintreau. *CRAWDAD dataset cambridge/haggle (v. 2009-05-29)*. Downloaded from http ://crawdad.org/cambridge/haggle/20090529. Mai 2009 (cf. p. 7).

[Sho+11] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec et Jean-Pierre Hubaux. « Quantifying location privacy ». Dans : *Security and privacy (sp), 2011 ieee symposium on*. IEEE. 2011, p. 247–262 (cf. p. 12, 61).

[Sto] *stormoRevolution, STORMO® rEVOLUTION Scientific Session*. `http://www.effettolarsen.it/Stormo_revolution_sci_eng.html`. 2016 (cf. p. 19).

[Toh01] Chai K Toh. *Ad hoc mobile wireless networks : protocols and systems*. Pearson Education, 2001 (cf. p. 4).

[Too+15] Jameson L Toole, Carlos Herrera-Yaqüe, Christian M Schneider et Marta C González. « Coupling human mobility and social ties ». Dans : *Journal of The Royal Society Interface* 12.105 (2015), p. 20141128 (cf. p. 6).

[Tou+11] Pierre-Ugo Tournoux, Jeremie Leguay, Farid Benbadis, John Whitbeck, Vania Conan et Marcelo Dias De Amorim. « Density-aware routing in highly dynamic DTNs : The rollernet case ». Dans : *Mobile Computing, IEEE Transactions on* 10.12 (2011), p. 1755–1768 (cf. p. 7).

[Ubi] *ubisense, location intelligence products*. `http://ubisense.net/en`. 2016 (cf. p. 16).

[Wu+08] Lynn Wu, Benjamin N Waber, Sinan Aral, Erik Brynjolfsson et Alex Pentland. « Mining face-to-face interaction networks using sociometric badges : Predicting productivity in an it configuration task ». Dans : *Available at SSRN 1130251* (2008) (cf. p. 6).

[YLN03] Jungkeun Yoon, Mingyan Liu et Brian Noble. « Random waypoint considered harmful ». Dans : *INFOCOM 2003. twenty-second annual joint conference of the IEEE computer and communications. IEEE societies*. T. 2. IEEE. 2003, p. 1312–1321 (cf. p. 9).

Liste de publications

Jesus Friginal, Marc Olivier Killijian, Roberto Pasqua, Matthieu Roy et Gilles Trédan. « Does Mobility Matter ? An Evaluation Methodology for Opportunistic Apps ». Dans : *Network Computing and Applications (NCA), 2014 IEEE 13th International Symposium on*. IEEE. 2014, p. 24–31.

Marc-Olivier Killijian, Roberto Pasqua, Matthieu Roy, Gilles Trédan et Christophe Zanon. « Souk : Spatial Observation of hUman Kinetics ». Dans : *Computer Networks* (2016).

Roberto Pasqua, Matthieu Roy et Gilles Tredan. « Loca : a location-oblivious co-location attack in crowds ». Dans : *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM. 2016, p. 535–544.

Résumé — La diffusion massive de dispositifs portables, de plus en plus utilisés pour le traitement et la communication de l'information, permet la collecte d'importantes masses de données liées à l'activité des utilisateurs sur des applications mobiles. Nous nous intéressons aux données de localisation (les traces de mobilité) qui sont issues de systèmes mobiles formés par un groupe d'utilisateurs. Les données de mobilité produites dans un système mobile sont étudiées suivant deux axes :

– L'utilisation des modèles de mobilité est à la base du développement d'algorithmes de communication dédiés au systèmes mobiles. Les données de mobilité réelles concernant les utilisateurs vont nous permettre de comparer les données de mobilité synthétiques utilisées dans la simulation avec la réalité qu'ils sont censés décrire.
– La manipulation des données de mobilité réelles implique une réflexion sur les conséquences que les informations extraites de ces données ont relativement à la protection de la vie privée des utilisateurs.

Les contributions sur ces deux fronts sont les suivantes :
– Une analyse fine des propriétés spatiales et sociales d'un ensemble de traces de mobilité réelles collecté expérimentalement à l'aide d'une plateforme à haute precision.
– Une comparaison exploratoire entre des traces de mobilité réelles et des traces de mobilité synthétiques générées à partir de sept différents modèle de mobilité.
– La conception et l'analyse d'un algorithme d'inférence par co-localisation décorrélée des informations sur la localisation des utilisateurs ciblés.
– La quantification du potentiel des données de co-localisation non-déterministes sur la perte de protection de la vie privée d'un ensemble d'utilisateurs.

Mots clés : Systèmes mobiles, systèmes distribués, modèles de mobilité, exploration des données, réseaux sociaux, protection de la vie privée.

Abstract — The wide diffusion of smart portable devices allows the collection of a big amount of data concerning the activities of users from mobile apps. We focus our attention on location data, *i.e.* mobility traces, of a set of users in a crowd. Data collected from these mobile systems are studied following two axies :

- Mobility models are used to simulate the behavior of users to develop opportunistic forwarding algorithms. We compare real and synthetic mobility traces to show the distance between the reality and the models.
- Information on mobility may endanger the privacy of users. We analyze the impact of such information on privacy of users.

The main contributions are :

- We analyze the spatial and social properties of human motion from real traces collected by a highly accurate experimental localization system.
- We compare the real traces with synthetic mobility traces generated from seven popular mobility models
- We provide an inference algorithm based on co-location of users and we show its efficiency on different datasets.
- We quantify the effect of probabilistic co-location information by means of a novel co-location attack.

Keywords : Mobile systems, distributed systems, mobility models, data mining, social network, privacy.

Laboratoire d'analyse et d'architecture des systèmes (LAAS-CNRS)
7, avenue du Colonel Roche 31031
Toulouse, France