HAL open science

Contacts between individuals: analysis and application to the study of infectious diseases propagation

Julie Fournet

▶ To cite this version:

Julie Fournet. Contacts between individuals: analysis and application to the study of infectious diseases propagation. Statistical Mechanics [cond-mat.stat-mech]. Aix Marseille University, 2016. English. NNT : . tel-03549476

HAL Id: tel-03549476

https://theses.hal.science/tel-03549476v1

Submitted on 2 Feb 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Aix-Marseille Université
Ecole Doctorale 352
Physique et Sciences de la Matière

Thèse de Doctorat présentée par

Julie FOURNET

pour obtenir le grade de

Docteur d'Aix-Marseille Université
Spécialité : Physique Théorique et Mathématique

A soutenir le 26 Septembre 2016

Contacts entre individus : analyse et application à l'étude de la propagation de maladies infectieuses

Directeur de thèse : Alain BARRAT

Thèse préparée au Centre de Physique Théorique

Jury :

Rapporteurs :	Renaud LAMBIOTTE	- Université de Namur
	Yamir MORENO	- Universidad de Zaragoza
Directeur de thèse :	Alain BARRAT	- CNRS
Membres du jury :	Ciro CATTUTO	- ISI Foundation
	Xavier LEONCINI	- Aix-Marseille Université
	Nicola PERRA	- Greenwich University
	Jean-François PINTON	- ENS Lyon
	Chiara POLETTO	- INSERM

Remerciements

Trois ans de thèse c'est avant tout trois ans de travail, c'est long, c'est difficile et je n'aurais pas pu aller jusqu'au bout de l'aventure sans le soutien et les encouragements de nombreuses personnes que je tiens à remercier ici.

Je remercie tout d'abord Alain Barrat, mon directeur de thèse pour m'avoir donné cette opportunité et m'avoir guidée tout au long de ces trois années, ainsi que pour ses précieux conseils.

Je remercie naturellement Renaud Lambiotte et Yamir Moreno qui m'ont fait l'honneur d'être les rapporteurs de ce manuscrit ainsi que Ciro Cattuto, Xavier Leoncini, Nicola Perra, Jean-François Pinton et Chiara Poletto qui ont accepté de faire partie de mon jury de soutenance.

Je remercie infiniment mes co-bureaux Christian, Mathieu, Rossana et Antoine qui ont fait du bureau 602 un lieu où j'avais hâte de venir chaque jour et sont devenus des amis très proches. Je leur souhaite de réussir là où la vie les mènera, que ce soit dans la recherche ou ailleurs. Merci également à François pour nos discussions enflammées aussi bien sur la physique que sur le cinéma et pour m'avoir accordé sa confiance et son amitié. Ensuite, merci à Maxime, Matteo, Sarah et Benjamin pour des rencontres moins régulières mais toujours intéressantes.

Je remercie tout le laboratoire du Centre de Physique Théorique, et en particulier le personnel du secrétariat pour leur accueil et leur disponibilité.

Mes remerciements vont aussi à mes deux meilleures amies Coline et Elise pour m'avoir parfois permis de penser à autre chose mais aussi m'avoir poussé à continuer même quand je me sentais découragée.

Ces remerciements ne seraient pas complets sans un immense merci à Thomas avec qui je partage ma vie et ma passion pour la physique depuis bientôt cinq ans. Je le remercie pour son soutien indéfectible et ses encouragements quotidiens et pour avoir été là tout simplement. Sans lui, je ne serais certainement pas arrivée jusque là. Mention spéciale à mes deux chats, Tapioca et Tequila, qui malgré toutes leurs tentatives pour s'allonger sur mon clavier d'ordinateur notamment, n'ont pas réussi à m'empêcher de terminer cette thèse.

Enfin un immense merci à mes parents, Agnès et Marc, et mes grands-parents, Janine et Louis qui m'ont toujours encouragée dans tout ce que j'ai entrepris et m'ont toujours fait me sentir comme quelqu'un de formidable. Si vous voulez mon avis c'est eux qui sont vraiment formidables. Un petit clin d'œil supplémentaire à mon père qui m'a transmis l'amour des sciences et de la physique si jeune que je ne saurais dire quand exactement.

Résumé

Les contacts face-à-face entre individus permettent de caractériser les réseaux sociaux et jouent un rôle important dans la compréhension des mécanismes de propagation des épidémies dans une population. De récentes avancées technologiques ont rendu possible l'acquisition de données précises sur les interactions humaines. En particulier, la collaboration SocioPatterns a développé une infrastructure basée sur des capteurs portables (badges) qui enregistrent les contacts entre les participants avec une grande résolution spatiale et temporelle. Cette infrastructure a été déployée dans divers contextes tels que des écoles, des hôpitaux ou des conférences. Cette thèse présente, dans un premier temps, l'analyse de données de contacts collectées trois années de suite (2011, 2012 et 2013) dans un lycée français entre des étudiants de classes préparatoires. L'analyse a montré que la plupart des contacts se produisent entre étudiants de même classe et que les structures des contacts sont très similaires d'un jour sur l'autre et également d'une année sur l'autre. De plus, les propriétés statistiques des contacts concordent avec les résultats obtenus dans d'autres contextes. Des méthodes de collecte plus traditionnelles basées sur l'auto-évaluation sont aussi utilisées pour étudier les relations humaines. Toutes ces méthodes ont chacune des avantages et des inconvénients mais sont rarement comparées dans un environnement donné. Ici, nous avançons dans cette direction en comparant différentes méthodes de collecte de données : les données collectées en 2013 concernent non seulement des données de contact obtenues avec les capteurs mais aussi des données de contacts rapportés de façon rétrospective par les étudiants eux-mêmes (registre de contacts), les relations d'amitié déclarées par les étudiants et les liens d'amitié sur Facebook. Malgré la faible participation des étudiants à ces autres méthodes, nous avons obtenu des résultats intéressants : les étudiants se souviennent plus facilement des contacts les plus longs et ont tendance à surestimer leur durée alors que la plupart des contacts de courte durée sont oubliés ; les contacts les plus longs correspondent à des amitiés déclarées, la plupart des amitiés débouchent sur des contacts enregistrés par les badges mais la plupart des contacts courts ne correspondent pas à des amitiés déclarées ; la comparaison du réseau de contacts avec le réseau Facebook a montré que la présence d'un lien sur Facebook ne donne pas d'information sur l'existence effective de contacts ou d'amitiés réelles. Finalement, l'utilisation des registres de contacts ou des données d'amitié dans des simulations de propagation épidémique mène à une sous-estimation du risque épidémique quand on compare

avec les résultats obtenus en utilisant les données des badges. Dans la dernière partie, nous étudions, dans le cas du réseau d'amitié, si cette sous-estimation peut être vue comme un biais résultant d'un processus d'échantillonnage réalisé sur le réseau des contacts, ce qui pourrait nous donner des indications sur comment compenser ce biais et comment utiliser les informations contenues dans un jeu de données incomplet pour obtenir des prédictions fiables sur le risque épidémique, même en l'absence de données sur le vrai réseau de contact.

Abstract

Face-to-face contacts between individuals contribute to shape social networks and play an important role in determining how infectious diseases can spread within a population. Recently, technological advances have made it possible to obtain accurate data on human interactions. In particular, the SocioPatterns collaboration has developed an infrastructure based on wearable sensors that record contacts between participants with high spatio-temporal resolution. This infrasctructure has been deployed in various contexts such as schools, hospitals or conferences. This thesis first presents the analysis of contact data collected three years in a row (2011, 2012 and 2013) in a French high school among students of "classes préparatoires" (i.e., studies taking place after high school and preparing for admission to higher education colleges). The analysis showed that most contacts occur within students of same classes and that contact patterns are very similar from one day to the next and also from one year to the next. Moreover, statistical properties of contacts are in good agreement with the results obtained in other contexts. More traditional methods based on self-reporting are also used to investigate human relationships. These methods have each advantages and limitations but are rarely compared in a given setting. Here, we make progresses in this direction by comparing different methods of data collection: the complete data set collected in 2013 gathers not only contact data obtained from sensors but also contact data obtained from retrospective contact diaries, friendship relations declared by students and Facebook links. Despite the low participation of students in the latter methods, we have obtained noteworthy results: the students remember more easily their longest contacts and have a tendency to overestimate their durations while most short contacts are forgotten; the longest contacts correspond to reported friendships, most friendships lead to actual encounters but most short contacts did not correspond to reported friendships; the comparison of the sensors network and the Facebook network showed that the existence of a Facebook link gives no information on the existence of actual contacts or real friendships. Finally, we have found that the use of contact diaries or friendship data in simulations of epidemic spreading leads to an underestimation of the epidemic risk when compared with results obtained using the sensors data. In the last part, we investigate, in the case of the friendship network, whether this underestimation may be seen as biases due to a sampling process performed on the contact network obtained from sensors, which might give hints on how to compensate

these biases and how to use the information contained in incomplete data sets to obtain accurate predictions of the epidemic risk, even in the absence of data on the actual contact network.

Synthèse en français

Introduction

Dans cette thèse, on s'intéresse aux réseaux d'interactions humaines qui constituent la toile de fond des phénomènes de propagation et en particulier la propagation des maladies infectieuses. Ce terme "interactions humaines" regroupe un large spectre de différents types d'interactions tels que les e-mails, les appels téléphoniques, les contacts face-à-face, entre autres. Lorsqu'on s'intéresse à la propagation des épidémies, les contacts face-à-face semblent avoir un rôle prépondérant car ils agissent comme un vecteur de transmission privilégié pour les maladies infectieuses.

Pour étudier ce type de contacts, on utilise des données récoltées par la collaboration SocioPatterns dans un lycée marseillais entre des étudiants de classes préparatoires. Ce jeu de données regroupe des données de contacts face-à-face collectées trois années de suite (2011, 2012 et 2013) enregistrées avec un système basé sur des capteurs portables (badges RFID). Cette méthode de collecte de données a l'avantage d'être objective contrairement à d'autres méthodes faisant intervenir la mémoire ou les ressentis des participants.

Pour étudier les avantages et inconvénients de ces différentes méthodes, on utilise d'autres types de données collectées en 2013 (dans la même population que précédemment), à savoir : (i) des données de contacts collectées par un questionnaire dans lequel les étudiants étaient invités à rapporter les contacts qu'ils avaient eu, avec qui et leur durée approximative, (ii) des données d'amitié collectées par sondage et (iii) le réseau des relations sur Facebook.

Enfin la comparaison de ces données nous poussera à chercher quelles informations nous pouvons tirer d'un jeu de données incomplet, en particulier quelles sont les informations contenues dans ces données incomplètes qui permettent d'obtenir une bonne estimation du risque épidémique dans une population.

Analyse des données de contacts

Dans cette première partie, on se concentre sur l'analyse des données de contacts récoltées trois années de suite dans un lycée. Ces données ont été récoltées parmi des étudiants de 3 (2011), 5 (2012) et 9 classes (2013). Pour éviter les répétitions, nous rapportons ici les résultats obtenus avec les données de 2013.

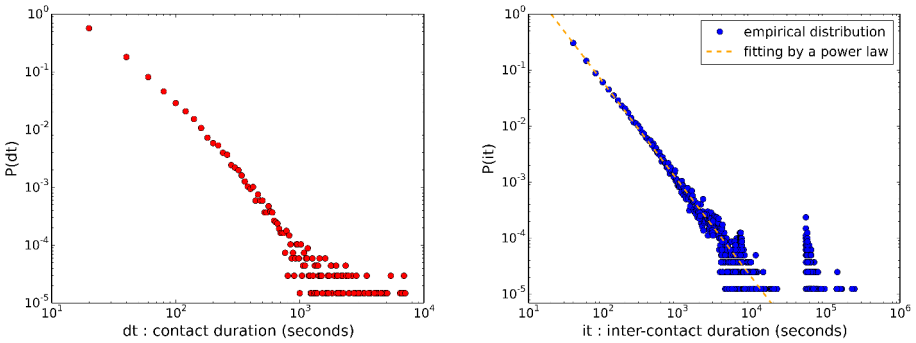


Figure 1: A gauche : distribution $P(dt)$ des durées de contacts pour les données de 2013 i.e., probabilité qu'un contact enregistré ait duré dt, à droite : distribution $P(it)$ des durées inter-contacts i.e., probabilité que le temps écoulé entre deux contacts successifs d'un individu soit it.

Sur les 5 jours qu'a duré le déploiement de 2013, 67613 contacts ont été détectés entre les 327 participants pour une durée cumulée d'environ 1047 heures de contacts. Un contact est défini comme une série ininterrompue d'intervalles pendant laquelle il existe un lien entre deux individus. Sur la figure 1, on montre la distribution des durées de contacts ainsi que la distribution des durées inter-contacts des nœuds i.e., durée pendant laquelle un nœud n'est en contact avec aucun autre individu. Ces deux distributions sont larges : la plupart des contacts sont de courte durée mais on observe aussi des contacts plus longs, voire de très longs contacts. Le réseau de contact agrégé sur toute la durée de l'étude a les propriétés d'un réseau "petit monde" : un petit diamètre (la plus grande distance entre deux nœuds) et un grand coefficient de clustering. La distribution des degrés des nœuds montre que le réseau de contact est homogène en terme de degrés comme c'est le cas dans beaucoup de réseaux d'interactions humaines. Au contraire, la distribution des poids des liens (durée agrégée des contacts entre deux individus) est une distribution large : la plupart des liens ont un faible poids mais on observe aussi des liens avec un poids beaucoup plus important.

La population des étudiants impliqués dans l'étude est divisée en 9 classes. On définit alors des matrices de contacts qui permettent d'avoir une idée générale de la structure du réseau de contact en divisant la population en groupes. Les matrices résultantes ont une structure quasi-diagonale montrant que la plupart des contacts ont lieu entre des étudiants faisant partie de la même classe. Ces résultats sont d'ailleurs en accord avec des résultats obtenus dans d'autres environnements scolaires. Au contraire, très peu de contacts ont lieu entre les différentes classes. Cependant, on remarque une sous-structure de trois groupes de classes étudiant des sujets similaires (biologie, mathématiques ou physique) : il y a plus de contacts à l'intérieur de ces trois groupes de classes qu'entre les groupes de classes.

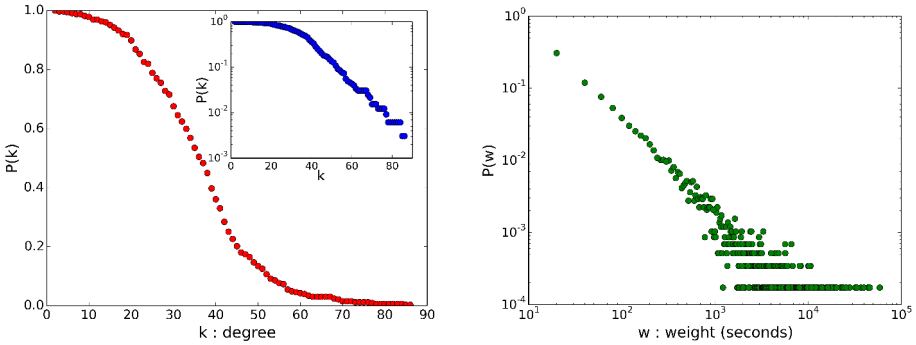


Figure 2: A gauche : Distribution $p(k)$ des degrés des nœuds (encart : distribution en échelle semi-log) i.e., probabilité pour un nœud d'avoir un degré supérieur ou égal à k, à droite : Distribution $p(w)$ des poids des liens.

Etant donné cette structure en classes, il est évident qu'un mélange homogène des nœuds ne serait pas une bonne représentation. En revanche, il a été observé qu'une division des individus en sous-groupes en fonction de leur genre n'apporte pas d'information supplémentaire puisque nous n'avons pas observé d'homophilie par rapport au genre (tendance à avoir plus de contacts avec des individus de même genre que soi) dans les données utilisées, contrairement à ce qui a été observé dans une école primaire. La division de la population à l'échelle des classes apparait comme un niveau adéquat de description du réseau, notamment si on souhaite concevoir un modèle des contacts pour évaluer le risque épidémique dans cette population.

Une autre partie importante de l'analyse concerne l'analyse longitudinale de notre dataset à deux échelles temporelles différentes : (i) d'un jour à l'autre et (ii) d'une année sur l'autre. En effet, lorsqu'on cherche à concevoir des modèles "data-driven" réalistes de contacts entre individus ou à renseigner des modèles de propagation, la robustesse des caractéristiques des contacts à plusieurs échelles temporelles représente une information cruciale. Sur la figure 2.11, nous montrons l'évolution du nombre de contacts au cours de la journée. On remarque que le nombre de contacts varie beaucoup au cours de la journée mais chaque jour présente une évolution très similaire aux autres. En effet, chaque journée est marquée par des pics d'activité déterminés par les récréations et les pauses repas. De plus, l'activité tombe à zéro pendant la nuit puisque les étudiants rentrent chez eux (l'enregistrement des contacts n'ayant lieu que dans l'enceinte du lycée).

Quand on s'attaque à la comparaison des propriétés des différents réseaux journaliers on remarque que ces propriétés sont extrêmement robustes : les distributions statistiques des caractéristiques des contacts (durées, durées inter-contacts, poids des liens) et des degrés des nœuds des différents jours se superposent, la comparaison des matrices de contacts des différents jours révèle des valeurs de similarités (similarités cosinus) très élevées. De plus, les étudiants

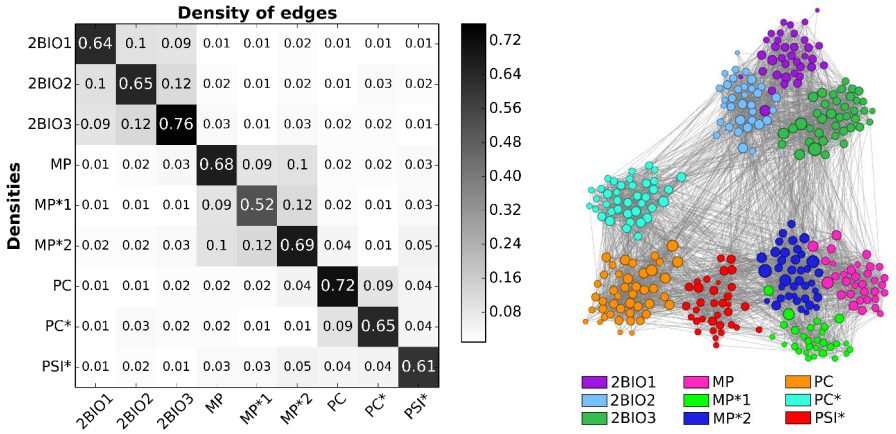


Figure 3: A gauche : matrice de densités de liens entre les différentes classes, à droite : représentation du réseau de contacts de 2013. Chaque nœud représente un étudiant, la couleur représente la classe de l'étudiant et la taille représente son nombre de voisins.

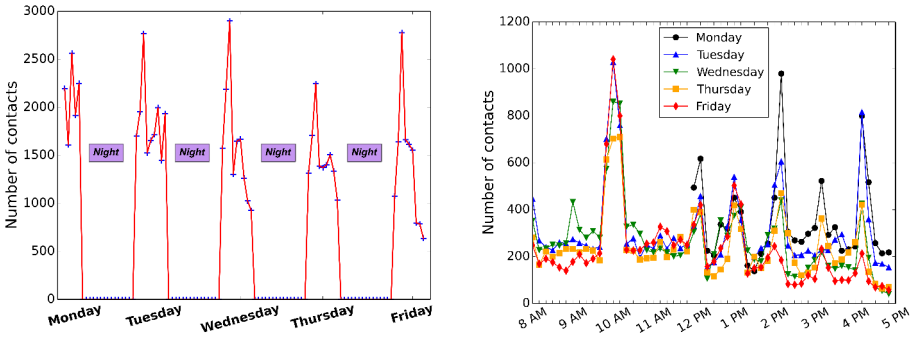


Figure 4: A gauche : Evolution du nombre de contacts par plage horaire de une heure au cours de l'étude, à droite : Evolution du nombre de contacts par plage horaire de 10 minutes pour chaque jour de l'étude.

changent de voisins (individus avec qui ils ont des contacts) d'un jour à l'autre mais beaucoup moins par rapport à ce qui serait attendu si les contacts avaient lieu au hasard.

De plus, cette robustesse est aussi observée à l'échelle annuelle. En effet, la comparaison entre les études réalisées en 2011, 2012 et 2013 avec des étudiants différents chaque année a montré de grandes similarités : encore une fois les distributions statistiques se superposent d'une année sur l'autre et les matrices présentent une structure diagonale indiquant que les contacts ont plutôt lieu entre des étudiants appartenant à la même classe. Les valeurs de clustering et de distances sont également similaires sur les trois ans et l'évolution temporelle du nombre de contacts au cours de la journée de cours présente des profils similaires chaque année avec notamment les pics d'activité aux moments où les étudiants ne sont plus en classe.

Nous avons aussi comparé les résultats obtenus avec nos données à ceux obtenus dans un environnement similaire : un lycée américain où la proximité entre 788 étudiants a été détectée par des méthodes similaires. Dans cette étude, la définition d'un contact est légèrement différente de la nôtre, cependant les résultats obtenus sont assez similaires : notamment les distributions des durées de contacts et des poids des liens sont des distributions larges avec des pentes similaires.

Comparaison des différentes méthodes de collecte de données

La première partie se concentre sur l'analyse de données collectées par une méthode objective, à savoir l'utilisation de badges qui enregistrent automatiquement les contacts (proximité physique et face-à-face) des étudiants. Certains biais sont ainsi évités. Cependant chacune de ces méthodes a des avantages et des inconvénients, par exemple les badges permettent d'enregistrer les contacts de façon objective mais des questionnaires bien étudiés peuvent permettre de récolter des informations supplémentaires sur les contacts tels que la nature du contact (professionnel, amitié) et s'il y a eu contact physique ou non. Cette partie est alors consacrée à la comparaison de différentes méthodes de collecte de données. Cette comparaison sert de point de départ pour quantifier les biais associés à chacune des méthodes par rapport à l'utilisation de badges considérée à l'heure actuelle comme la méthode la plus efficace, et permet aussi de se faire une idée de la quantité d'informations qui peut être récupérée dans des données incomplètes et qui pourrait être utilisée dans le contexte de l'étude de la propagation des maladies infectieuses.

Lors du déploiement de 2013, les contacts ont donc été enregistrés avec l'infrastructure développée par la collaboration SocioPatterns (badges RFID qui détectent la proximité spatiale de face avec un pas de temps de 20 secondes) mais d'autres données ont également pu être collectées : (i) à la fin du quatrième jour de déploiement, les étudiants ont été invités à remplir une fiche

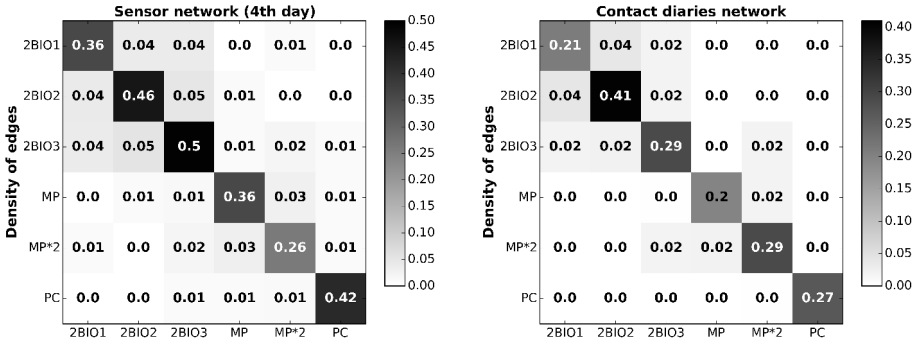


Figure 5: Matrices de contacts des densités de liens pour le réseau de contact obtenu avec les badges le 4ème jour du déploiement (gauche) et pour le réseau des contacts rapportés par les étudiants.

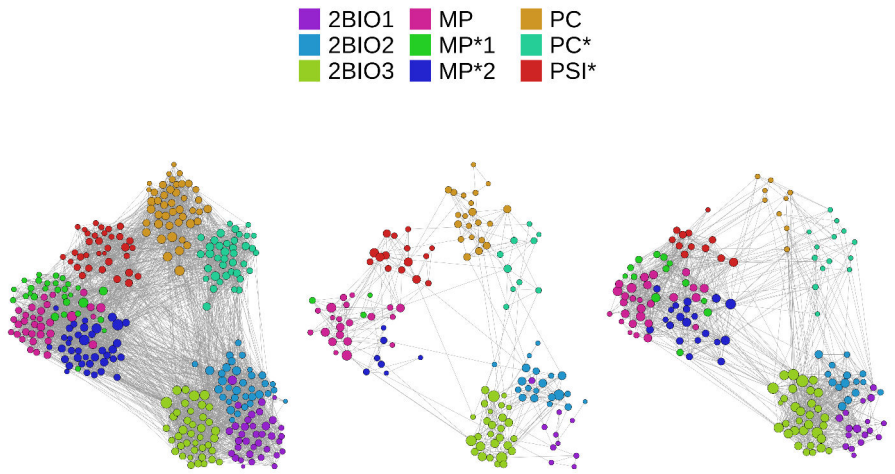
dans laquelle ils devaient rapporter les noms de ceux avec qui ils avaient été en contact au cours de la journée et la durée totale approximative (parmi 4 catégories de durée) de contact avec chacun de ces individus, (ii) les étudiants ont également pu rapporter les noms de leurs amis parmi les autre étudiants participants, (iii) enfin certains étudiants ont accepté de fournir leur réseau d'amitié Facebook. Le premier inconvénient rencontré avec ces méthodes de collecte est que peu d'étudiants ont participé en comparaison du nombre de participants au déploiement SocioPatterns.

Dans un premier temps, nous avons comparé le réseau de contact avec le réseau obtenu avec le sondage "mémoire" ainsi que les différentes réponses des étudiants à ce sondage. En effet, un contact peut avoir été rapporté par l'un des deux ou les deux étudiants impliqués dans ce contact et avec une durée approximative différente. En fait, lorsque deux étudiants ont rapporté un contact entre eux, dans la majeure partie des cas, les étudiants ont rapporté ce contact avec la même durée approximative. Du côté de la comparaison avec le réseau des badges, il apparait que la plupart des contacts de courte durée, tels qu'ils ont été mesurés par les badges, n'ont pas été rapportés alors que les contacts plus longs ont une probabilité plus grande d'avoir été rapportés ; de plus tous les contacts enregistrés avec une durée supérieure à environ une heure ont été rapportés. En contrepartie, la durée de contact rapportée par les étudiants a tendance à surestimer la durée enregistrée par les badges (ce résultat est d'ailleurs en accord avec plusieurs études sociologiques qui affirment que les individus ont tendance à percevoir le temps différemment de la réalité et souvent à le surestimer). Malgré le faible taux de participation au sondage "mémoire" et la faible densité du réseau résultant, on a remarqué que la structure générale du réseau était bien préservée par le processus de sondage, ce qui peut être notamment observé à travers la similarité des matrices de contacts (Figure 3.4). En conséquence, les données tirées du sondage pourraient éventuellement contenir assez d'information pour informer des modèles décrivant les contacts humains.

Les données rapportées par le sondage mémoire correspondent au même type de relations que celles détectées par les badges, à savoir des contacts ayant eu lieu. D'un autre côté, les données obtenues par le sondage d'amitié correspondent à un type différent de relation. En effet, on peut s'attendre à ce que des amis se rencontrent plus souvent mais une amitié ne mène pas forcément à des rencontres physiques et les contacts n'ont pas lieu qu'entre deux amis. La comparaison des réseaux de contact (badges) et d'amitié a montré que les contacts les plus longs ont eu lieu entre des étudiants s'étant déclarés comme amis, la plupart des amitiés déclarées correspondent à des contacts détectés par les badges, en revanche, beaucoup de contacts de courte durée ne correspondent pas à des amitiés. De plus, comme dit précédemment, le nombre de participants au sondage d'amitié est largement inférieur au nombre d'individus qui ont accepté de porter les badges. Ainsi le réseau d'amitié a beaucoup moins de nœuds et est beaucoup moins dense que le réseau de contacts. Néanmoins, comme pour le réseau obtenu avec le sondage mémoire, la structure en classes du réseau de contact est bien préservée. Encore une fois, on pourrait alors considérer que le réseau d'amitié contient assez d'information pour informer des modèles.

Le réseau d'amitié Facebook est quant à lui beaucoup plus difficile à analyser. En effet, étant donné le nombre négligeable d'étudiants ayant accepté de communiquer leur réseau Facebook, le réseau résultant n'est pas, à proprement parler, un réseau puisque pour de nombreuses paires de nœuds, on ne sait pas s'il existe un lien entre eux ou pas. Ce réseau ci contient bien trop peu d'information pouvant être utilisée pour la modélisation.

Enfin, des simulations de propagation d'épidémies réalisées sur les réseaux obtenus avec les badges et avec les deux types de sondages ont montré que l'utilisation des réseaux obtenus par sondage mène à une grande sous-estimation du risque épidémique tel qu'il est calculé en utilisant le réseau des badges. La comparaison des résultats obtenus est intéressante car elle pourrait nous aider à quantifier et comprendre les biais contenues dans les données auto-rapportées ainsi que peut-être des pistes pour compenser de tels biais. Cela pourrait également nous aider à comprendre comment utiliser des données incomplètes pour informer des modèles de propagation d'épidémies.



(a) Contact Network (b) Friendship Network (c) Facebook Relationships*

Figure 6: Réseaux de contact, d'amitié et Facebook. Les trois réseaux sont montrés en utilisant la même spatialisation des nœuds. La couleur du nœud représente la classe de l'étudiant et sa taille représente son degré dans le réseau correspondant. *Note : les données Facebook ne donnent pas accès à un réseau au sens strict du terme car pour certaines paires de nœuds nous ne savons s'il y a ou pas un lien Facebook entre les deux.

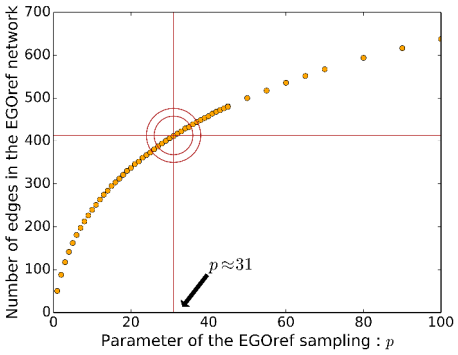


Figure 7: Nombre de liens dans le réseau échantillonné obtenu avec la méthode EGOref en fonction du paramètre p. Le nombre de nœuds est fixé égal à celui du réseau d'amitié. La droite horizontale rouge représente le nombre de liens présent dans le réseau d'amitié.

Equivalence entre le réseau d'amitié et un échantillonnage non-uniforme du réseau de contact

Dans la partie précédente on a vu que l'utilisation des réseaux obtenus par sondage dans des simulations de propagation d'épidémie mène à une sous-estimation du risque épidémique. Dans cette partie, on se concentre sur le réseau d'amitié et on cherche à simuler les biais dus au processus du sondage d'amitié par un échantillonnage réalisé sur le réseau de contact. On a vu que le réseau d'amitié avait beaucoup moins de nœuds que le réseau des contacts et était également beaucoup plus dilué.

Pour reproduire les biais du processus de sondage et en particulier ceux observés sur le risque épidémique, nous avons testé plusieurs méthodes d'échantillonnage. Dans un premier temps, nous avons utilisé des méthodes qui contrôlent uniquement le nombre de nœuds du réseau échantillonné. Sans surprise, ces méthodes ne permettent pas de reproduire les résultats de simulations d'épidémies obtenus avec le réseau d'amitié. Dans un second temps, nous avons alors testé des méthodes qui permettaient de contrôler le nombre de nœuds et le nombre de liens du réseau échantillonné pour pouvoir choisir ces grandeurs égales à celle du réseau d'amitié. En particulier, nous avons conçu une méthode (méthode EGOref) qui sélectionne les liens du réseau de contact avec une probabilité dépendant de leur poids dans le réseau de contact (avec l'hypothèse que les contacts les plus longs correspondent à des amitiés et que toutes les amitiés correspondent à des contacts) et d'un paramètre p. Ce paramètre peut être choisi de telle façon que l'on obtienne le nombre voulu de liens dans le réseau résultant (Figure 4.4).

Cette méthode a permis de reproduire avec beaucoup de fidélité les résultats obtenus avec le réseau d'amitié et surtout bien mieux que d'autres méthodes

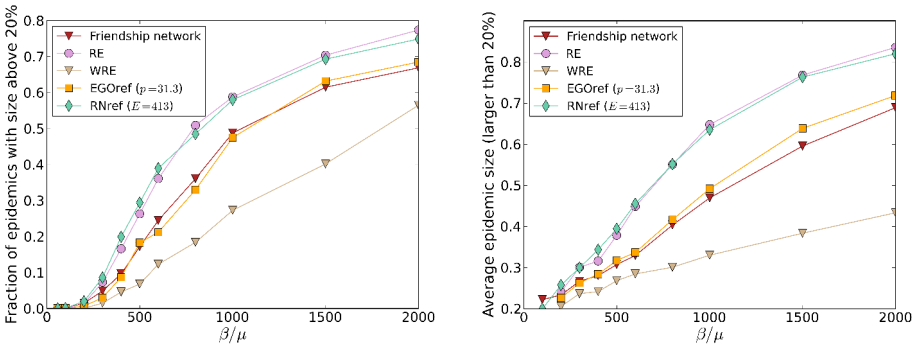


Figure 8: Résultats de simulations d'épidémies réalisées sur le réseau d'amitié et les réseaux échantillonnés ayant le même nombre de nœuds et le même nombre de liens que le réseau d'amitié. A gauche : fraction d'épidémies avec une taille finale supérieure à 20% de la population en fonction des paramètres de propgation, à droite : taille moyenne des épidémies ayant une taille supérieure à 20% en fonction des paramètres de propagation.

contrôlant également les nombres de nœuds et de liens (Figure 4.5). Le réseau résultant a aussi des caractéristiques très proches du réseau d'amitié ; seul le coefficient de clustering est remarquablement plus petit que celui du réseau d'amitié mais bien plus grand que pour les autres réseaux échantillonnés. Une méthode encore plus raffinée de ce modèle notée EGOref-het ne choisit pas les nœuds entièrement au hasard (elle choisit les nœuds dans chaque classe de façon à avoir la même division des nœuds en classe que dans le réseau d'amitié où il n'y a pas du tout le même nombre de nœuds dans chaque classe). Cette seconde méthode améliore quelque peu les résultats obtenus mais pas de façon significative. De plus, le coefficient de clustering est toujours bien en dessous de celui du réseau d'amitié.

Notre modèle EGOref dépend en fait de deux paramètres : le nombre de nœuds N et le paramètre p. Dans le contexte d'un sondage, ces nombres vont en fait correspondre au nombre de participants au sondage et à la quantité de contacts rapportés. Cela rend son application possible dans de nombreux contextes : en effet, nous avons commencé à étudier l'impact de ce type d'échantillonnage sur d'autres données de contacts collectées dans d'autres environnements, malheureusement la comparaison avec un réseau d'amitié correspondant est impossible car ces autres jeux de données ne combinent pas des données de contacts et des données d'amitié.

Finalement, nous avons aussi commencé à étudier des stratégies de reconstruction du réseau de contact à partir du réseau échantillonné. En effet, si nous pouvons reconstruire un réseau de contact, qui ne serait pas le vrai réseau de contact mais reproduirait ses caractéristiques, nous pourrions obtenir une bonne estimation du risque épidémique à partir de données incomplètes. Dans le cas de processus d'échantillonnage uniformes, des stratégies simples suff-

isent à reconstruire un réseau de contact de substitution et à évaluer le risque épidémique dans une population. Cela étant, notre processus d'échantillonnage reproduisant les caractéristiques du réseau d'amitié n'est pas uniforme et ces stratégies simples échouent. En effet, la stratégie utilisée préserve la densité du réseau échantillonné lors de la reconstruction, or, la méthode EGOref a justement été conçue pour modifier la densité du réseau de contact pour qu'elle corresponde à celle du réseau d'amitié. Ainsi, la reconstruction devrait, à partir du réseau échantillonné reconstruire un réseau ayant une densité bien supérieure. Les résultats des tests menés sur les réseaux échantillonnés avec la méthode EGOref dépendent en fait des deux paramètres N et p. En effet lorsque p est très grand, la méthode de reconstruction est efficace ; cela s'explique par le fait que lorsque p est très grand la méthode d'échantillonnage s'apparente à un échantillonnage uniforme.

Conclusion

Dans cette thèse, nous avons contribué à fournir une image plus complète des réseaux d'interactions humaines. Dans un premier temps, nous avons analysé des données de contacts collectées de façon objective à l'aide de badges dans un lycée. L'analyse a révélé les caractéristiques d'un réseau petit monde comme c'est le cas pour beaucoup de réseaux d'interactions humaines, de plus on a observé une grande robustesse des caractéristiques du réseau à plusieurs échelles temporelles. Ensuite, nous avons comparé plusieurs manières de collecte de données. Cette comparaison a montré que la méthode de collecte utilisant les badges peut être considérée comme la meilleure à ce jour mais que les méthodes utilisant des sondages permettent d'obtenir une bonne image de la structure générale du réseau au moins dans le cas d'un établissement scolaire où la plupart des contacts ont lieu entre des étudiants de même classe. Cependant, l'utilisation des réseaux construits à partir de données de sondage pour la simulation de processus épidémiques mène à une sous-estimation du risque épidémique. Enfin nous avons reproduit les résultats sur le risque épidémique obtenus avec le réseau d'amitié avec un échantillonnage du réseau de contact. Cette méthode d'échantillonnage permettant de reproduire les biais du réseau d'amitié pourrait nous aider à compenser de tels biais et à apprendre à utiliser l'information contenue dans des données incomplètes pour obtenir une bonne estimation du risque épidémique.

Contents

1 Introduction **3**
1.1 How to describe networks ? 4
1.2 Epidemic spreading processes on networks 7
1.2.1 Compartmental models in epidemiology 7
1.2.2 How are simulations performed? 8
1.3 Data collection 9
1.3.1 The SocioPatterns collaboration 10
1.3.2 Description of datasets 10
1.4 Overview of the following chapters 12

2 Analysis of face-to-face proximity data **13**
2.1 Study context 13
2.2 Number and durations of contacts 15
2.3 Contact matrices 16
2.4 Contact network 20
2.5 Gender homophily 22
2.6 Longitudinal analysis at daily scale 28
2.6.1 Temporal evolution of contact patterns 28
2.6.2 Comparison of daily patterns 30
2.7 Long-term stability of patterns 37
2.8 Comparison with another similar study 43
2.9 Conclusion 45

3 Comparison of methods of data collection **47**
3.1 Data analysis 48
3.1.1 Contact diaries 48
3.1.2 Friendships 49
3.1.3 Facebook 49
3.2 Contact diaries 51
3.2.1 Analysis of the contact diaries network 51
3.2.2 Comparing contact diaries and sensors data 51
3.3 Multiplex network of students' relationships 59
3.3.1 Contact network versus friendship-survey network 60
3.3.2 Face-to-face contacts and Facebook links 67

3.3.3 Contacts and friendship networks as a multiplex 69
3.4 Epidemic risk from different methods of data collection 70
3.5 Conclusion 75

4 Equivalence between friendship network and a non-uniform sampling of contact network 79
4.1 Methodology 80
4.1.1 Sampling methods 81
4.1.2 How are simulations performed ? 83
4.2 Properties of sampled networks and outcome of SIR simulations . 84
4.2.1 Simple sampling methods 84
4.2.2 More refined methods of sampling 87
4.3 Sampling model exploration 91
4.4 Impact of weight assignment 93
4.4.1 Contact network and sampling procedures independent from weights 94
4.4.2 Friendship network 98
4.4.3 WRE sampling procedure 99
4.4.4 EGOref sampling procedure 101
4.5 The EGOref sampling in other contexts 103
4.6 Conclusion and outlook 108

5 Conclusion 113

Appendices 117
A Introduction 119
B Analysis of face-to-face proximity data 121
C Equivalence between friendship network and a non-uniform sampling of contact network 125

List of publications 131

Bibliography 133

Chapter 1

Introduction

Biological systems, human relationships, airport networks and the Internet are a few examples of physical systems that all have one thing in common: they are composed by a large number of interconnected units and can be described by complex networks [1]. The study of complex networks and their applications have interested the research community for many years, since the mathematician Leonhard Euler and the famous Königsberg bridges problem in the context of graph theory. Recently, the study of complex networks has been extended to various topics in which complex networks are characterized by complex topology and heterogeneous structures.

In this thesis, we are interested in the networks of human interactions as a substrate for spreading processes and in particular spreading of infectious diseases. The term "human interactions" covers a large number of different types of interactions, for instance: emails, phone calls, online social networks, scientific collaborations or face-to-face interactions. In the context of epidemic spreading, face-to-face contacts are expected to play a crucial role as transmission routes for infectious diseases. However, contrary to e.g. emails, face-to-face interactions do not leave any digital trace and thus are more complicated to measure. As such, the development of measurement strategies has been a major challenge of the last decades [2, 3].

Traditional methods consist in for instance, time-use data, video monitoring, surveys or diaries. The benefit with surveys and diaries is that they can ask to participants, in addition to reporting their contacts, an estimation of contact durations, the possible presence of physical interactions during these contacts and they might also allow to gather information about other types of relationships surch as friendships for instance. These self-reporting methods are however prone to biases due to their subjective character [4, 5].

Recent approaches take advantage of the development of new technologies based on wearable sensors, which record contacts between participants with a high spatio-temporal resolution allowing to collect data in an objective way. The SocioPatterns collaboration [6] has developed an infrastructure based on RFID (Radio Frequency Identification) which has been deployed in various contexts

[7–12] such as schools, hospitals or conferences participating to the building of an atlas of human contacts. The main limitation of these methods arises from the fact that they allow to collect data only in a closed population and contacts with the individuals not participating to the data collection cannot be recorded.

In both cases, sampling issues might arise if individuals in the target population do not want to participate and wear the sensors or do not fill in the surveys. Moreover, in the case of diaries, individuals might forget about some of their contacts, especially short ones, yielding a network with a highly underestimated number of contact links between individuals [13,14]. Indeed, many datasets are incomplete samples of the actual network of interest. Thus, understanding if and how incomplete data can be used in data-driven models describing human interactions or the spreading of infectious diseases and for the design of containment strategies is of great interest.

The first chapter of this thesis is dedicated to the introduction of basic concepts used in the study of complex networks and the presentation of datasets used further in the thesis.

1.1 How to describe networks ?

In the following, we present some useful measurable quantities to classify networks. These properties of networks can help us to perform quantitative analysis when the structure of networks becomes complex.

- A **graph**, denoted $G(N, E)$, is the representation of a network. In this thesis, we will the two terms equivalently. N is the number of nodes (or vertices) in the network and E is the number of edges (or links) in the network. An edge between the two nodes i and j is denoted (i, j). The **density** d of a network is defined as the ratio of the number of edges to the number of possible edges. Thus, the density is a number between 0 and 1: when $d << 1$ the network is sparse, when $d \approx 1$ the network is dense. For an undirected network $G(N, E)$ the density is given by $d = E/[N(N-1)/2]$. A **directed** graph is defined as a set of nodes and a set of ordered pairs of nodes i.e., edges with a specific direction: (i, j) is not equivalent to (j, i). As the number of possible edges in a directed graph is twice higher than in an undirected graph, the density is $dd = E/[N(N-1)]$.

- The **adjacency matrix** of G is a $N \times N$ matrix: the element e_{ij} is 1 if there is a link between i and j, 0 otherwise. The adjacency matrix of an undirected network is symmetric while it is asymmetric for a directed network.

- A (simple) **path** between two nodes n_1 and n_k is a sequence of nodes $(n_1, n_2, ..., n_k)$ in which subsequent nodes are neighbors and all the nodes are distinct from each other. The shortest path between two nodes is the path with the smallest number of nodes. The length of the shortest path

between two nodes is called the **distance**. The diameter D of a graph is the largest distance between two nodes of the graph.

- The **degree** k of a node i is the number of its neighbors $V(i)$. It corresponds to the number of edges incident to this node. The average degree $< k >$ of a network is $2E/N$. The degree distribution $P(k)$ is defined as the probability that any randomly chosen node has degree k. Usually, we preferably plot the complementary cumulative distribution function (CCDF) of degrees which is defined as the fraction of nodes with degree greater than or equal to k. We measure the level of heterogeneity of networks by defining $\kappa =< k^2 > / < k >$ and we distinguish scale-free networks (i.e., heterogeneous networks) with $\kappa >>< k >$ and homogeneous networks with $\kappa \sim< k >$. In a directed graph, we also define the in-degree of a node k_{in} which is the number of edges that arrive to this node and the out-degree of a node k_{out} which is the number of edges coming from this node.

- The **clustering coefficient** of a graph G is a measure of the tendency of nodes to cluster together. There are two definitions of the clustering: a global one which is called transitivity and a local one c_i. The local clustering coefficient for each node is the ratio of the number of triangles between this node and its neighbors $T(i)$ to the number of possible edges between its neighbors,

$$c_i = \frac{2T(i)}{k_i(k_i - 1)}.$$

  The definition of the network clustering coefficient that we will use is the average of the local clustering coefficients of all the nodes.

$$C = \frac{1}{N} \sum_{i \in G} c_i.$$

- **Small-world networks** refer to a category of networks in which most nodes are not direct neighbors of one another but most nodes can be reached from every other node by a small number of steps. This small-world phenomenon can be found in many empirical graphs e.g., social networks, networks of brain neurons, the Internet. This type of networks is characterized by a large clustering coefficient and a small average shortest path length: the typical distance is typically of the order of the logarithm of the number of nodes N in the network $L \propto \log N$.

- A **weighted** network is a network in which a weight $w_{i,j}$ is assigned to each edge (i, j). The weight of a link is a numerical value which depends on the nature of the network e.g., if the network represents an airline network, the weights might represent the number of weekly flights between two airports. In this thesis, the weight of an edge will represent (most of the

time) the time spent in contact by two nodes.
The **strength** of a node i is the sum of the weights of edges incident to the node: $s_i = \sum_{j \in V(i)} w_{i,j}$.

- A **temporal** network is a network that evolves over time. To investigate the temporal evolution of the network we define new quantities:
  - The **activity** of nodes (resp. edges) is the number of nodes in contact (resp. number of pairs of nodes in contact) at a specific time interval and can be plotted over time.
  - A **contact** is defined as an uninterrupted sequence of timesteps in which two nodes are linked, we define as well the contact duration dt as a length of this sequence. The inter-contact duration it is the duration during which a node is not linked to any other node. The distributions of these quantities help to characterize the dynamics of the network. In the time-varying networks of face-to-face contacts, these distributions are broad with large variations.
  - A temporal network can be **aggregated** over time to obtain a static picture of a time-varying network. In networks depicting face-to-face interactions, the aggregate contact duration i.e., the total time spent in contact by two nodes during the aggregation time window corresponds to the weight of the edge between these two nodes.
- When the nodes of a network can be separated into categories (e.g., gender, school classes, countries...) we can study the mixing patterns between the different categories using **contact matrices**.

  We denote the number of nodes in the category X by n_X. Then we define the following possible contact matrices:
  - the number of edges between nodes of category X with nodes of category Y: $E_{XY} = \sum_{i \in X, j \in Y} e_{ij}$ for $X \neq Y$ (and $E_{XX} = \frac{1}{2} \sum_{i,j \in X} e_{ij}$);
  - the density of edges between category X and category Y: $\rho_{XY} = E_{XY}/E_{XY}^{max}$, where $E_{XY}^{max} = n_X n_Y$ is the maximum possible number of edges between category X and category Y ($E_{XX}^{max} = n_X(n_X - 1)/2$);
  - the total number of contacts between nodes of category X with nodes of category Y: $N_{XY} = \sum_{i \in X, j \in Y} n_{ij}$ (for $X = Y$ we have $N_{XX} = \frac{1}{2} \sum_{i,j \in X} n_{ij}$);
  - the average number of contacts of a nodes of category X with nodes of category Y $n_{XY} = \frac{N_{XY}}{n_X}$;
  - the total time spent in contact between nodes of category X with nodes of category Y: $W_{XY} = \sum_{i \in X, j \in Y} w_{ij}$ (for $X = Y$ we have $W_{XX} = \frac{1}{2} \sum_{i,j \in X} w_{ij}$);
  - the average time spent by a node of category X in contact with nodes of category Y: $w_{XY} = \frac{W_{XY}}{n_X}$;

In the context of comparing two matrices, we define the cosine similarity between 2 matrices M and N as $\sigma(M,N) = \frac{\Sigma_{i,j} m_{ij} n_{ij}}{\sqrt{\Sigma_{i,j} m_{ij}^2}\sqrt{\Sigma_{i,j} n_{ij}^2}}$. This number between 0 and 1 is close to 1 for similar matrices.

1.2 Epidemic spreading processes on networks

1.2.1 Compartmental models in epidemiology

The spread of infectious diseases is a complex phenomenon that depends on many factors. A category of mathematical models used to study the dynamics of outbreaks is the case of compartmental models. In this class of models, the population is divided into comparments describing the health state of the individuals. The number of nodes in each compartment fluctuates, whereas the total number of nodes remains constant (when studying a closed population). We give two examples of compartmental models: the SI model and the SIR model.

The SI model

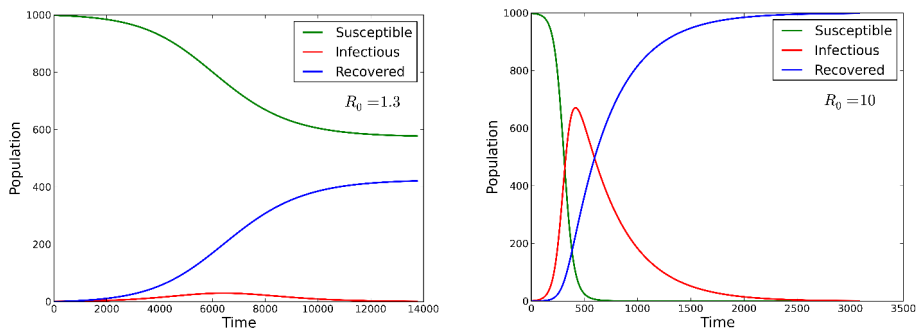
In the simplest epidemic model, the nodes can have two different states: Susceptible (S) or Infectious (I). The number of individuals in each state is denoted by S and I. The total number of nodes remains constant: $N = S + I$. A susceptible node in contact with an Infectious node changes its state to Infectious with probability β: $S + I \xrightarrow{\beta} 2I$. The Infectious nodes cannot return to the Susceptible state. We use the mean field approximation i.e., we consider that all nodes are linked to $< k >$ (average degree) neighbors. In this approximation, each Infectious individual (I) has $< k > S/N$ Susceptible neighbors, thus the number of new Infectious individuals per unit time is $\beta I < k > S/N$. The differential equations describing the evolution of the number of individuals in each compartment are given here (note that here the variables S, I and the time are continuous):

$$\frac{dS}{dt} = -\beta < k > \frac{IS}{N}$$

$$\frac{dI}{dt} = \beta < k > \frac{IS}{N}$$

The SIR model

In the SIR model, besides the S and I states, the Infectious nodes transfer to the immune Recovered state at rate μ: $I \xrightarrow{\mu} R$ ($N = S + I + R$). The Recovered individuals cannot infect other individuals and cannot be infected anymore. As a consequence, the outbreak can stop before infecting all the nodes. The final fraction of Recovered nodes depends on the parameters β and



(a) $R_0 = 1.3$ (b) $R_0 = 10$

Figure 1.1: Numerical solution of the differential equations: evolution of the numbers of Susceptible, Infectious and Recovered nodes over time for two different R_0. The dynamic stops when the number of Infectious nodes goes to zero.

μ. The basic reproduction number defined as $R_0 =< k > \beta/\mu$ can be seen as the average number of individuals an Infectious node will infect before reaching the Recovered state, it describes the competition between the timescales of recovery and of transmission. If $R_0 < 1$, the infection will die out before reaching a significant portion of the nodes, if $R_0 > 1$, the infection is more likely to spread in a population. The corresponding differential equations are:

$$\frac{dS}{dt} = -\beta < k > \frac{IS}{N}$$

$$\frac{dI}{dt} = \beta < k > \frac{IS}{N} - \mu I$$

$$\frac{dR}{dt} = \mu I$$

Figure 1.1 shows the numerical solution of the differential equations of the model for different values of R_0. In the case $R_0 \approx 1$, the final fraction of Recovered nodes (i.e., the size of the outbreak) is far below 50%, whereas in the case where $R_0 >> 1$, the whole population has been infected.

1.2.2 How are simulations performed?

In this thesis, the simulations of epidemic spreading are performed with the SIR model applied in an agent-based approach: the process is stochastic. We perform the simulations either on dynamic or static networks. For each simulation, we choose at random one node (i.e., the seed) to be the first Infectious

node. The simulation process is quite different if we choose to use a dynamic or a static network.

In the case of a dynamic network, we test the transmission of the infection for each interaction between an Infectious and a Susceptible node. Actually, we look at the number $\beta \times \tau$ (where τ is the duration of the time step) and compare this number with a random number between 0 and 1. If this number is higher, the Susceptible node gets infected. For each time step, we also test the recovery of the Infectious nodes by comparing the number $\mu \times \tau$ to a random number between 0 and 1. As previously, the change of state happens if this number is higher than the random number. In our datasets, we have only access to daytime contacts and consider that nodes are no longer in contact when they leave the place of data collection. As a consequence, in the case of epidemic spreading we add "nights" divided in time steps during which Infectious nodes can recover but cannot transmit the infection to Susceptible nodes.

In the static case, we use the weighted aggregated network where the weight of an edge is the total time spent in contact by the two nodes during the time of aggregation. Each weight w_{ij} of the network is divided by the total duration T of the aggregation time. Then we create "days" and "nights" divided in discrete time steps. For each daytime step, we test the transmission of the infection for each Susceptible-Infectious edge by calculating the probability $\beta \frac{w_{ij}}{T} \tau$ and the recovery of the Infectious nodes with the probability $\mu \times \tau$. For each nighttime step, we test only the recovery of the Infectious nodes as above.

In each case, we pursue the simulation until the number of Infectious nodes is zero by looping back on the data if necessary. For the results to be stastically significant, we perform at least 1000 simulations for each set of parameters β and μ. To study the outcome of simulations, we will measure the whole distributions of the epidemic sizes for each set of parameters, the fraction of epidemics with size larger than 20% and the average size of epidemics with size larger than 20% (the cut-off of 20% is chosen as a way to distinguish between small and large epidemics, we can change the value of this cut-off without altering our conclusions).

1.3 Data collection

In the last years, many efforts have been dedicated to the collection of data on human behaviours and contact patterns in various contexts [15]. The recent technological advances have helped the research community to move from traditional methods ranging from diaries and surveys [13, 16–24], to new methods based on wearable sensors able to detect close proximity [13, 25–29] and even face-to-face contacts between individuals [7–12, 30, 31]. These methods allow to collect high-resolution data in an objective way, avoiding biases due to self-reporting [13, 14, 22].

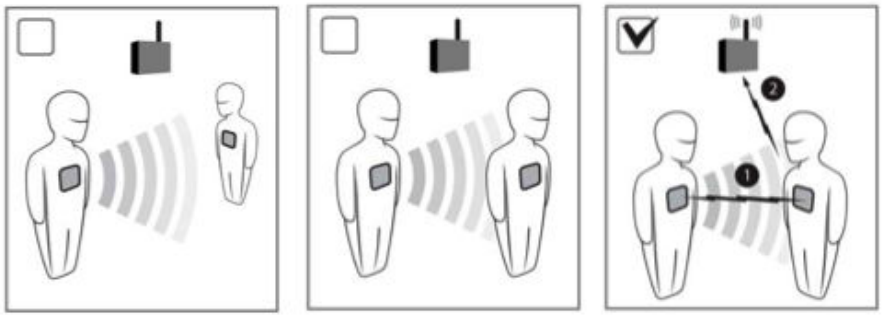


Figure 1.2: Schematic illustration of the RFID sensor system. RFID badges are worn by the participants to the study. A contact is detected when two individuals are close enough and facing each other. The signal of the contact is then sent to an antenna.

1.3.1 The SocioPatterns collaboration

SocioPatterns is a collaboration for interdisciplinary research between researchers and developers from the following institutions and companies: ISI Foundation (Turin, Italy), CNRS - Centre de Physique Théorique (Marseille, France) and Bitmanufactory (Cambridge, UK). It was originally created by Alain Barrat (CNRS and ISI Foundation), Ciro Cattuto (ISI Foundation), Jean-François Pinton (ENS Lyon) and Wouter Van den Broeck (ISI Foundation) in 2008.

The SocioPatterns collaboration developed an infrastructure able to obtain accurate data on face-to-face contacts with high temporal and spatial resolution. This infrastructure is based on wearable wireless sensors working on RFID technology. The participants to the deployment wear the sensors as badges. A contact is detected between two individuals when they are close enough (less than 1.5-2 meters) from each other and are facing each other (the signal is tuned so that the human body acts as an obstacle). The signal is then sent to an antenna (devices are shown in Appendix: Figure A.1). This process is schematized in Figure 1.2. The temporal resolution is 20 seconds. The system is used to gather data in various real-world environments as schools, conferences, workplaces etc.

Further details and information can be found at the SocioPatterns website [6]. SocioPatterns makes available most of the collected data on a dedicated webpage: http://www.sociopatterns.org/datasets/

1.3.2 Description of datasets

The most used datasets in this thesis are face-to-face contacts data collected by the SocioPatterns collaboration in a French high school in Marseille (Lycée Thiers) over three years from 2011 to 2013. The number of participants involved and the duration of study vary from one year to another. The data were collected

in the same environment over the three years but the students changed from one year to the next.

- In 2011 (Thiers11), the study lasted for 4 days (Tuesday to Friday in December 2011) and involved 118 students divided into three different classes.
- In 2012 (Thiers12), the study lasted for 7 days (from a Monday to the Tuesday of the following week in November 2012) and involved the same three classes already involved in the study of 2011 plus two other classes (gathering 180 students).
- Finally in December 2013 (Thiers13), the study lasted for 5 days (from Monday afternoon to Friday) and involved 327 students of nine classes (five classes of 2012 plus four other classes).

For each year, metadata about participants were also collected such as class or gender. Moreover, the dataset of 2013 contains data of different nature:

- At the end of the fourth day (the 5th of Dec.), students were asked to fill in paper contact diaries giving the list of other students they had had contact with (where contact was defined as close face-to-face proximity) during the day in the high school, and to give the approximate aggregated duration of the contacts with each nominated individual, to choose in one of four possible categories: at most 5 minutes, between 5 and 15 minutes, between 15 minutes and 1 hour, more than one hour. 120 students returned a filled in diary (Contact diaries).
- During the period of the deployment, students were asked to give the names of their friends within the high school. Such friendship surveys were obtained from 135 students (Friendship).
- Finally, students were asked to use the Netvizz application to create their local network of Facebook friendships (i.e., the use of the application by a student yields the network of Facebook friendship relations between this student's Facebook friends). 17 students gave access to their local network, from which we removed all users who were not concerned by the data collection (Facebook).

We will also use two other datasets collected by the SocioPatterns collaboration in two different settings.

- InVS (workplace): the study lasted for two weeks (24 June - 5 July 2013) and took place in the office building of the "Institut de Veille Sanitaire" (the French institute for public health surveillance). The study involved 100 individuals structured in 5 departments.
- SFHH: the study took place during the Congress of the "Société Française d'Hygiène Hospitalière" (3-4 July 2009) and involved 403 individuals.

Data set	Type	N (participation rate)	Duration	Dates	Notes
Thiers11	high school	126	4 days	6-9 Dec. 2011	3 classes + teachers
Thiers12	high school	180	7 days	19-27 Nov. 2012	5 classes
Thiers13	high school	327 (86%)	5 days	2-6 Dec. 2013	9 classes
Contact diaries		120 (37%*)	1 day	5 Dec. 2013	
Friendship		135 (41%*)			
Facebook		17 (5%*)			
InVS	workplace	92 (63%)	2 weeks	24 Jun.-5 Jul. 2013	5 departments
SFHH	conference	403 (34%)	2 days	3-4 Jun. 2009	

Table 1.1: Information about data sets used in this thesis. *Participation rate with respect to the number of students who participated to the "Thiers13" data collection.

The main characteristics of each data set is given in Table 1.1. Note that I have actively participated to the deployment of the Sociopatterns infrastructure in Lycée Thiers in 2013 (the locations of antennas is shown in Appendix: Figure A.2) and performed some of the data cleaning on the combined data set of 2013, moreover I have participated to the discussion leading to the data collection performed in the Institut de Veille Sanitaire in 2013 as well as to the analysis performed on the resulting data set (results reported in [32]).

1.4 Overview of the following chapters

This thesis tackles the following topics. In Chapter 2, we present the analysis performed on the three data sets collected with wearable sensors in Lycée Thiers in Marseille. We consider the statistical properties of the networks and study the structure of the network determined by classes. We also compare the contact patterns on two different timescales: we investigate similarities and differences between the different days of one specific data set and between the three years of study. In Chapter 3, we compare different methods of data collection using the combined data set of 2013 which gathers face-to-face contact data obtained with sensors, contact data obtained with contact diaries, reported friendships and Facebook links. We question the possibility of using the information contained in contact diaries and friendship data to obtain an accurate estimation of the epidemic risk. In Chapter 4, we investigate whether the underestimation of the epidemic risk obtained with the friendship network may be seen as biases due to a sampling process performed on the contact network obtained from sensors. In Chapter 5, we give a short conclusion and future perspectives.

Chapter 2

Analysis of face-to-face proximity data

In this chapter, we present the results of the quantitative analysis performed on the three high school datasets (Thiers11, Thiers12, Thiers13) describing the contacts between the students. We investigate the mixing patterns of students and how it is driven by the repartition of students into classes.

We compare the contact patterns on two different timescales: on the one hand, we examine the similarities and differences between the different days of one single deployment; on the other hand, even if people change from one year to the next, we take advantage of the fact that we collected data in the same environment in three consecutive years and study the long term stability of contact patterns in this high school.

We finally investigate if gender differences have an impact on contact patterns, as observed in primary school [33].

This chapter covers the results reported in the following paper: *Contact Patterns among High School Students*, published in PLoS ONE in September 2014 [34], in which we analyzed only the data sets of 2011 and 2012; this chapter presents the analysis performed on the data set of 2013 and additional results.

2.1 Study context

The students involved in the study were all part of classes called "classes préparatoires". This type of studies is specific to the French schooling system and gathers students after the end of the end of the usual high school studies. During the two years of these studies, the students prepare for competitive exams yielding admission to various higher education colleges. The studies take place in a high school environment but students are separated from the younger high school students. In fact, classes are located in a different part of the high school building and they take lunches separately. The study gathered only students of second year of this specific school cycle. Indeed, students in second year of these

Class name	Number of individuals	Male	Female
PC	31	16	15
PC*	45	32	13
PSI*	42	32	10
teachers	8	5	3
Total	126	85	41

Table 2.1: Classes involved in the 2011 data collection. The study lasted for 4 years in December 2011 (from a Tuesday to the Friday).

Class name	Number of individuals	Male	Female
PC	38	24	14
PC*	35	26	9
PSI*	41	29	12
MP*1	31	27	4
MP*2	35	27	8
Total	180	133	47

Table 2.2: Classes involved in the 2012 data collectionThe study lasted for 7 years in November 2012 (from a Monday to the Tuesday of the following week).

studies have to prepare a small project. Some of them based this project on their participation to the data collection as well as the use of collected data for some small scale analysis and numerical simulation. Thanks to their involvement in the data collection, the participation of students was close to 100%.

The various classes participating to the study focus on different topics: "MP" classes focus on mathematics and physics, "PC" classes on physics and chemistry, "PSI" classes on engineering and finally "BIO" classes focus on biology. Tables 2.1-2.3 report the class names and the number of individuals of each gender in each class involved in the study. We can notice that there are much more males than females in "hard sciences" classes while the contrary is observed in biology classes. Note that all classes involved in the study of 2011 (resp. 2012) are involved in the study of 2012 (resp. 2013) and that the three studies took place at similar periods of the year; these two facts make the comparison possible between the different studies.

In order to avoid repetitions, we mainly report results of analysis performed on the 2013 dataset. Supplementary results for 2011 and 2012 can be found in the Appendix.

Class name	Number of individuals	Male	Female
PC	44	26	18
PC*	39	23	14
PSI*	34	24	10
MP*1	29	23	6
MP*2	38	32	6
MP	33	18	12
2BIO1	36	8	28
2BIO2	34	13	20
2BIO3	40	8	32
Total	327	175*	146*

Table 2.3: Classes involved in the 2013 data collection. The study lasted for 5 years in December 2013 (from a Monday to the Friday).*6 students did not give information about their gender.

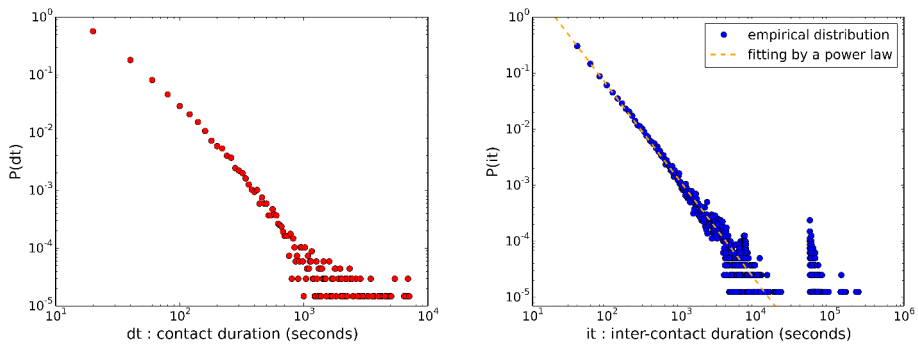
2.2 Number and durations of contacts

During the 5 days of data collection of 2013, 67,613 contact events were registered, corresponding to a cumulative duration of 3,770,160s (approximately 1,047 hours). Table 2.4 reports the number and cumulative duration of contacts registered in each day. Figure 2.1 reports the distributions of contact and inter-contact durations (time intervals between successive contacts of an individual) measured over the whole data collection. We recall that a contact event is defined as an uninterrupted sequence of 20-seconds time steps where the two individuals are in contact. Most contacts are short, but contacts of very different durations are observed, including very long ones. While the average duration of a contact is 56 seconds, and 83% of the contacts last less than 1 minute, approximately 2% of the contacts last at least 5 minutes. The strong variability in contact durations is shown by the large value of the squared coefficient of variation of the distribution, $CV^2 = 7.4$. In fact, the distribution is heavy-tailed and can be approximated by a power law: as already observed in previous studies (e.g., [7, 26, 32]) measuring the durations of contact events between individuals. We cannot define a characteristic contact time scale. In other words, the average contact duration is not a good representation of the actual duration of contacts because both much shorter and much longer contacts can be observed with non-negligible probabilities. That being said, the duration of contacts play a leading role when studying how outbreaks spread within a population. In fact, the transmission probability of an infectious disease between two individuals is generally assumed to depend on the time they spend in contact: the longer they are in contact, the higher is the probability of infection. However, it is also well known that the weak ties should not be neglected in the models of epidemic spreading [35].

On the other hand, the distribution of inter-contact durations is also very broad: most intervals between periods of activity are very short, but very long

	Number of contacts	Cumulative duration of contacts			N	E
Day	Number (% of total)	Seconds (% of total)	Minutes	Hours		
Monday (afternoon)	10,539 (15.6)	575,600 (15.3)	9,593	160	312	2242
Tuesday	16,702 (24.7)	946,760 (25.1)	15,779	263	310	2573
Wednesday	14,499 (21.4)	803,480 (21.3)	13,391	223	303	2161
Thursday	13,317 (19.7)	745,580 (19.8)	12,426	207	295	2162
Friday	12,556 (18.6)	698,740 (18.5)	11,646	194	299	2075
Total	67,613	3,770,160	62,836	1,047	327	5818

Table 2.4: Number and duration of contacts in the different days of the 2013 data collection that lasted 5 days.



(a) (b)

Figure 2.1: (a) Distribution $P(dt)$ of contact durations for 2013 data i.e., probability for a registered contact to last dt, (b) Distribution $P(it)$ of inter-contact durations i.e., probability that the time elapsed between two successive contacts of a node is it. The peak on the right correspond to inter-event durations of one or two nights.

durations are also observed. It can be approximatively fitted by a power law. This highlights the burstiness of human contacts, already observed in various contexts where human behavior is involved [36, 37].

2.3 Contact matrices

Figure 2.2 reports, in the form of contact matrices, the cumulative durations and the total numbers of contacts between classes of individuals, measured over the whole study duration. The second and third rows take into account the different numbers of individuals in each class yielding asymmetric matrices. The matrices have a strong diagonal strucure: most contacts occur between students of the same class (92% of all contacts). These results shows the strong assortativity of contacts with respect to class and are in agreement with the results obtained in other school environments [10, 30, 31], moreover in many complex agent-based models this assortativity is assumed as an important feature of contact networks

[38]. On the contrary, very few contacts occur between students of different classes.

Moreover, we can notice more contacts within three groups of classes: the classes of Biology (2BIO1, 2BIO2, 2BIO3), the classes of Mathematics (MP, MP*1, MP*2) and the classes of Physics (PC, PC*, PSI*). This substructure can have two origins: first, the topics studied by classes of each group are similar; moreover, the classrooms of each group are physically close in the high school.

However, if the number and duration of contacts are very different inside and outside classes, the distributions of contact durations are very similar: these distributions are broad with strong fluctuations. The main difference is that the maximal contact duration within classes is higher than the maximal duration of inter-classes contacs (Figure 2.3).

The contact matrix representation of the data in Figure 2.2 gives a picture in which all students in class X are in contact with all students of class Y, even if the numbers and durations of contacts vary strongly. Figure 2.4 shows the numbers and densities of edges between pairs of classes. We observe that the density of edges is instead very small for distinct classes, and that it is still far from 1 inside each class, even if it takes much larger values. This shows the interest of investigating contact patterns in more detail by studying the contact network structure.

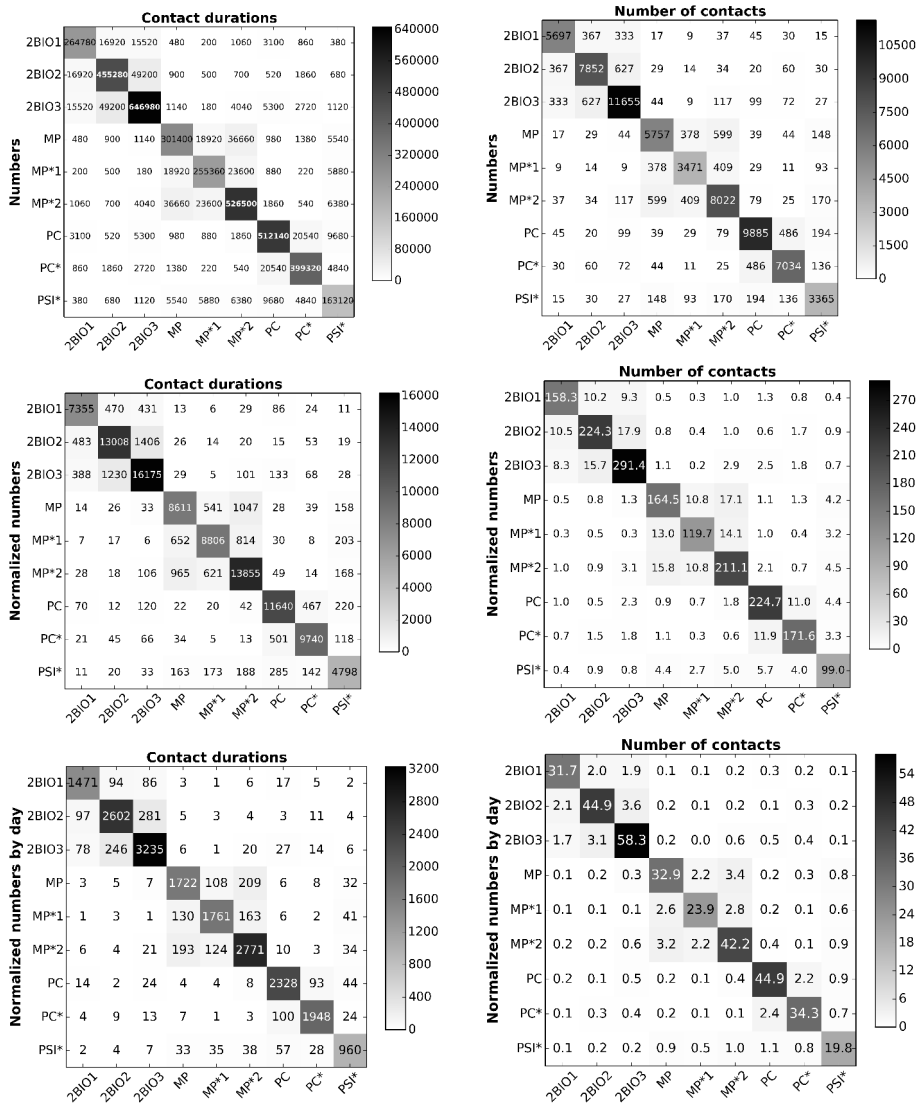


Figure 2.2: Contact matrices giving the cumulated durations in seconds (first column) and the numbers (second column) of contacts between classes during the whole study. In the first row, the matrix entry at row X and column Y gives the total duration (resp. number) of all contacts between all individuals of class X with all individuals of class Y. In the second row, the matrix entry at row X and column Y gives the average duration (resp. number) of contacts of an individual of class X with all individuals of class Y. In the third row, we normalize each matrix element of the second column matrices by the duration of the study, in days, to obtain at row X and column Y the average daily duration (resp. number) of contacts of an individual of class X with individuals of class Y.

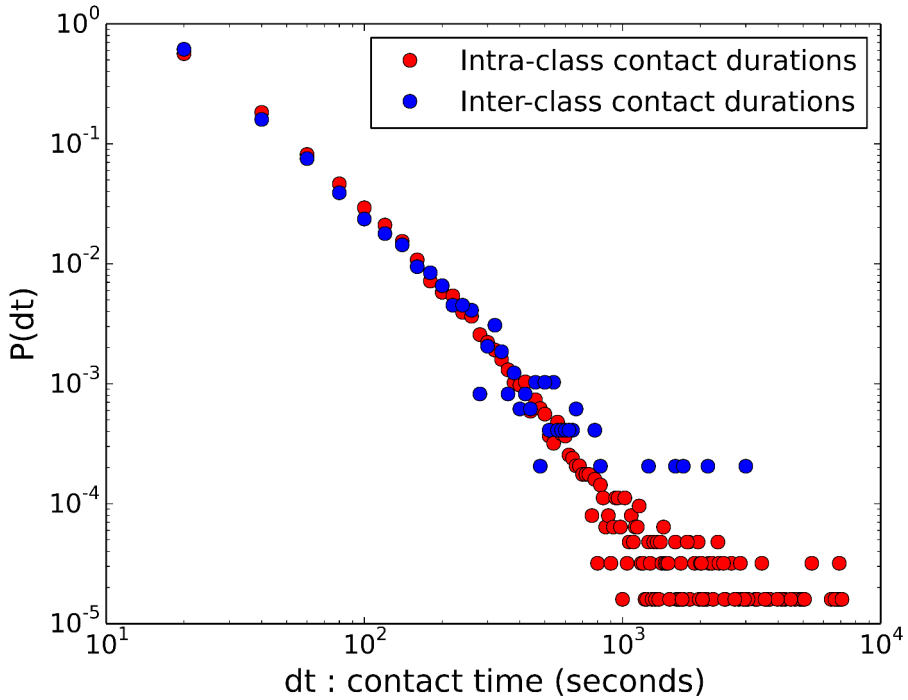


Figure 2.3: Distributions $P(dt)$ of contact durations for intra-class contacts and inter-class contacts.

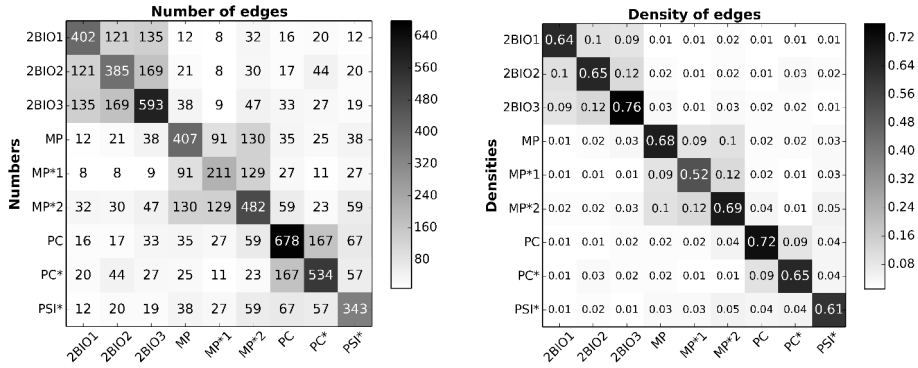


Figure 2.4: Contact matrices of edge numbers and densities. Left: the matrix entry at row X and column Y gives E_{XY}, i.e., the number of pairs of individuals of classes X and Y who have been in contact at least once during the study. Right: the matrix entry at row X and column Y gives ρ_{XY}, i.e., E_{XY} normalized by the maximal possible number of pairs of individuals of classes X and Y.

2.4 Contact network

The contact network representing the data set of 2013, aggregated over the whole study duration, has 327 nodes representing the 327 students, and 5818 edges corresponding to the pairs of students who have been in contact at least once during the data collection. The contact network displays the properties of a small-world network: the average shortest path length in this network is equal to 2.15, and its clustering coefficient is equal to 0.503 (a random network with the same number of nodes and edges would have a clustering coefficient ≈ 0.11).

Figure 2.5 shows a spatialization of the network obtained with the Force Atlas algorithm of the Gephi software; this algorithm helps to visualize network's structure by using a model in which nodes repulse each other (like magnets) while edges attract the nodes they connect (like springs). This layout highlights the strong modular structure of the network in classes. Besides the structure in classes, we can distinguish the additional substructure of the 3 groups of classes: only the class PSI* seems to be equally close to two groups (MP classes and PC classes).

Figure 2.6 displays the distributions of nodes' degrees and of edges' weights in the global aggregated network (i.e., aggregated over the whole duration of the study). The average degree of nodes in the aggregated network is equal to 35.6. The contact network is homogeneous in terms of degrees as seen in the distributions of nodes' degrees: the degree distribution is narrow ($CV^2 = 0.14$) as observed in many empirical networks of human contacts in various contexts [8, 10, 11, 26, 32]. The distribution of weights, on the other hand, shows strong fluctuations ($CV^2 = 14.9$). It is a heavy-tailed distribution that can be fitted by a power law. The average amount of time spent in interaction by two persons is 648 seconds (10 min 48 s) during the whole study duration, but very different values of weights can be observed. Most cumulated durations are short (62% of the pairs of individuals who have interacted at least once have been in contact less than 2 minutes over the whole data collection period), but large values are also observed: 16% have spent more than 10 minutes in contact and 4% more than 1 hour. This strong heterogeneity is not entirely due to the structure in classes of the network. In fact, the same heterogeneity is also observed if we restrict the distributions to the links between students of the same classes or between students of two different classes as shown in Figure 2.7. The main difference is the maximal edge weight which is higher for intra-class edges than for inter-class edges.

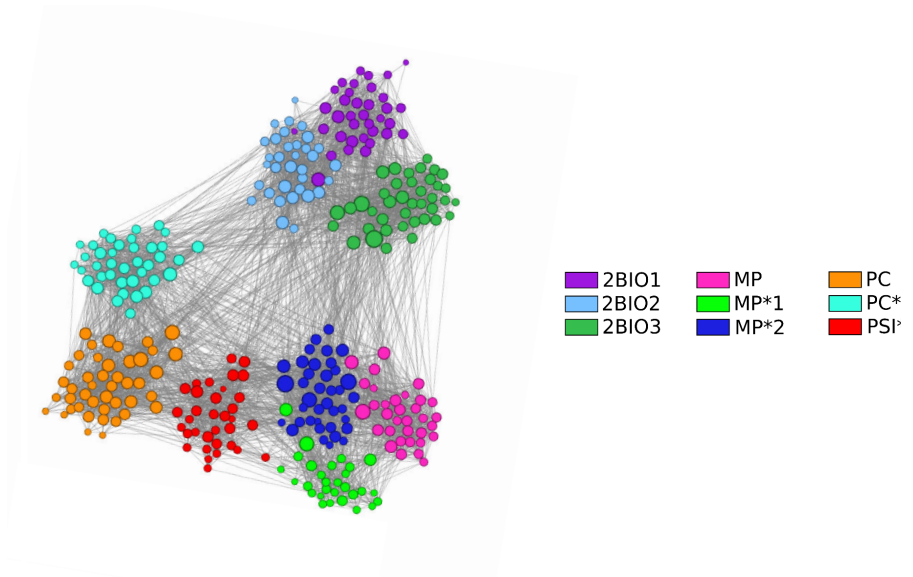
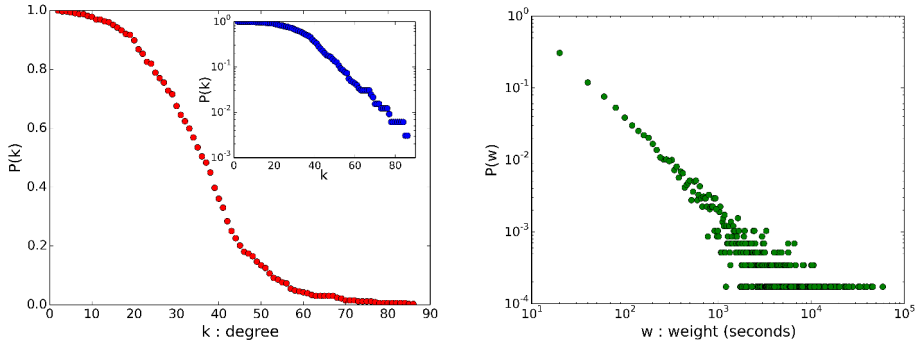
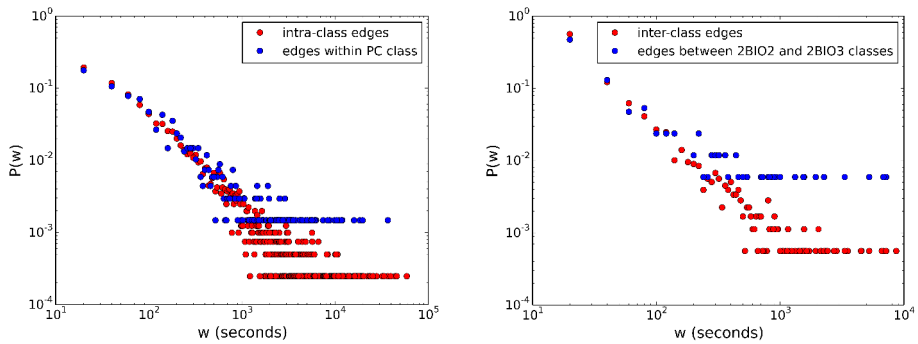


Figure 2.5: Representation of the network of contacts between students, aggregated over the whole study duration for 2013. Each node represents a student, its color corresponds to the student's class and its size represents its degree. Created using the Gephi software: http://www.gephi.org.



(a) (b)

Figure 2.6: (a) Complementary Cumulative Distribution Function (CCDF) $p(k)$ of nodes' degrees (inset: same distribution in log-lin scale) i.e., probability for a node to have degree greater or equal to k. (b) Distribution $p(w)$ of edges' weights.



(a) (b)

Figure 2.7: Distributions $p(w)$ of edges' weights for (a) intra-classes edges and edges within one specific class and for (b) inter-classes edges and edges between a specific pair of classes.

2.5 Gender homophily

In this section, we investigate if new specific structures emerge from the individual characteristics of students such as gender, as already observed for classes. The term homophily refers to the preference that individuals exhibit when they interact and build social ties with peers they consider to be alike. It is a well-known feature of human behavior and has been studied in many contexts [31,39]. This homophily can be associated to the gender, the age or the political opinions. In sociology, it is a well known fact that children tend to exhibit gender homophily at school, and that this tendency decreases with age, especially during adolescence. Statistical evidence of gender homophily has also been obtained in a high-resolution time-resolved data set describing face-to-face proximity of children in a primary school [33]. The present data set describes the interactions of young adults in a high school context, and it is therefore of interest to investigate the possible presence of gender homophily with the same methods.

Figure 2.8 displays contact matrices giving the normalized numbers of contacts and the densities of edges between individuals of given units (class + gender). Each unit is obtained by dividing each class in two groups according to the students' gender. Note that we remove the 6 students for whom the gender is unknown. The resulting network has 321 nodes and 5615 edges. Moreover, we look at each group of classes (groups: Biology stands for 2BIO1, 2BIO2 and 2BIO3 classes, Mathematics stands for MP, MP*1 and MP*2 classes and Physics stands for PC, PC* and PSI* classes) separately for more clarity; it is not a big issue because there are very few contacts between different groups. As the numbers of male (M) and female (F) students are strongly different (see Table 2.3), with much more male than female students (except for Biology classes where it is the contrary), we consider normalized contact matrices: for the number of edges, we normalize by the maximum number of edges between

Group of classes	Biology	Mathematics	Physics	**All classes**
MM links	8.9%	58.5%	40.7%	32.9%
FF links	53.1%	6.3%	15.1%	26.4%
MF links	38%	35.2%	44.2%	40.7%
Null model				
MM links	6.8%	55.6%	40.9%	32.3%
FF links	54%	6.2%	12.7%	24.8%
MF links	39.2%	38.2%	46.4%	42.9%

Table 2.5: Percentage of male-male links (MM), female-female links (FF) and male-female links (MF) in each group of classes for the original data and in the null model. Here we consider only links within each group of classes and not links between two different groups of classes.

each pair of groups and, for the contact durations, each matrix element at row X and column Y is normalized by the number of individuals in group X in order to give the average time spent by a member of group X with individuals of group Y.

The contact matrices display a block diagonal form, each 2x2 block corresponding to a class. The number and durations of contacts among female students is slightly lower than among male students; in the global aggregated contact network, 32.9% of the edges join two male students, 26.4% join two female students, and 40.7% are between students of different gender. In Table 2.5, we show these percentages within each group of classes.

These values seem to indicate a preference for contacts with students of the other gender among male students and for students of the same gender for female students in Biology classes while it is the contrary in MP and PC classes. This appears as well through the distribution of the same gender preference index P_{sg}: for each individual, this index is defined as the fraction of edges, in the aggregated contact network, with individuals of the same gender. The corresponding distributions are shown separately for male and female students as boxplots and for each group in Figure 2.9 for the contact network aggregated over the whole data collection, averages of these distributions are shown in Table 2.6. In Biology classes, the fraction of same-genders neighbors is much higher for female students than for male students while the opposite is observed in Mathematics classes. In Physics classes, the P_{sg} is higher for males than for females but the difference is not as striking. For the whole network, the P_{sg} is a little higher for males than for females.

To interpret these values, we need however to take into account the very strong imbalance between male and female students inside classes, which clearly plays an important role here. For instance, we recall that there are much more female students in Biology classes. In the limit of a very small fraction of male students, the fraction of male-male interactions would be negligible even in the case of a fully homogeneous, gender-indifferent, mixing of individuals. We therefore consider a simple null model, given by a graph with the same

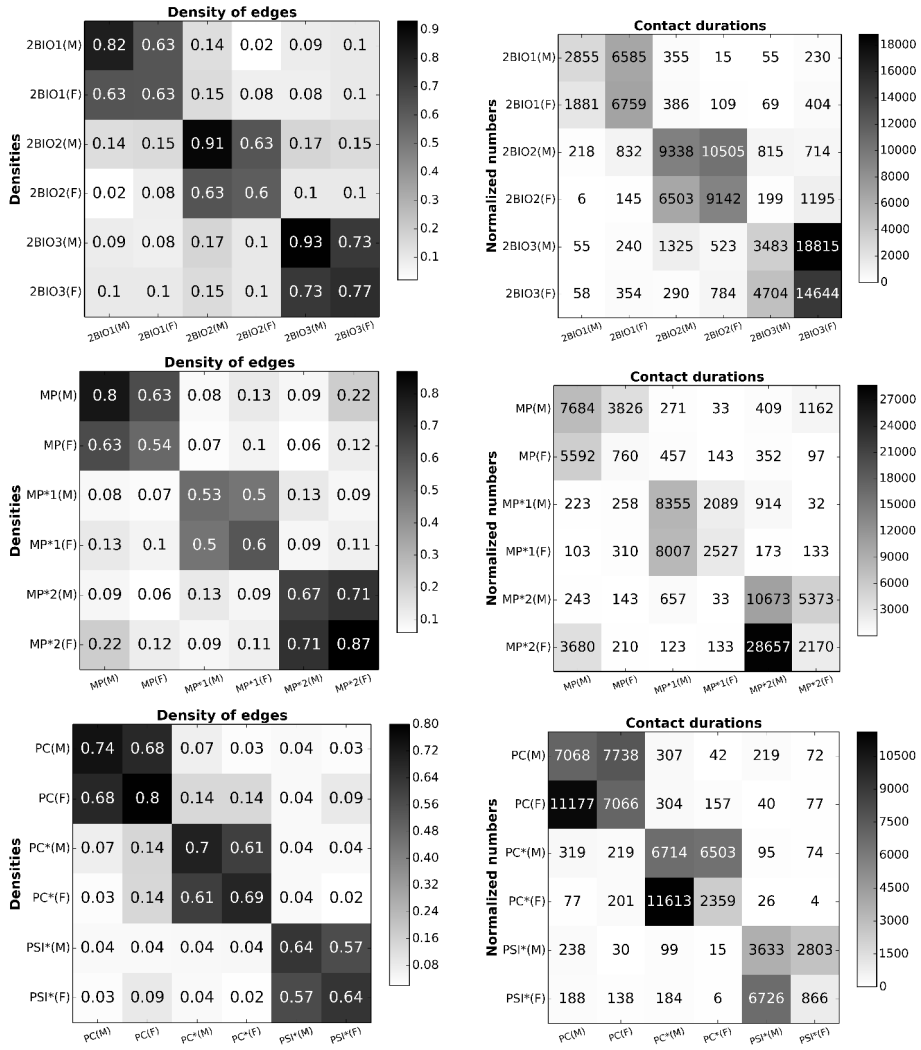


Figure 2.8: First column: Densities of edges between groups of individuals (class + gender) of the aggregated network for Biology (1st row), Mathematics (2nd row) and Physics classes (3rd row). Second column: Normalized numbers of contacts for the whole study duration for Biology, Mathematics and Physics classes.

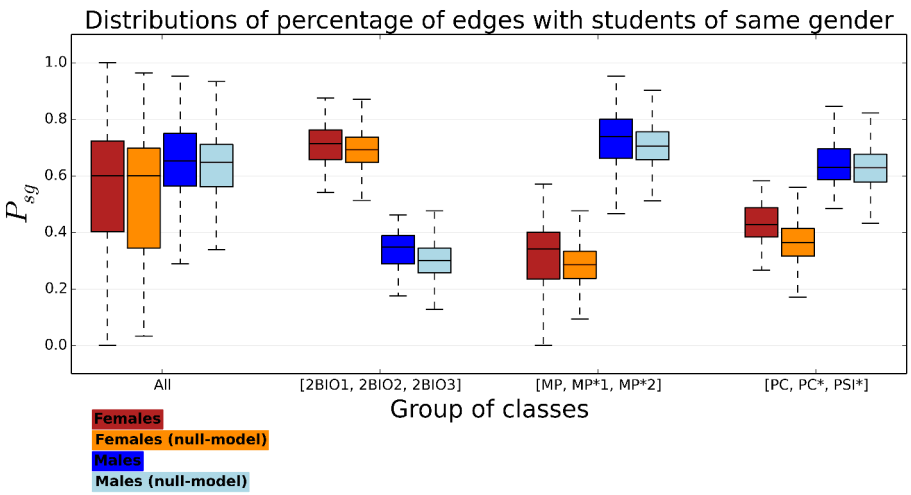


Figure 2.9: Boxplots showing the distributions of the fraction of edges with students of same gender for males and females with empirical data and using the null-model described below for each class. The centre of each box indicates the median of the distribution, its extremities the 25% and 75% quartiles.

Group of classes	Biology	Mathematics	Physics	**All classes**
Males	34.5	72.1	63.3	62.2
Females	70.4	32.5	43.9	56.5
Males (reshuffled network)	30.1	70.5	62.7	60.6
Females (reshuffled network)	69.0	28.6	36.6	53.0

Table 2.6: Average same gender preference index for males and females of each group of classes, for the original data and in the null model.

number of nodes and edges but randomly placed edges with a supplementary condition: we keep the matrix of number of edges between groups of classes unchanged, e.g., an edge between a Biology class and a Mathematics class will remain between a node of Biology classes and a node of Mathematics classes. As a result, the fractions of edges joining male students, female students and students of different gender in the reshuffled network are given in Table 2.5. The results are averaged over 1000 realizations of the null model.

Moreover, the distributions and averages of the same gender preference index in the reshuffled network are shown in Figure 2.9 and Table 2.6. The average value of the same gender preference index is systematically slightly lower in the null model than in the empirical data for both male and female students. This tends to indicate a slight tendency towards gender homophily for both genders. The boxplots displayed in Figure 2.9 show however that this tendency is not statistically significant, and that the observed data is in fact compatible with a null hypothesis of absence of homophily and of gender indifference in the contact patterns of the students. We also note that keeping only the links corresponding to a weight larger than a given threshold, corresponding e.g., to an aggregated interaction duration of 2 or 5 minutes over the study duration, changes the number of edges of each type but does not change the results concerning the absence of gender homophily.

Another way (and reason) to assess the presence of homophily in the classes is related to the information of epidemiological models by data on human contacts. As appears clearly from the contact matrix and contact network analysis, the population of high school students is far from being homogeneously mixed, and it is certainly relevant to use a level of description in which students are divided into groups corresponding to their respective classes. If a strong gender homophily were to be observed, corresponding to the presence of a strong group substructure inside classes, it would as well be important to consider such substructure in models of contact patterns in order to describe spreading phenomena in such population. The assessment of such properties is thus related to the issue of the amount of information needed to inform models, as discussed in [8, 40, 41]. While a detailed analysis of spreading processes in the population under study goes beyond the scope of the present investigation, we investigate this point through numerical simulations of a simple SIR spreading process: we have performed these simulations on two different representations of the network. In the CM representation, the weights of links are taken from the matrix of contact durations in which nodes are divided into classes: all links corresponding to a cell of the matrix have the same weight which the average weight. The CMg representation is almost the same, except we take the matrix where nodes are divided into classes and gender. The results are shown in Figure 2.10. We do not observe any difference between the two representations: hence, a description of the contact patterns at the level of classes corresponds to a sufficient resolution and there is no need to represent the students population at the finer level of a division into gender groups in each class.

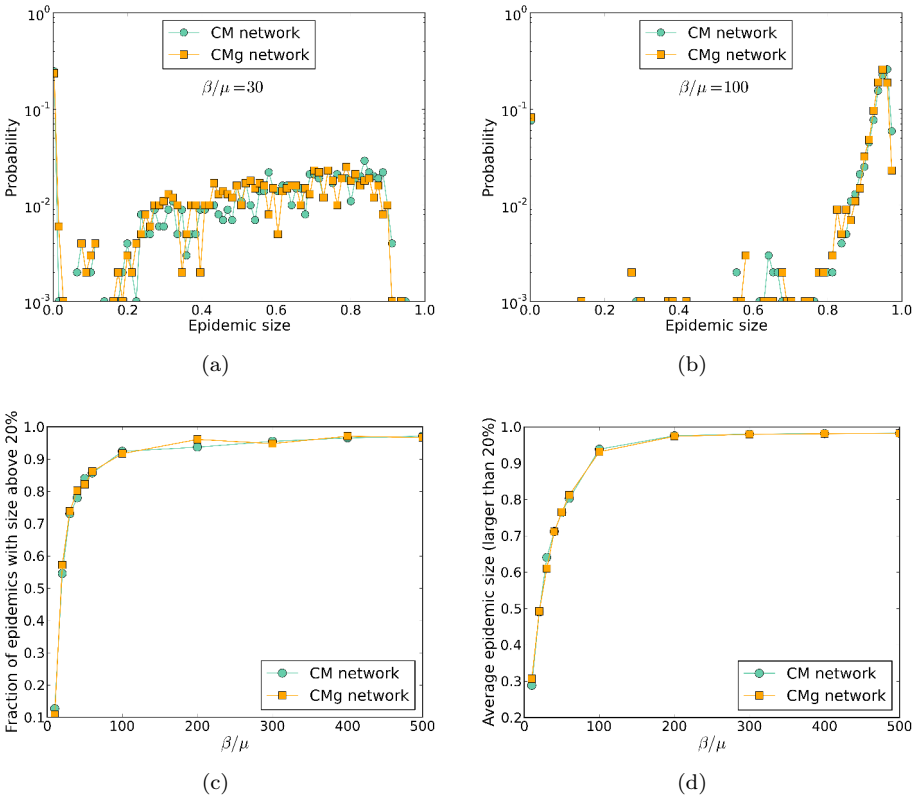


Figure 2.10: Results of the simulation of a stochastic SIR process with different values for β/μ using different representations of the contact patterns between students. (a) Distributions of epidemic sizes for $\beta/\mu = 30$. (b) Distributions of epidemic sizes for $\beta/\mu = 100$. (c) Fraction of epidemics with size above 20% (at least 20% of recovered individuals at the end of the SIR process) as a function of β/μ. (d) Average size of epidemic with size above 20% as a function of the parameter of spreading β/μ. (Representations: In the CM network, the weight of links are taken from the matrix of contact durations in which nodes are divided into classes. In the CMg network, the weight of links are taken from the matrix of contact durations in which nodes are divided into classes + gender.)

2.6 Longitudinal analysis at daily scale

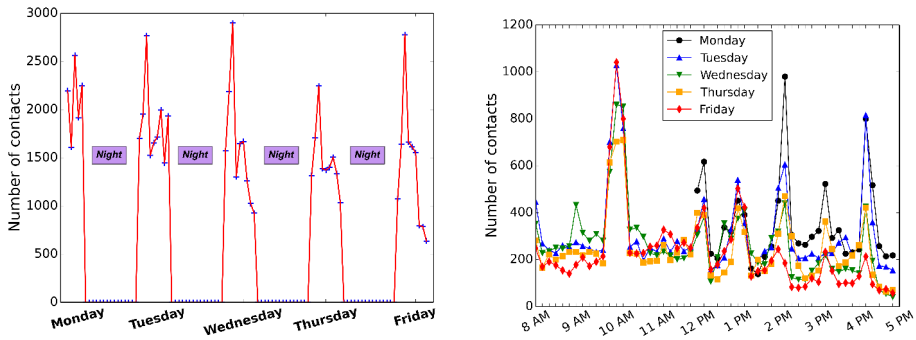
Let us now turn to the longitudinal analysis of the dynamic network. Actually, our data sets allow us to have an instantaneous picture of the contact network every 20 seconds. Then we can aggregate these snapshots over different time windows. The study of the similarities and differences of contact patterns depending on the length of the aggregation window is of particular interest as it can help us to understand for instance how much information is lost if data are gathered only during one single day or few days, and how much data gathering is needed to inform models of human behavior.

2.6.1 Temporal evolution of contact patterns

In this section, we look at the time evolution of properties of the network. Figure 2.11 reports the evolution of the number of contacts at two different temporal resolutions: on the left we show this evolution over the whole study duration per one-hour time windows, on the right we look at the evolution over the course of each day discretized in 10-minutes time windows.

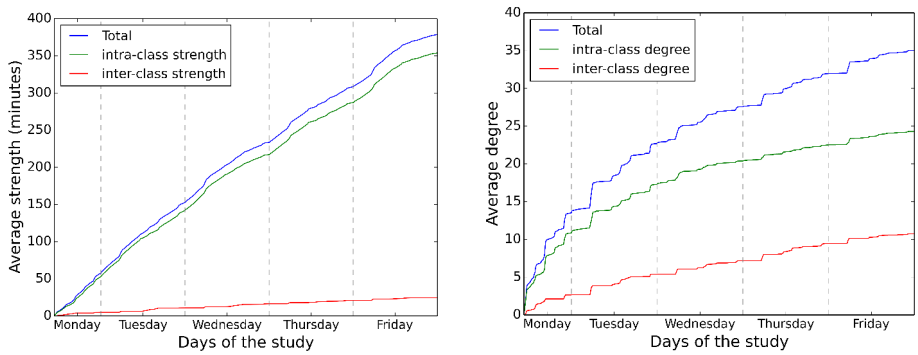
The number of contacts fluctuates strongly over the course of each day. Class breaks and lunch breaks are determined by strong peaks of activity. As the students leave the area of deployment after the end of lectures (around 5PM), the activity drops to zero at night. However, the evolution pattern is very similar from one day to another with peaks at the same time of the day, a feature already observed in other contexts [12].

Figure 2.12 shows the time evolution of the average strength and degree during the five days of study. The total time spent in contact by an average student grows regularly over time, both with students of the same class and with students of different classes, showing that the average amount of time spent in contact each day by a student does not fluctuate strongly from one day to the next, as also observed in a primary school [10]. However, the average strength is much higher for intra-class edges than for inter-class edges. Notably, the average number of distinct individuals with whom a student has been in contact also displays a strictly increasing behavior over the whole study duration, with no clear saturation trend. Note that Figure 2.12 shows average values: at each time the distribution of degrees is similar to the one displayed in Figure 2.6(a), the average value of this distribution increases as time increases but the shape remains similar. This means that an average student continues to meet new persons each day, and that his/her neighborhoods in the contact network, i.e., his/her individual contact patterns, change from one day to the next. The figure also shows that students meet a larger number of distinct individuals of the same class than of other classes, as expected from the previous analysis of contact matrices and networks, and that both numbers continue to grow during the whole study.



(a) (b)

Figure 2.11: (a) Evolution of the number of contacts per one-hour time-windows during the study. (b) Evolution of the number of contacts per 10-minutes time-windows for each day of the study.



(a) (b)

Figure 2.12: (a) Time evolution of the average time spent in contact (strength) by a student during the study. (b) Time evolution of the average degree during the study.

(a)

Cosine similarity b/t contacts matrices	Monday	Tuesday	Wednesday	Thursday	Friday
Monday	1	**0.98**	**0.98**	0.97	0.95
Tuesday	**0.98**	1	0.96	0.95	0.94
Wednesday	**0.98**	0.96	1	**0.93**	**0.93**
Thursday	0.97	0.95	**0.93**	1	0.94
Friday	0.95	0.94	**0.93**	0.94	1

(b)

Cosine similarity b/t contacts matrices	Monday	Tuesday	Wednesday	Thursday	Friday
Monday	1	**0.83**	0.75	0.80	0.72
Tuesday	**0.83**	1	**0.64**	0.77	0.68
Wednesday	0.75	**0.64**	1	0.71	0.73
Thursday	0.80	0.77	0.71	1	0.82
Friday	0.72	0.68	0.73	0.82	1

Table 2.7: (a) Cosine similarities between contact matrices of the first column of Figure 2.13. (b) Cosine similarities between contact matrices in which diagonal elements are ignored. The numbers in blue (resp. red) are the maxima (resp. minima) of the table.

2.6.2 Comparison of daily patterns

We define contact matrices for each day, each morning (before 12PM) and each afternoon (after 12 PM) in Figure 2.13. Note that the study started at 12 PM the monday; as a consequence, we do not define morning or afternoon matrices for this special day. The number of contacts between classes fluctuate from one day to the other and between morning and afternoon. However, the structure of the contact matrices presents a robust pattern, with higher values on the diagonal and the additional substructure corresponding to the 3 groups of classes, already observed in the contact matrix aggregated over the whole study duration. To quantify more precisely this observation, we compute the cosine similarities between (i) pairs of daily contact matrices and (ii) morning and afternoon contact matrices for each day. The values are given in Table 2.7(a), they are very large with a minimum of 0.93. The similarities between morning and afternoon contact matrices are also very high with minimum at 0.85 (Table 2.8). We also check if these large values are only due to the diagonal elements of the matrices as they take much larger values than the off-diagonal elements. The corresponding values, given in Table 2.7(b), are still large with a minimum of 0.64 and a maximum of 0.83. This shows that, despite the fluctuations in the number of contacts, the structure of the contacts between classes is very robust across different days, and is well captured by a data collection performed on any given day.

This robustness is also observed in the statistical properties of the contact networks of different days. We investigate this point in Figure 2.14. The distributions of weights, durations of contact events and inter-contact durations overlap from one day to the other; the fitting by a power law yields similar ex-

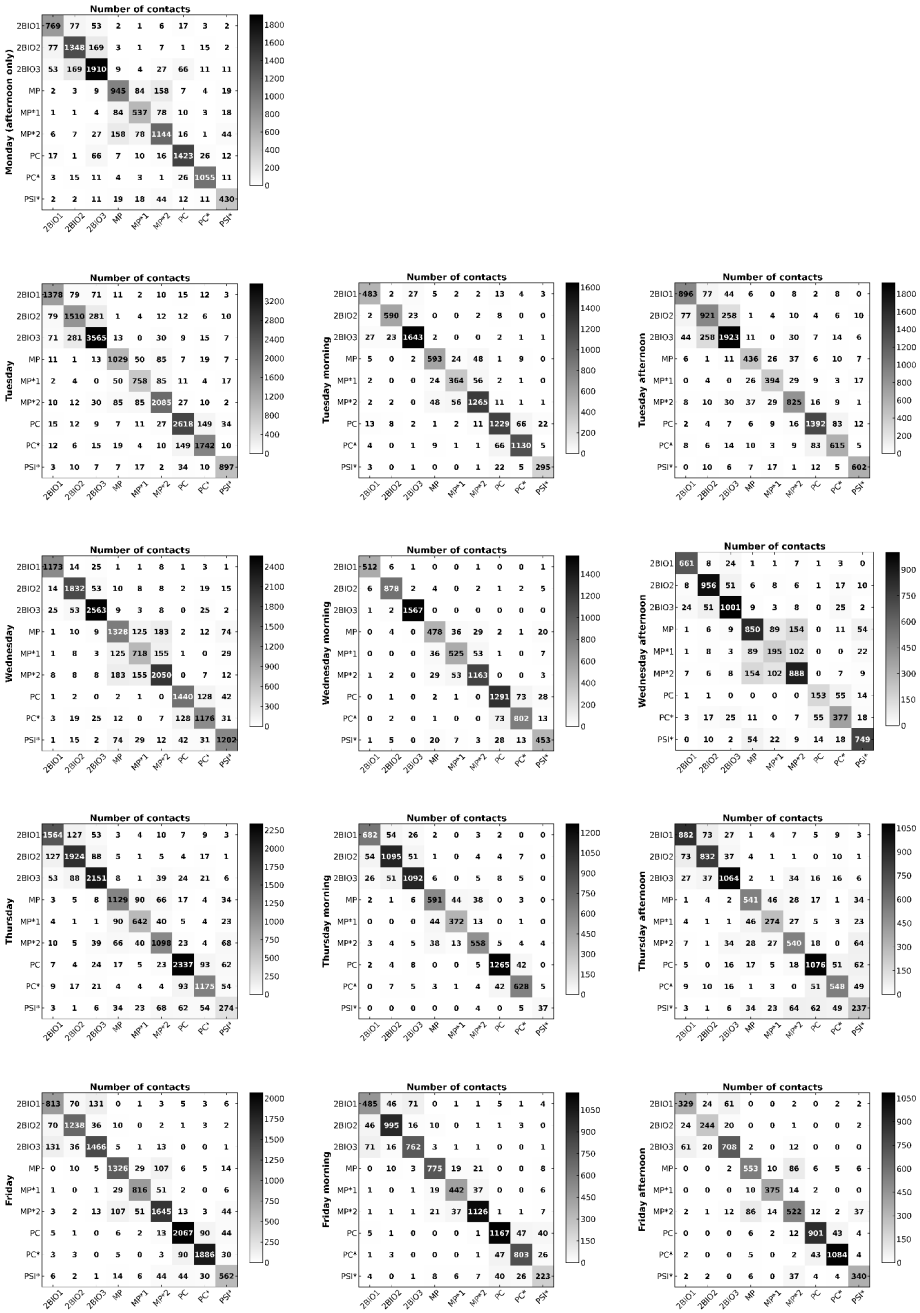


Figure 2.13: First column: Contact matrices giving the numbers of contacts between classes for each day of the study. Second column: contact matrices for each morning (before 12:00 p.m.). Third column: same for each afternoon (after 12:00 p.m.). The matrix entry at row X and column Y gives the total number of contacts between all individuals of class X with all individuals of class Y during the aggregation interval (one day, one morning or one afternoon).

Cosine similarity b/t morning and afternoon	
Tuesday	0.94
Wednesday	0.85
Thursday	0.98
Friday	0.90

Table 2.8: Cosine similarities between each pair of morning-afternoon contact matrices of Figure 2.13 (2nd and 3rd columns).

ponents for different days. The degree distributions have different average values but the shapes are similar; when rescaled by the average value the distributions are superimposed.

The previous figures display distributions and aggregated measures of the network; we now turn to a comparison at the individual level. Figure 2.15 displays boxplots of distributions of cosine similarities. These cosine similarities measure the change in nodes' neighborhoods between each pair of different days of the study. The similarity between the neighborhoods of an individual i in the contact networks measured in two different days denoted 1 and 2 is measured through the cosine similarity defined this way:

$$\sigma^{1,2}(i) = \frac{\Sigma_j w_{ij,1} w_{ij,2}}{\sqrt{\Sigma_j w_{ij,1}^2}\sqrt{\Sigma_j w_{ij,2}^2}}$$

The distributions shown in Figure 2.15 are obtained in the following way: in each class, for each person we calculate the cosine similarities of his/her neighborhood for each pair of days. Note that we distinguish intra-class and inter-class neighborhoods. Each distribution thus corresponds to 10(couples of days)*N(number of students in the class) similarity values. Cosine similarities restricted to intra-class neighborhoods tend to be larger than the ones restricted to inter-class neighborhoods, indicating a slightly larger stability of intra-class neighborhoods.

In both cases, the values of cosine similarities are rather far from 1, indicating than the neighborhoods of each student change significantly accross the days. However, without comparison to any reference values, we cannot state on the character of these values. In other terms, to know if these empirical values should be considered "small" or "large" , we need to compare them to values obtained with null models. Table 2.9 lists the null models used. We first consider null models in which the network edges are placed at random between the nodes of the networks. We also consider edge rewirings which conserve the degree of each node ("Sneppen-Maslov" null model [42]). Finally, we keep the network topology unchanged but we reshuffle the weights on the edges. In the (b) version of each null model, the additional constraint is to keep the contact matrix of link densities unchanged.

The distributions of the cosine similarities obtained through the use of null models are shown as boxplots in Figure 2.16. The empirical values are much larger than the ones obtained with the null models, even with the last null

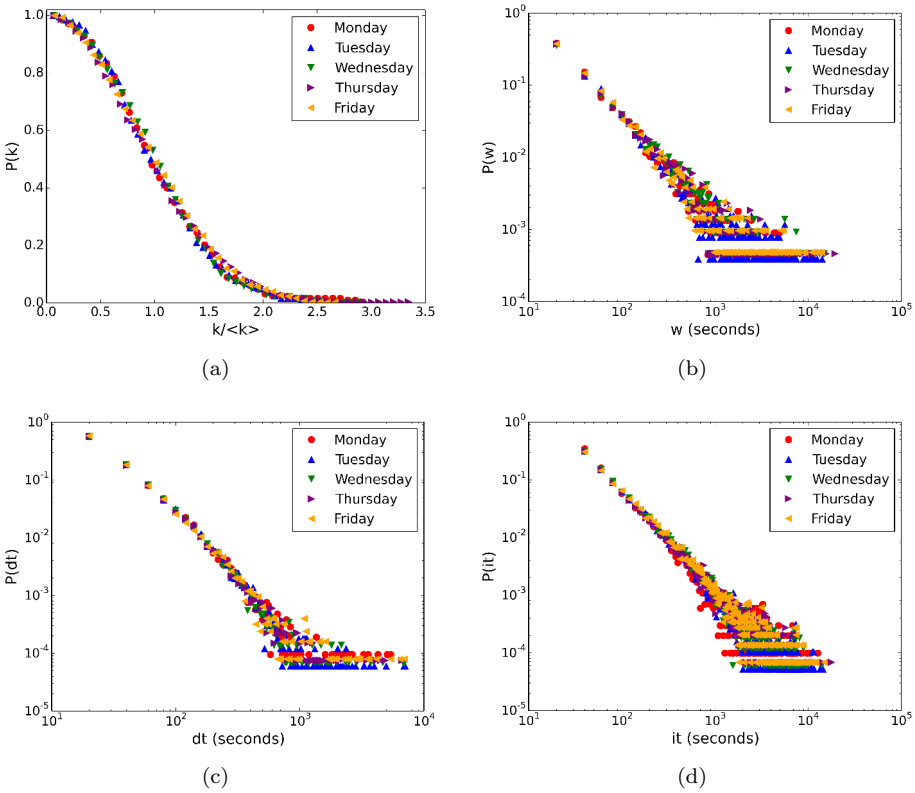


Figure 2.14: Properties of daily aggregated networks. (a) CCDF of nodes' degrees of the daily aggregated networks rescaled by the average degree $< k >$ of each daily network. (b) Distribution of edge weights for each daily network. (c) Distribution of contact durations in each day. (d) Distribution of inter-contact durations.

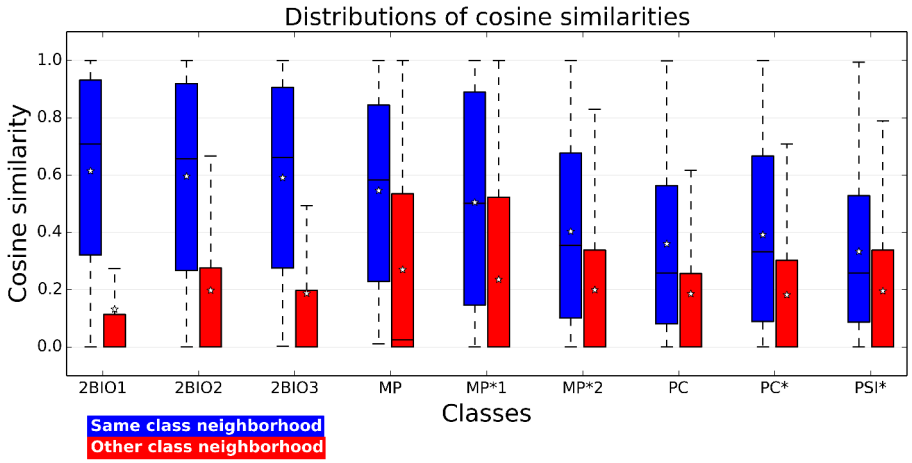


Figure 2.15: Boxplots showing the distributions of cosine similarities of neighborhoods of nodes in pairs of daily contact networks, for each class and for the whole population, and restricting the neighborhoods to intra-class (blue) and inter-class (red) neighborhoods. The center of each box gives the median (value given above each box) and its extremities correspond to the 25% and 75% quartiles. The star symbols show the mean value of each distribution.

model in which the topological structure of the contact network and the statistical properties of the cumulative contact durations are kept at the level of each class. This comparison helps us to conclude that the changes of students' neighborhoods from one day to another observed in the empirical network are substantial, but much less than in a situation in which contacts would occur at random. Moreover, not only the topological structure of the networks is important but also the attribution of weights in the network. This emphasizes the need to take this robustness of contact patterns into account in models of contacts between individuals, as done e.g., in [8].

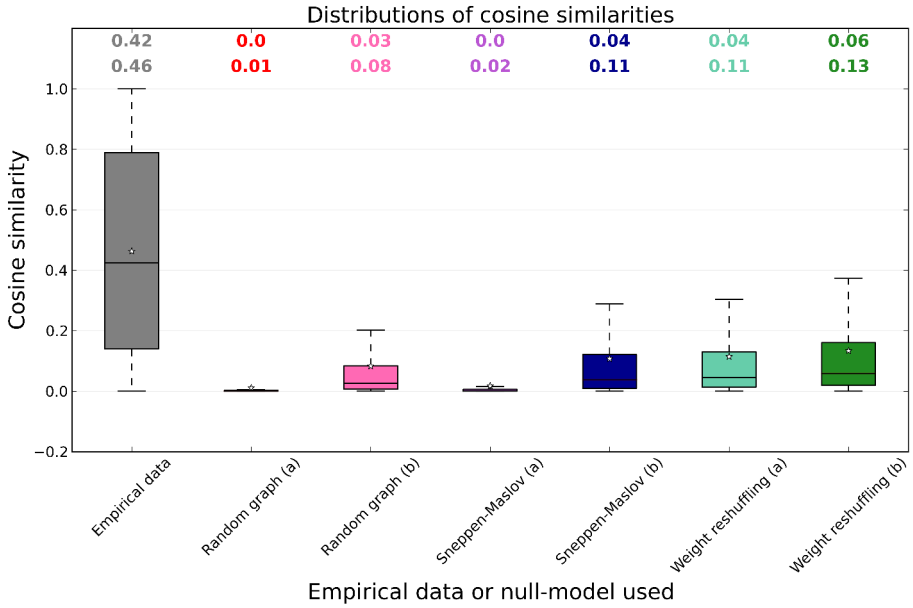


Figure 2.16: Boxplots showing the distributions of similarities of nodes' neighborhoods in different days (for all pairs of days in the data set), for the empirical data and for the various null models (1000 realizations for each). The center of each box gives the median (first value given above each box) and its extremities correspond to the 25% and 75% quartiles. The star symbols show the mean value (second value given above each box) of each distribution.

Name of the null-model	Description
Random graph (a)	All edges with their weight are replaced randomly in the graph.
Random graph (b)	Same as above although each edge between class X and class Y is replaced randomly remaining between class X and class Y.
Rewiring Sneppen-Maslov (a)	Choose 2 edges A-B and C-D such that A is not linked to D and B is not linked to C; Remove these edges replacing them by edges A-D (with weight of edge A-B) and C-B (with weight of edge C-D). Repeat this procedure approximately $3 * E$ (E: number of edges) times.
Rewiring Sneppen-Maslov (b)	Same as above although we do this separately for each pair of classes (and inside each class). Between two classes X and Y: A and C must be in class X and B and D must be in class Y. Inside a class X: A, B, C, D must be in class X. Repeat the procedure approximately $3 * E_{XY}$ (E_{XY}: number of edges between class X and class Y)/ $3 * E_{XX}$ (E_{XX}: number of edges inside class X) times. Note: when we cannot find enough nodes that meet the eligibility criteria in a specific cell, we do not do the rewiring.
Weight reshuffling (a)	The topology of the graph remains unchanged but the weights of the edges are reshuffled randomly.
Weight reshuffling (b)	Same as above although weight reshuffling is done for each pair of classes (or within each class) separately.

Table 2.9: Description of the null-models used.

2.7 Long-term stability of patterns

We take here advantage of the fact that data was collected in the same context in three different years and that each class participating to the study on a specific year is involved in the data collection of the following year to investigate the long term stability of the contact patterns between students in the high school. As students participating in the data collection in the three years were not the same, we cannot study the change in individual behaviors, but focus on the overall structure of the contact networks and matrices.

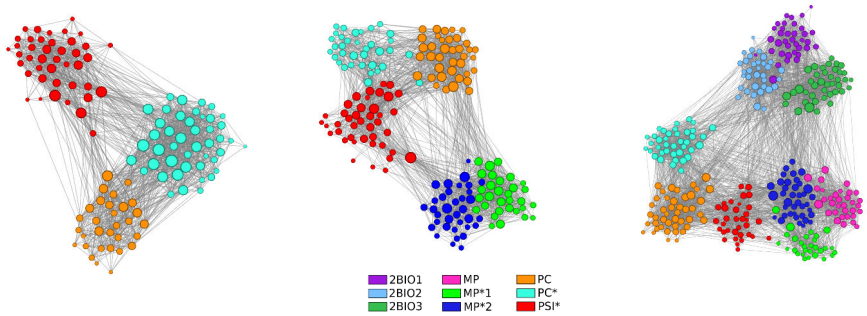


Figure 2.17: Representation of the three networks of 2011, 2012 and 2013 (from left to right). Each node represents a student, its color corresponds to the student's class and its size represents its degree. Created using the Gephi software: http://www.gephi.org.

Table 2.10 compares the main statistics of the aggregated contact networks of 2011, 2012 and 2013, both at the global level and for each class. Despite fluctuations in the absolute values, which can be expected as the data concerns different sets of individuals and different durations of study, similar properties are observed, with high values of the clustering coefficient and small average path lengths. Higher edge densities are also observed within each class, consistently with the strong class structure observed previously.

Figure 2.18 displays the distributions of nodes' degrees and link weights of the contact networks obtained in 2011, 2012 and 2013, aggregated over the whole data collection duration, as well as the distributions of contact durations. All distributions are very similar, with an exponential decrease at large degree values for the degree distributions, and very broad weights and contact duration distributions which collapse on top of each other for the data obtained in the different years.

Figure 2.19 moreover displays the contact matrices describing the structure and numbers of contacts between classes. We evaluate the similarity between the contact patterns of two specific years by computing the cosine similarity between the matrices of the two years. As the classes involved in the three studies are not entirely the same we compute the similarities between two years

Year of study	**2011**	**2012**	**2013**
N	126	180	327
E	1,710	2,220	5,818
Number of contacts	10,432	19,774	67,613
Cumulative duration of all contacts	561,010s (156 hrs)	900,940s (250 hrs)	3,770,160s (1,047 hrs)
Density	0.22	0.14	0.11
Density inside classes	0.57	0.51	0.66
Coefficient of clustering	0.58	0.48	0.50
Coefficient of clustering inside classes	0.71	0.65	0.77
Average shortest path length	1.95	2.15	2.15
Average shortest path length inside classes	1.44	1.53	1.31

Table 2.10: Comparison of the properties of the global aggregated networks of the 2011, 2012 and 2013 data collections.

on contact matrices restricted to classes the two data sets have in common. The results are given in Tables 2.11(a)-(c) for each type of matrix (contact durations, contact numbers, link densities) and for each pair of data sets (2011 with 2012, 2011 with 2013 and 2012 with 2013). We obtain very high values; even the smallest value is around 80%.

The overall contact structures are therefore very robust from one year to the next, despite the different populations involved. In order to investigate in more detail these similarities between years, we show in Figure 2.20 the temporal evolutions of the number of contacts registered in one-hour time windows in the three cases. The number of contacts vary strongly within each day, but the temporal patterns are very similar from one day to another in the three cases, with daily rhythms due to class and lunch breaks. In the three cases however, the contacts of each individual in different days are not completely the same, as shown for 2013 in Figure 2.12 and through the measure of the cosine similarities of neighborhoods in different daily aggregated networks (Figure 2.15). Interestingly, the distributions and average values of these cosine similarities take similar values in both years: for instance, the average values of the cosine similarities of individual neighborhoods in different days vary from 0.35 to 0.43 in the 2011 data set, from 0.29 to 0.44 in the 2012 data set and from 0.35 to 0.57 in 2013. The rates of renewal of the contact neighborhood of each individual is thus as well a robust property of these data sets.

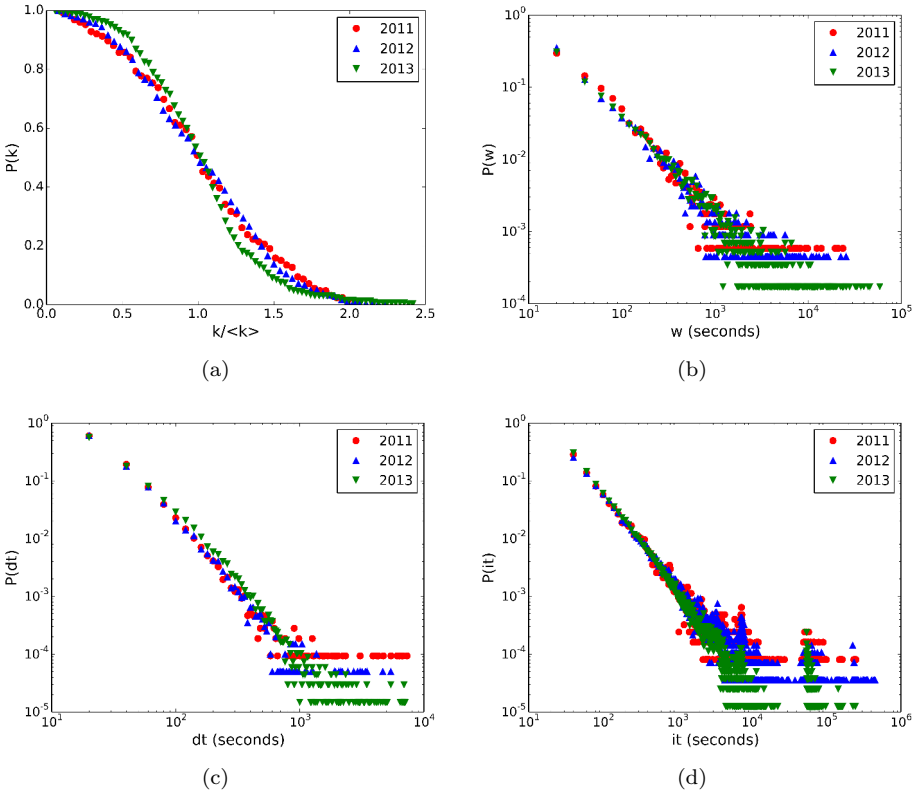


Figure 2.18: Properties of aggregated contact networks of 2011, 2012 and 2013. (a) CCDF of nodes' degrees. (b) Distribution of edge weights. (c) Distribution of contact durations. (d) Distribution of inter-contact durations.

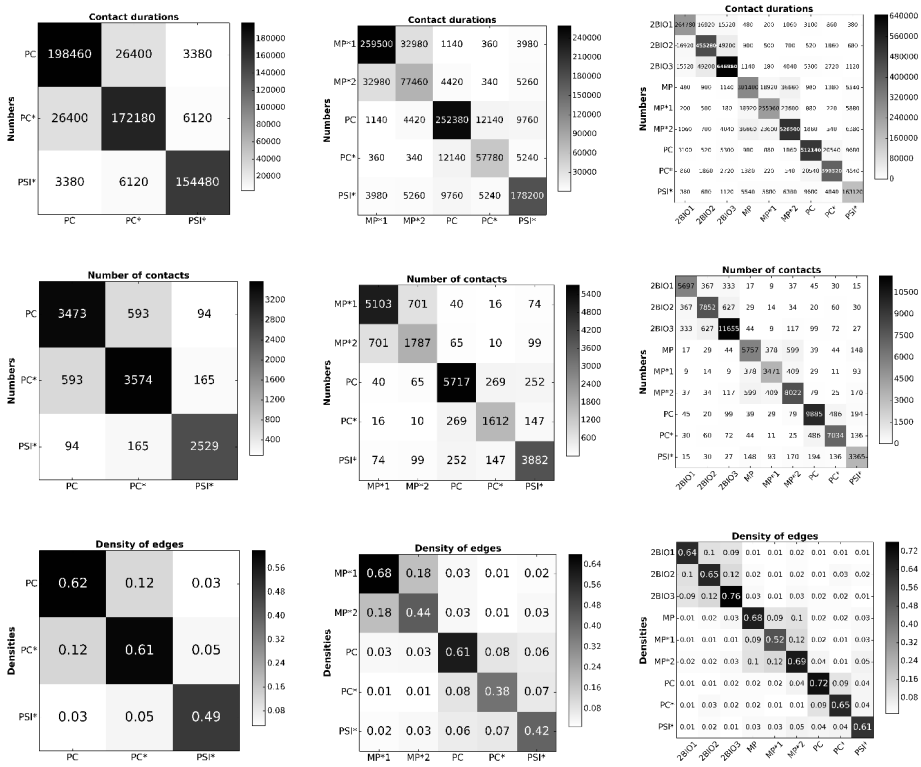


Figure 2.19: Contact matrices giving the cumulated durations of contacts (1st row), the number of contacts (2nd row) and densities of edges between classes (3rd row) for 2011 (1st column), 2012 (2nd column) and 2013 (3rd column).

(a)

Cosine similarities for contact durations matrices			
	2011	2012	2013
2011	1	0.91	0.96
2012	0.91	1	0.78
2013	0.96	0.78	1

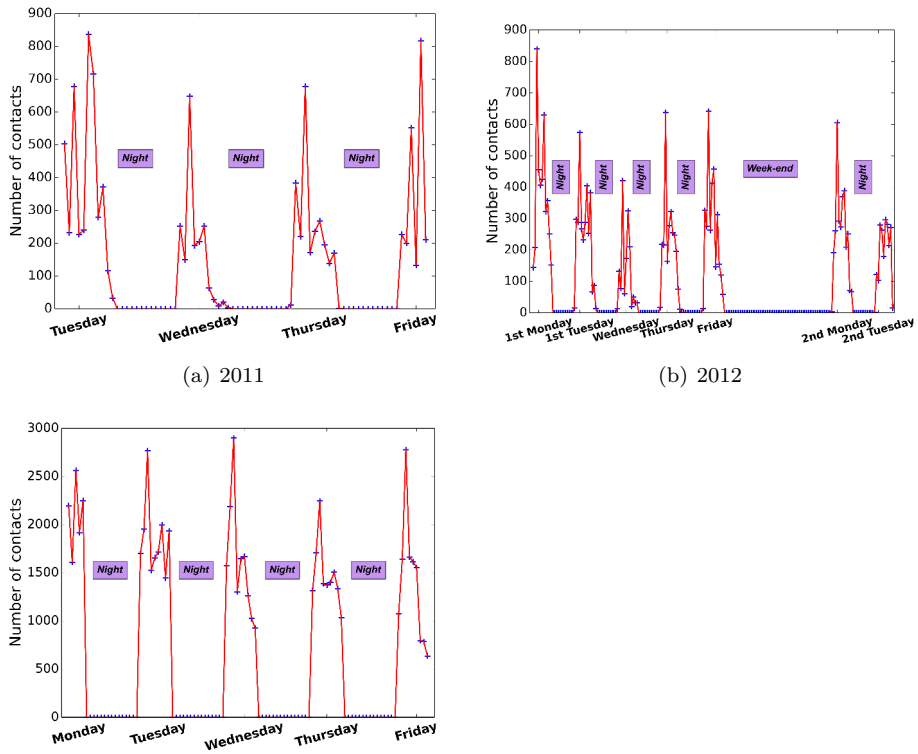
(b)

Cosine similarities for contact numbers matrices			
	2011	2012	2013
2011	1	0.89	0.96
2012	0.89	1	0.82
2013	0.96	0.82	1

(c)

Cosine similarities for link densities matrices			
	2011	2012	2013
2011	1	0.98	0.99
2012	0.98	1	0.96
2013	0.99	0.96	1

Table 2.11: Cosine similarities between contact matrices of different years for (a) contact durations matrices, (b) contact numbers matrices and (c) link densities matrices (Figure 2.19).



(a) 2011

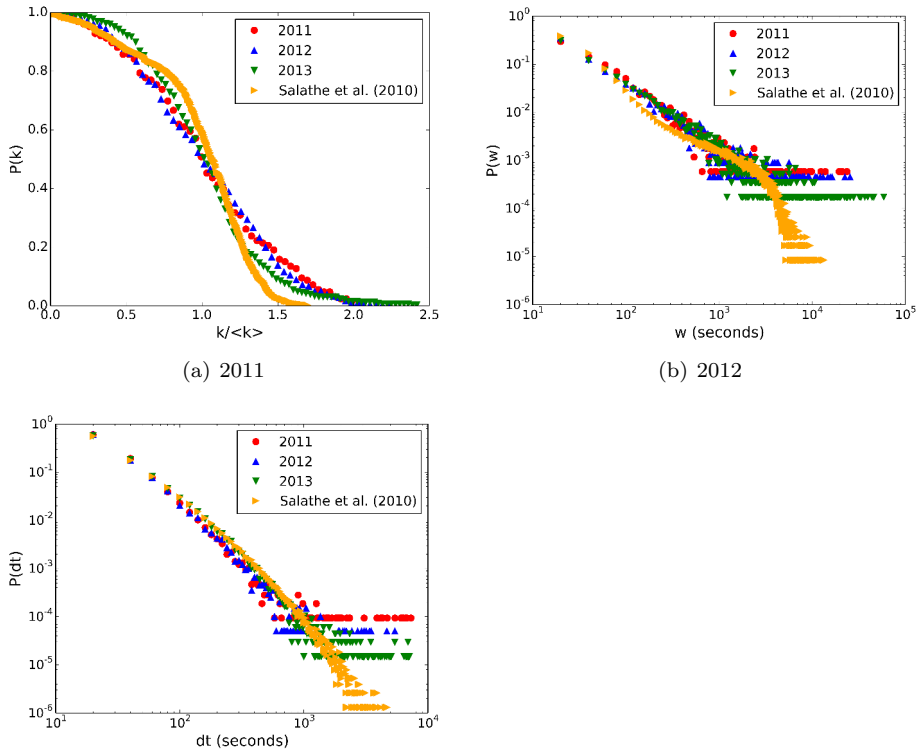
(b) 2012

(c) 2013

Figure 2.20: Evolution of the number of contacts per one-hour periods the (a) 2011, (b) 2012 and (c) 2013 data sets.

2.8 Comparison with another similar study

Comparison of data collected in different environments of similar nature is also important, in particular to highlight similar patterns in specific structures and therefore to inform mathematical models. Few studies using wearable sensors, and giving access to high-resolution data on contacts between individuals, are however available. Here, we compare our data sets (Thiers11,12,13) with the data set made public by Salathé et al. [26], which gives the durations of close proximity events between 788 individuals (mostly high school students, but also teachers, high school staff and others) during a typical day in an american high school. This contact network has 118,291 links, a coefficient of clustering equal to 0.5 and a small average shortest path length (1.62). In Figure 2.21, we compare some statistical properties of the contact networks. The definition of a contact is slightly different in the data set of Salathé et al. as the sensors detect a contact within a distance of 3 meters. As a consequence, the average degree is much higher (299.5) than in our data sets (between 24 and 36 for the three different years). Once rescaled however, the distributions of degrees of the aggregated networks are similarly short-tailed, albeit with slightly different functional shapes. Most importantly, the distributions of the contact durations and of the edge weights in the contact networks are very similar in the four studies, with similar slopes and heavy-tails. Moreover, the average duration of a contact is similar in the four studies: between 45 seconds and 1 minute. Finally, contrary to our data sets, no specific structure (e.g., classes, functions: students or teachers...) can be defined from the analysis of the contact network.



(a) 2011

(b) 2012

(c) 2013

Figure 2.21: Comparison of the distributions of (a) degree in the aggregated network, (b) cumulated contact durations, (c) durations of contact events, for the data sets analyzed above and the data set of [26])

2.9 Conclusion

In this chapter, we have presented the analysis of three high-resolution face-to-face contact data sets collected in a high school three years in a row using wearable sensors. We have focused some parts of our analysis on the third one, which was collected in December 2013, to avoid repetitions.

Many results can be drawn from this analysis. The network of contacts aggregated over the whole study has a small diameter and a high coefficient of clustering, which is common in many human interactions networks. Moreover, the network is highly structured by classes, similarly to what was observed in a primary school [10]. The distribution of degrees is narrow i.e., the number of distinct individuals with whom a student has been in contact does not show strong fluctuations. On the contrary, the distributions of contact durations and edges' weight are heavy-tailed with strong variations i.e., very short contacts occur as well as very long ones. As a consequence, no particular timescale can be defined. This is in agreement with results obtained in other environments [9–12].

The contact matrices highlight the structure in classes of the network with much larger values on the diagonal (which corresponds to the contacts within each class). Moreover, an additional substructure has been found gathering classes in 3 distinct groups of 3 classes each: Biology classes (2BIO1, 2BIO2, 2BIO3), Mathematics and Physics classes (MP, MP*1, MP*2), and Physics, Chemistry and Engineering classes (PC,PC*,PSI*). While the large values of the number of contacts inside each class is expected due to the school structure and schedule, the off-diagonal structure reflects patterns which are more due to either spatial arrangements of classes inside the high school or to similarities in the dominant subjects studied by the students. Given this structure in classes and groups of classes, it is clear that a homogeneous mixing of nodes would not be a good representation of the network. That being said, we have seen that taking into account an additional division of nodes in subgroups determined by the gender does not bring additional information as no strong gender homophily was observed in our data set, contrary to the case of a primary school [33]. The division of the population at the level of classes appears as an adequate level of description, for instance when designing a model of contacts to evaluate the outcome of a spreading process in this population [38].

In the context of the design of data-driven realistic models for human contacts, or for the information of models of spreading processes, the robustness of contact patterns at different timescales represents also a crucial information. For instance, the number of contacts fluctuates strongly over the course of the day while the evolution of the number of contacts is very similar from one day to another with, for instance, peaks of activity determined by lunch and class breaks. In addition, other properties of the network are extremely robust between the different days of the study: the statistical distributions of contact characteristics and nodes degrees of different days overlap, and contact matrices have very high similarity from one day to the other. This robustness of patterns is observed at a daily scale, as well as at a yearly scale with the comparison of these properties between the studies performed on different years, namely 2011,

2012 and 2013, with different students.

Interestingly, at the individual level, the contact patterns are not the same in different days, as we see from the measure of the average cosine similarity of students' neighborhoods in daily aggregated networks. However, the values of cosine similarities obtained for the empirical contact network are much larger than the ones obtained with null models where edges are reshuffled randomly, and even when the null model only reshuffles the weights of edges but keeps the topology of the network unchanged, showing that the specific position of weights in the network is somehow correlated to the topology of the network.

To our knowledge, few studies have performed similar analyses on contact patterns between high school students. In [26], Salathé et al. have collected data about close proximity events between 788 students of an American high school during one specific day. Given the different definitions of contact in the two studies, the average number of distinct neighbors in the aggregated contact networks are different: 35.6 for the present study and 299.5 for [26]. As a consequence, the distributions of degrees of the two different studies are not comparable, however, once rescaled by the average degree, these distributions are similarly short-tailed, albeit with slightly different functional shapes. Most importantly, the distributions of the contact durations and of the edge weights in the contact networks are very similar in the two studies, with similar slopes and heavy-tails.

The use of wearable sensors appears as an appealing method to measure contacts between individuals as it allows us to have time-resolved contact data without some biases of other traditional methods such as surveys or time-use data [15, 17, 19, 20, 24]. However, all these methods have both advantages and limitations.

In the next chapter, we compare the contact data collected through the use of wearable sensors with data of different nature collected at the same period within the same students : face-to-face contacts reported in contact diaries, friendship relations collected with a survey and Facebook ties. First, we compare the contact networks obtained from wearable sensors and contact diaries. Then we compare the sensor network with the friendship network and with the Facebook network separately. Finally, we study the full network of students' relationships as a multiplex network with 3 layers : face-to-face contacts, friendships and Facebook ties. These comparisons aim to quantify the biases of the different methods of data collection with respect to the use of wearable sensors and to have to have an idea of what amount of information can be drawn from incomplete data sets, especially in the context of simulations of epidemic spreading.

Chapter 3

Comparison of methods of data collection

In the previous chapter, we have analyzed face-to-face contact data obtained through the use of wearable sensors in a population of high school students population.

Many studies have provided similar analyses in various contexts [7–12]. This type of technology used to collect high-resolution data avoids the biases due to self-reporting. However, these methods do not allow us to access contacts with individuals not participating to the data collection but only in a closed population. Moreover, sampling issues can also arise if the participation rate is low in the population of interest [43].

These technologies are however recent and deployments are not always feasible. Many datasets have been obtained with other methods such as contact diaries. In contact diaries, participants are asked to report the individuals with whom they have been in contact and can give access to other characteristics of their contacts. In particular, participants may be asked to specify for each contact an estimated duration, if it involved physical contact or not and distinguish periods of well-being and illness of the respondent. Moreover, surveys can also be used to report other types of relationships than direct contacts, such as friendships or work collaboration. This type of data can be helpful when studying spreading processes such as spreading of infectious diseases: in particular, it is often assumed that friendships yield actual encounters.

However, this type of procedures is often costly and it can be difficult to recruit participants. Finally the main limitation are the biases inherent to self-reporting procedures. These biases are difficult to estimate and come from various reasons. First, the participants might not recall all their contacts (especially very short ones) or make incorrect estimates of their durations, especially in retrospective collection of data [4]; then, the individual perceptions and feelings or the apparent ambiguities of questionnaires can affect participants' answers [5]. Thus, the comparison between human relationships data coming from different

methods in the same population is of great interest to investigate the importance of theses biases. To our knowledge, very few studies have been performed such analyses on combined data sets [13, 31].

The complete data set collected in 2013 in Lycée Thiers gathers not only face-to-face contact data, analyzed in the previous chapter, but also contact data obtained with contact diaries, friendship relations obtained through surveys and Facebook relationships. In this chapter, we perform a comparison between the networks of these various types of interactions. First, we investigate similarities and differences between face-to-face contact data collected through the use of wearable sensors and contact diaries. Then, we study the multiplex network obtained by the superposition of differents types of students' relationships: face-to-face interactions (obtained with wearable sensors), friendships and online friendships (Facebook friendships). Finally, we investigate if the information contained in self-reported data is sufficient to obtain an accurate estimation of the epidemic risk computed with the contact network of sensors by performing simulations of epidemic spreading on the corresponding networks. This might help to quantify the biases and give hints on how to compensate for them.

This chapter covers the results reported in the following paper: *Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys*, published in PLoS ONE in September 2015 [14], and investigates more deeply some aspects of the analysis.

3.1 Data analysis

3.1.1 Contact diaries

In the previous chapter, we have analyzed data collected through wearable sensors; the deployment of the SocioPatterns sensing platform lasted for 5 days. On the fourth day of the study, students were asked to report the list of individuals with whom they have been in contact that particular day and to give an estimated aggregated duration of those contacts. As a consequence, the resulting contact diaries do not yield temporally resolved data, contrary to SocioPatterns data. Actually, we can build an aggregated weighted network where w_{ij}^{diary} is the aggregated duration of contact between i and j as reported by i. We note that links are not necessarily reciprocated (it can happen that i reports a contact with j but j did not report a contact with i) and even if both i and j reported a contact between them, the reported durations might not match: w_{ij}^{diary} can be different from w_{ji}^{diary}. We will first perform a systematic study of this data set as was done in [22].

As a second step, we will compare information contained in contact diaries with the contact network (obtained with wearable sensors) aggregated over the fourth day of study in the same spirit as [13]. To this aim, we will build a symmetrized version of the network of reported contacts, in which a link exists between i and j if at least one of the two reported a contact and the weight of this link is the maximum of w_{ij}^{diary} and w_{ji}^{diary}. Finally, for such symmetrized

network we can build, as for the sensor network, the contact matrices of the numbers of edges and of contact durations.

3.1.2 Friendships

As the contact diaries, the friendship survey yields a directed network: a node i can declare a friendship with node j without being reported as a friend by j. Actually, students were asked to give the names of their friends without specifying what is considered as a friendship relation. Moreover, the friendship network is unweigthed as the survey did not ask to quantify the intensity of a reported friendship. Note that we restrict the friendship network to participants to the survey who filled in correctly this survey (e.g., participants who declared friendship with all students of a given class were removed). We can also symmetrize the friendship network in order to compare it with the aggregated contact network, and obtain a link density contact matrix between classes.

3.1.3 Facebook

The analysis of the data gathered from the local Facebook network reveals a more complex character. Actually, the local egocentric network obtained for each student participating gives his/her friends and the relations of friendship between his/her friends. As very few students (17) provide such data, the presence or absence of a Facebook link for many pairs of students remains unknown yielding a very incomplete picture of the network. Figure 3.1 explains this point. We take the example of two students A and B giving access to their local Facebook network. We know if there is a link between A and any other students, idem for B. We know if there is a link between A's friends (idem for B's friends): for example, we know there is a link between i and j and we know there is no link between j and k. On the contrary, we are not aware of relationships between individuals who are only friends with A and individuals who are only friends with B: we do not know if there is a link or not between i and k. Thus, in our case, Facebook data cannot be represented as a network but consists in a list of "known-pairs" (pairs of students for which we know if they are Facebook friends or not) and a list of "unknown-pairs" (pairs of students for which we do not know if there is a Facebook link or not).

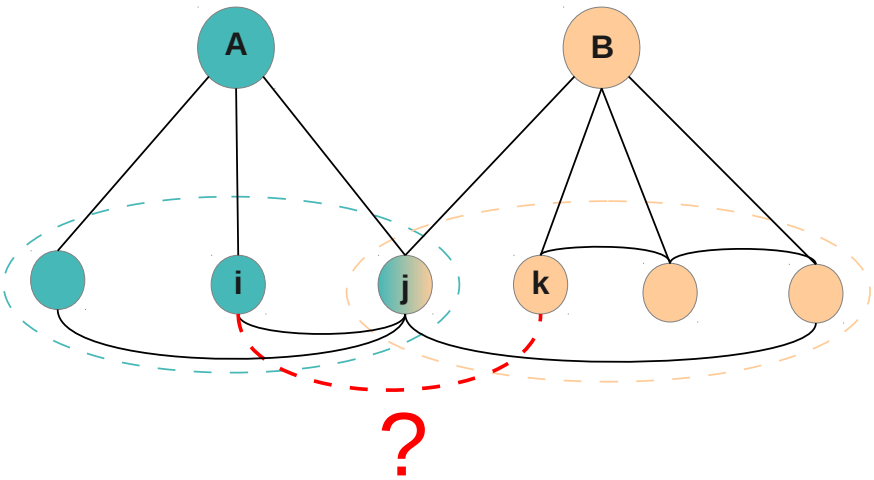


Figure 3.1: Facebook ego-networks. The local Facebook friendship networks provided by students A and B are shown in black. In particular, we know that i and j are friends on Facebook but not j and k, as i and j are both friends of A and j and k are both friends of B. On the other hand, we do not know if i and k are friends or not: the red dashed line represents the lack of knowledge about the potential existence of this relationship.

3.2 Contact diaries

3.2.1 Analysis of the contact diaries network

The network obtained from contact diaries is a directed and weighted network, it has 120 nodes (corresponding to the 120 students who filled in a contact diary) and 502 directed weighted links. We ignore contacts reported by respondents with non-respondents i.e., students who did not participate to the memory survey. In a second step, we keep only the students for whom the sensors registered at least one contact on the fourth day of study (Figure 3.2).

The resulting contact diaries network has 109 nodes and 416 weighted and directed links (158 links correspond to contacts reported by only one student i.e., non-reciprocated links, the other 258 links correspond to 129 pairs of students who both reported a contact with each other). Table 3.1 reports the corresponding statistics. Among the 129 contacts reported by both students, 81 (63%) were reported with the same estimated duration. In 35 cases, the reports of the two involved students differed by only one category. Among the 72 reciprocated links reported in the highest category of duration by at least one student, 71% were also reported in this category by the other student. These results are in agreement with the ones obtained in [22]. Moreover, following the work of [22], we compute the probability P to report a contact of a certain duration, under the hypothesis that such probability depends only on the duration. If N_c is the real number of contacts i.e., the number of pairs of students who have been in contact, and N_{both} is the number of pairs of students reporting both the contact, then $N_{both} = N_c P^2$, while the number of contacts reported by only one student is $N_{one} = 2N_c P(1-P)$; as a result, P is given by $N_{both}/(N_{both} + N_{one}/2)$. We obtain that the overall reporting probability is $P \approx 62\%$. Assuming that the correct duration of a reported contact is the highest reported value, we obtain that the probability to report a contact is 40% for contacts of less than 5 min, 54% for contacts between 5 and 15 min, 61% for contacts between 15 and 60 min, and 72% for contacts with aggregate duration longer than one hour.

3.2.2 Comparing contact diaries and sensors data

In this section, we compare the data collected by the wearable sensors with the contacts reported by the students using contact diaries. To this end, we consider on the one hand the aggregated weighted network of the contacts registered by sensors the 4th day of the study and on the other hand the symmetrized network of contact diaries. In this version of the network, a link exists between two students if at least one of the two students reported a contact between the two; if both students reported the contact the highest value of the aggregated contact duration reported by the students is retained.

The network of contact diaries has much less nodes than the network obtained from sensors as many students did not fill in the diaries. Moreover, the respondents were not uniformly distributed in classes: 20 were in 2BIO1, 11 in 2BIO2, 13 in 2BIO3, 23 in MP, 1 in MP*1, 18 in MP*2, 23 in PC, none in PC*

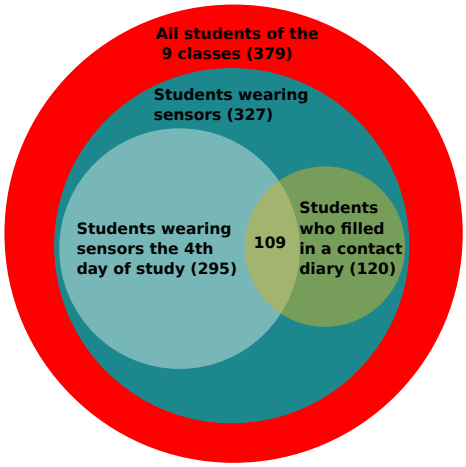


Figure 3.2: Venn diagram reporting the number of students who filled in a contact diary and the ones who were detected by sensors the 4th day of study as well as the number of students present in both networks.

Lower value \ Higher value	<5 min	5-15 min	15-60 min	>1 hour	**Row Tot**
Not reported	**38** (24%) (75%)	**31** (20%) (63%)	**33** (21%) (56%)	**56** (35%) (44%)	**158** (100%) (55%)
<5 min	**13** (42%) (25%)	**10** (32%) (21%)	**5** (16%) (9%)	**3** (10%) (2%)	**31** (100%) (11%)
5-15 min		**8** (32%) (16%)	**12** (48%) (20%)	**5** (20%) (4%)	**25** (100%) (9%)
15-60 min			**9** (41%) (15%)	**13** (59%) (10%)	**22** (100%) (8%)
>1 hour				**51** (100%) (40%)	**51** (100%) (17%)
Column Tot	**51** (18%) (100%)	**49** (17%) (100%)	**59** (20.5%) (100%)	**128** (44.5%) (100%)	**287** (100%) (100%)

Table 3.1: Cross-tabulation of pairs of contacts reported by students in the contact diaries. Each pair of participants with at least one contact reported gives a single observation. For instance, there were 12 pairs of students (i, j) such that i reported contacts with j with total duration between 15 minutes and 1 hour while j reported a duration between 5 and 15 minutes. Each percentage within a cell represents the percentage with respect to the row (right of the cell entry) and column (below the cell entry) totals.

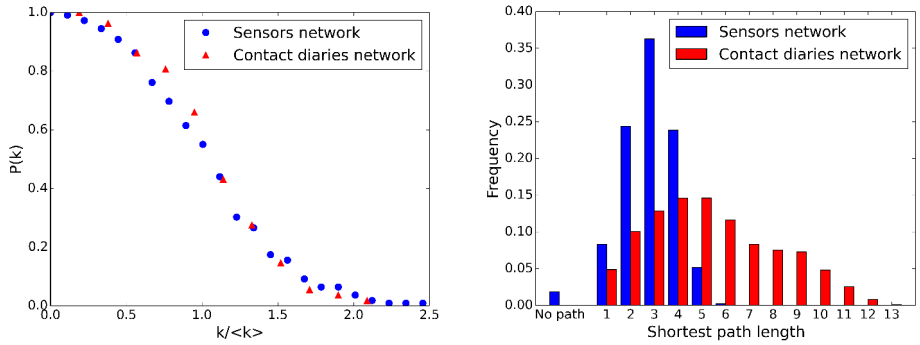
	Sensors	Contact diaries	Sensors	Contact diaries
Number of nodes	295	120	109	109
Number of edges	2162	348	488	287
Density	0.05	0.05	0.08	0.05
Average degree	14.7 (0.30)	5.8 (0.17)	9.0 (0.23)	5.3 (0.17)
Average clustering	0.38 (0.22)	0.45 (0.31)	0.45 (0.22)	0.44 (0.34)
Average SPL	2.81 (0.08)	5.36 (0.26)	2.94* (0.12)	5.36 (0.25)
Clique number	9	5	8	5

Table 3.2: Comparison of properties for the contact networks obtained from sensors and diaries. On the day of collection of the contact diaries (4th day of the study), only 295 students out of the 327 participating were present in sensor data. All network properties for the contact diaries network are computed on its symmetrized version (undirected network). In this summary table we assume that if a contact is reported by at least one of the two nodes, it exists. The right side of the table is performed after matching the population of the two networks. Matching is done by removing the nodes who did not participate to the survey and the ones who did not have contacts recorded by sensors on the 4th day of the study. *After the match, the graph is no longer connected; that means a path cannot always be found between two nodes. In this case, we computed the average on the connected pairs only. Standard deviations are given in parentheses.

nor PSI*, and this has consequences on the overall structure of the network as some classes are not or almost not represented. We therefore perform a short comparison of the sensor contact data of respondents and non-respondents: we have compared the properties (numbers, durations and aggregate durations of contacts, degree and centrality in the aggregated contact network) of respondents and non-respondents using Wilcoxon tests and did not find any significant difference.

Table 3.2 reports some properties of networks of sensors and contact diaries. On the one hand, we consider the entire networks obtained with both methods i.e., for sensor data, the network with all the nodes for whom at least one contact has been registered on the 4th day; for contact diaries, the network with all the students who have filled in a contact diary (first two columns). On the other hand, we consider the two networks restricted to the nodes present in both networks i.e., we take the subgraph induced by these 109 matching nodes on the two original networks (last two columns). The density and average degree of the contact network obtained by the sensors are almost twice as large as the ones obtained using contact diaries, but the degree distributions have similar shapes (Figure 3.3(a)). The cliques are also larger in the sensor contact network, while the average shortest path length is smaller (Figure 3.3(a)): nodes seem farther apart in the contact diary network than in the sensor data network. The average clustering coefficient is high and very similar from one network to the other.

Figure 3.4 shows the matrices of edge densities of the network obtained from



(a) (b)

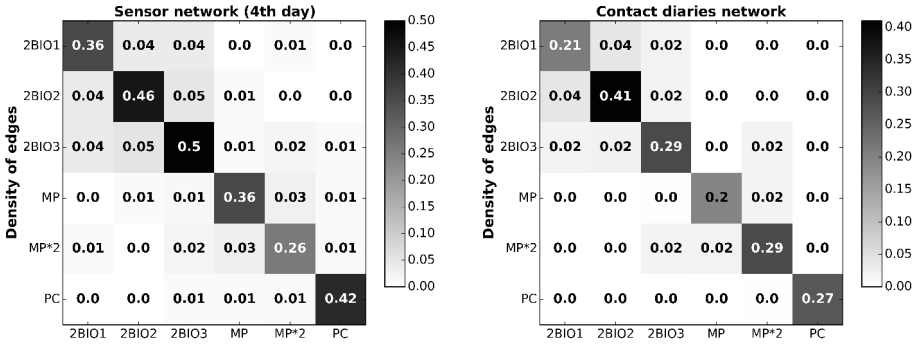
Figure 3.3: Comparison of the contact networks obtained by sensors and by contact diaries. (a) CCDF of nodes' degrees. (b) Shortest path length distributions. "No path" corresponds to disconnected pairs of nodes.

sensor data and the network obtained from contact diaries. The strong structure of the sensors network in classes is well preserved in the contact diaries network. Actually, the matrices are marked by the dominance of the diagonal elements together with the existence of groups of classes. Despite the low sampling of the contact diaries, a sensible information on the mixing patterns between classes is obtained. The similarity between the link density contact matrices is very high (97%).

Given the different numbers of links in the two networks, we expect discrepancies between sensor data and contact diaries. Overall, 70.4% of the links obtained from contact diaries correspond to contacts registered by the sensors, while only 41.4% of the contacts registered by the sensors find a match in the contact diary. We now investigate these discrepancies in more details.

Figure 3.5(a) compares the distributions of the cumulative durations of the links registered by the sensors, distinguishing between the links which were reported in the contact diaries and those which were not. For reference, the figure also reports the distribution of durations for all the links registered by the sensors. Both distributions are broad and heavy-tailed; most links have short cumulated durations, but large values are also observed in both cases. However, the distribution of durations for the links finding a match in the contact diaries is much broader, with much larger average duration and standard deviation (Wilcoxon tests for each pair of distributions reject the null hypothesis of equality of the distributions). In particular, links not reported tend to correspond to smaller durations, and all the links with a duration above a certain threshold (close to 1 hour) were reported in the diaries.

We moreover investigate in Figure 3.5(b) the diversity of the cumulative durations registered by the sensors for the links reported in the diaries in each duration category. Strikingly, all distributions are rather broad and, given a



(a) (b)

Figure 3.4: Contact matrices of link densities from (a) the network of contacts obtained using the sensor data collected on the 4th day and (b) the network of contacts as reported in the contact diaries. In order to compare easily the two matrices we discarded here the data corresponding to the MP*1, PC* and PSI* classes as too few students from these classes filled in a contact diary (1 for MP*1, 0 for PC* and PSI*).

reported category, both much shorter and much longer durations are registered by the sensors. In particular, the distributions corresponding to the two first categories (less than 5 min and between 5 and 15 min) are similar. However, the distributions become consistently broader for categories corresponding to larger durations, and links with durations (as registered by the sensors) above a certain threshold are reported only in the highest duration category of the diaries.

Table 3.3 gives more details through a cross-tabulation of the aggregate durations of the contacts as registered by the sensors or reported in the diaries. If we consider the duration registered by the sensors as accurate, we obtain that 68% of the contacts in the first category (less than 5 min) were not reported, against 30% and 31% for the next categories, while all the contacts lasting more than 1 hour were reported. Above all, 59% of the contacts detected by sensors were not reported in the contact diaries. On the other hand, the number of reported contacts which were not registered by the sensors does not strongly depend on the reported duration.

202 links are common to both networks, yet there are discrepancies between the durations reported by students and the durations detected by sensors. In particular, 49 (24%) links were reported and detected in the same duration category, while 146 (72%) links have a reported duration overestimated with respect to the one found in sensor data, and only 7 (4%) were reported with a shorter duration than the detected duration (overall, the Kendall's τ computed for the list of links ranked according to the durations either registered or reported yields a rank-correlation of $\approx 26\%$). Note that, if we use a symmetrized version

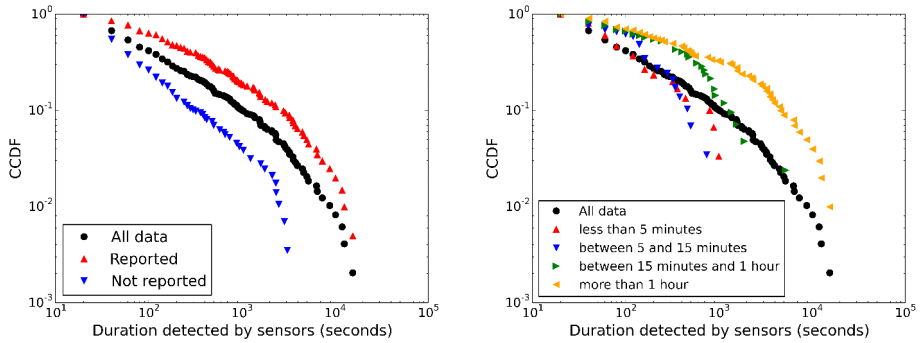


Figure 3.5: Sensors versus contact diaries. (a) CCDF of the aggregate durations of contacts registered by the sensors for (i) all 488 links between the 109 nodes belonging to both networks; (ii) the 202 links that were also reported in the diaries; (iii) the 286 links that were not reported in the diaries. (b) CCDF of aggregate durations of contacts registered by the sensors for the different categories of links reported in the diaries.

of the contact diaries network in which we retain the lowest value of the reported durations by each pair of individuals (including 0 for links non-reported by one of the individuals i.e., if the contact was not reported by both students, we remove the link) instead of the highest, the number of links found in both networks drops to 102, but the other results are robust. In particular, 31 links (30%) correspond to the same duration category in both networks, 64 (63%) to an overestimation in the diaries, and 7 (7%) to an underestimation.

We now turn to the comparison of the two datasets at the individual level. Contrary to the data presented in [13], we observe a significant correlation (0.4) between the degree of a node in the network obtained with sensor data and its degree in the symmetrized version of the contact diaries network, despite important fluctuations are present (Figure 3.6). On the other hand, when considering the directed network obtained with contact diaries, we do not find a significant correlation (0.14) between the degree of a node in the sensor network and its out-degree in the contact diaries network (number of contacts reported by the corresponding student). However, we note a significant correlation (0.39) between the degree of a node in the sensor network and its in-degree in the contact diaries network (number of students who declared a contact with this particular student). Similar correlation values are obtained when considering only contacts larger than a given threshold (Table 3.4). This interesting result indicates that we obtain a better picture of contacts of a student, at least in terms of number of contacts, by considering the contacts reported by other individuals with this one instead of the contacts actually reported by the student.

Even though the correlation between the out-degree of a node in the contact

Survey \ Sensors	Not detected	<5 min	5-15 min	15-60 min	>1 hour	**Row Tot**
Not reported	unknown (n/a) (n/a)	**257** (90%) (68%)	**17** (6%) (30%)	**12** (4%) (31%)	**0** (0%) (0%)	**286** (100%) (50%)
<5 min	**21** (41%) (25%)	**24** (47%) (6%)	**5** (10%) (9%)	**1** (2%) (3%)	**0** (0%) (0%)	**51** (100%) (9%)
5-15 min	**20** (41%) (23%)	**23** (47%) (6%)	**6** (12%) (11%)	**0** (0%) (0%)	**0** (0%) (0%)	**49** (100%) (9%)
15-60 min	**17** (29%) (20%)	**24** (40.5%) (6%)	**11** (18.5%) (20%)	**6** (10%) (15%)	**1** (2%) (7%)	**59** (100%) (10%)
>1 hour	**27** (21%) (32%)	**51** (40%) (14%)	**17** (13%) (30%)	**20** (16%) (51%)	**13** (10%) (93%)	**128** (100%) (17%)
Column Tot	**85** (15%) (100%)	**379** (66%) (100%)	**56** (10%) (100%)	**39** (7%) (100%)	**14** (2%) (100%)	**573** (100%) (100%)

Table 3.3: Cross-tabulation of the number of links detected by sensors and reported by students in each duration category. For instance, there were 23 links that were reported by students with duration between 5 and 15 minutes which were detected by sensors with an aggregated duration below 5 minutes. The percentages within a cell are computed with respect to the row (right of the cell entry) and column (below the cell entry) totals.

diaries network and its degree in the sensor network is not statistically significant, we investigate more in detail this relation with the idea that it could depend on the characteristics of contacts of each individual. The comparison of the two networks showed that students tend to remember more easily their longest contacts. Given this fact, we compute for each student the coefficient of variation (CV) of his/her longest contacts recorded by the sensors: for each student and each of his/her k neighbors we keep only the duration of the longest contact, then we compute the CV on this list of k durations. Then we separate the students into two groups: students with $CV_i \leq 1$ and students with $CV_i > 1$. $CV_i \leq 1$ means that the student i has contacts of similar durations with other students, while $CV_i > 1$ corresponds to a large variability, i.e., that i divides his/her contact time in a heterogeneous way among the other students s/he has met during the day. No particular grouping of individuals with $CV_i > 1$ was observed with respect to the various features (gender, class, field of study) of the students. We obtain 47 students in the first group and 62 in the second one. We find a significant correlation (close to 0.38) between the out-degree in the contact diaries network and the degree in the sensor network in the first group, while in the second group, no significant correlation was found. Note that the correlation for the in-degree is similar in both groups (close to 0.4). In other terms, for students who have encounters of similar maximum durations with other individuals ($CV \leq 1$), the contact diaries data reported by these students correlate with the data registered by the sensors. For students whose maximum contact durations are heterogeneous ($CV > 1$) on the other hand, no correlation in the diary-reported and sensor-measured degrees is observed. This is perfectly in line with our initial hypothesis: for $CV \leq 1$ the number of links remembered is then correlated with the real number of links in the contact network while, if $CV > 1$, there might be an arbitrary large number of links of small weight that

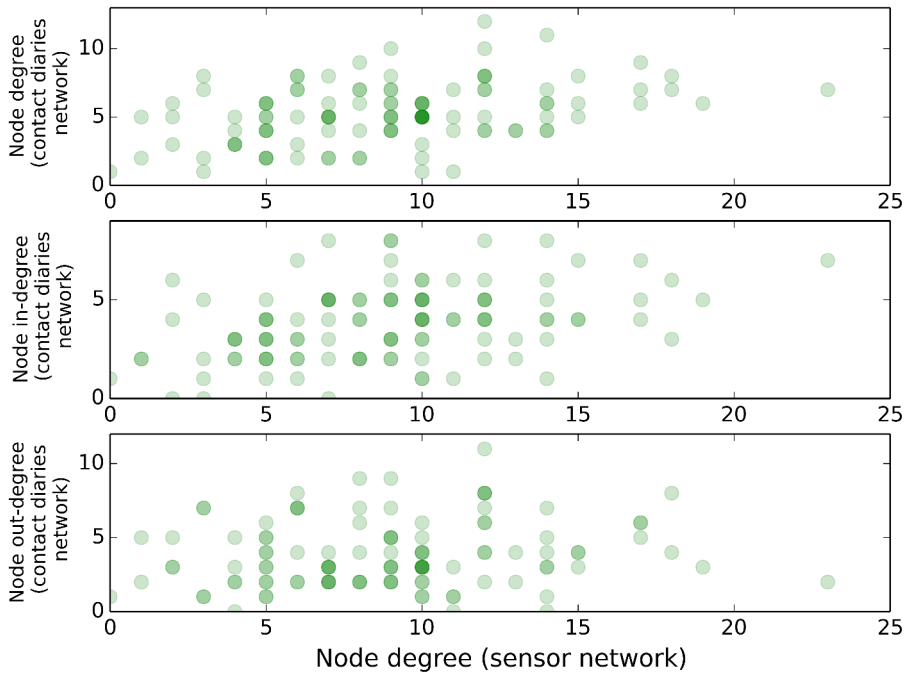


Figure 3.6: Comparison of the degree of individuals in the contact networks obtained by sensors and by contact diaries. Scatterplot of the number of links of each node in both networks. As the network built from contact diaries is directed, we consider the degree in its symmetrized version (top), the in-degree (middle) and the out-degree (bottom), vs. the degree in the contact network obtained from sensor data.

are not remembered in the diaries as the number of links that are reported is not correlated with the total number of links registered by the sensors.

Threshold	k_{in}	k_{out}
No threshold	0.39	0.14
40 seconds	0.44	0.22
60 seconds	0.50	0.22
80 seconds	0.49	0.16
100 seconds	0.48	0.16

Table 3.4: Coefficient of correlation between the degree of a node in the contact network built from the sensor data and the in- or out-degree of the same node in the network built from contact diaries. Each row corresponds to keeping only links with an aggregate duration above a certain threshold in the contact network built from sensor data.

3.3 Multiplex network of students' relationships

The previous sections concern only face-to-face proximity data which represents one type of relationship between individuals. However, other types of social ties exist; for instance, friendships and online relationships represent another aspect of human interactions. The gathering of these different types of links contribute to form a multiplex network in which each node represents a student and each pair of nodes can be linked by one to three links of different nature. These links might a priori be related: for instance, one has more physical encounters with a friend, or becomes friend with someone because of frequent encounters, etc, but might also differ substantially: one can be very good friend with someone and meet him/her only rarely because of specific constraints (such as different schedules of classes in our case) or can be friend with someone online only for communication facilities without being "true" friends. Therefore, we need to compare the different layers of this multiplex network to understand what information on actual contacts can be gathered from data about friendship relations. In particular, friendship survey data might be more reliable than contact diaries: it might be easier to remember the names of one's friends than the contacts occurred during a day and moreover, friendships evolve on slower timescales so that friendship surveys can more easily be gathered on several days without memory biases (however, other types of biases can arise because of personal perceptions i.e., considering someone as a friend is very subjective).

As described before, we have collected data about two types of friendships among students participating to the study in the high school: students gave the names of their friends through surveys and used a Facebook application to compute their local friendship network and gave us access to this network. However the participation to these two data collections was substantially lower than for the collection of face-to-face contacts data: 135 students correctly answered the friendship survey (41%) and only 17 students gave access to their local Facebook network. In the latter case, we end up with 156 nodes (48%). The number of students present in each data set is given in Figure 3.7. Figure 3.8 displays the network of contacts registered by the sensors during the week of

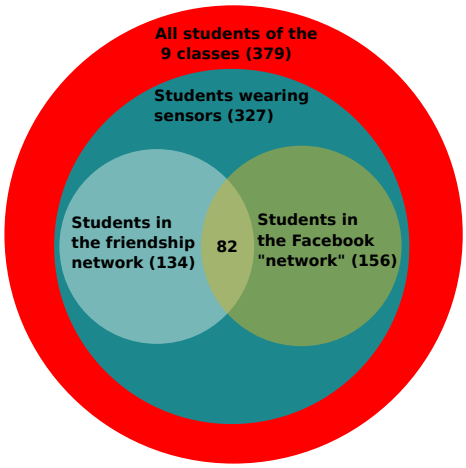


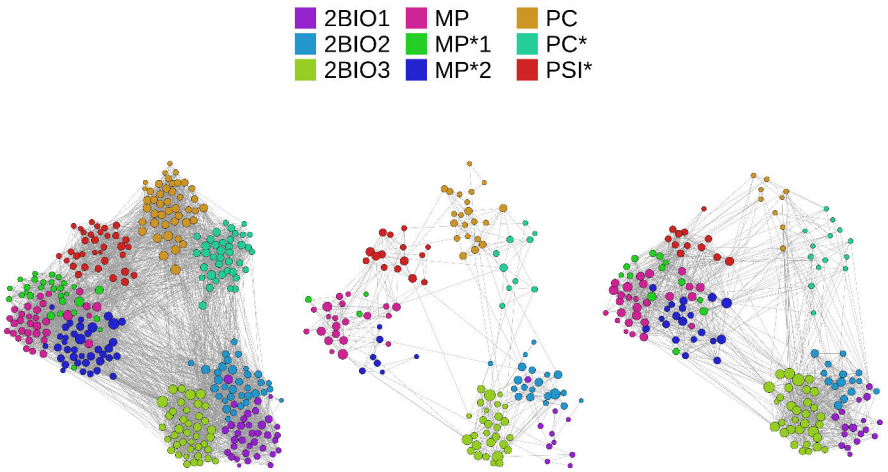
Figure 3.7: Venn diagram reporting the number of students present in each network. The student who did answer the friendship survey but who is not present in the sensor data was removed from the diagram.

data collection, as well as the network of reported friendships and the Facebook links, using the same position for the nodes in the three representations (as already explained, in the case of Facebook other links than the ones represented might exist: we are only representing the "known-pairs", so that Figure 3.8(c) might be an underestimation of the real number of existing Facebook links. For this reason, standard network metrics cannot be computed in this case). Overall, the friendship and Facebook networks have clearly much less nodes and appear substantially different, however, the grouping of nodes in classes is still relevant. We perform a more detailed comparison in the next paragraphs.

3.3.1 Contact network versus friendship-survey network

We build the friendship network in the same way as the network obtained from contact diaries. The friendship network is directed: a student A might report a friendship with another student B while B does not mention A as a friend, or A and B might both report each other as a friend. The directed network of friendships has 689 directed links of which 137 are not reciprocated and 552 corresponding to 276 pairs of students who both declared a friendship with each other. In the following, we will consider the symmetrized version of the network, in which a link is drawn between two students if at least one of the two reported a friendship with the other. The resulting network has 135 nodes and 413 links.

The students who filled in the friendship survey were not spread evenly in the various classes: 11 were in 2BIO1, 18 in 2BIO2, 28 in 2BIO3, 21 in PC, 3 in MP*1, 7 in MP*2, 21 in PC, 10 in PC*, 15 in PSI*. However, as in the case of the contact diaries, we have checked using Wilcoxon tests that respondents and non-respondents do not show significant differences in their contact characteristics



(a) Contact Network (b) Friendship Network (c) Facebook Relationships*

Figure 3.8: Contact and friendship networks. The three layers of the multiplex are shown using exactly the same layout: each node is placed at the same position in the three panels. The color of each node represents its class and size represents its degree in the corresponding network (here we consider a symmetrized version of the network of reported friendships). * Strictly speaking, the Facebook data do not provide a network as we do not have information about the presence or absence of a link between many pairs of nodes (see Figure 3.1). Figure created using the Gephi software http://www.gephi.org.

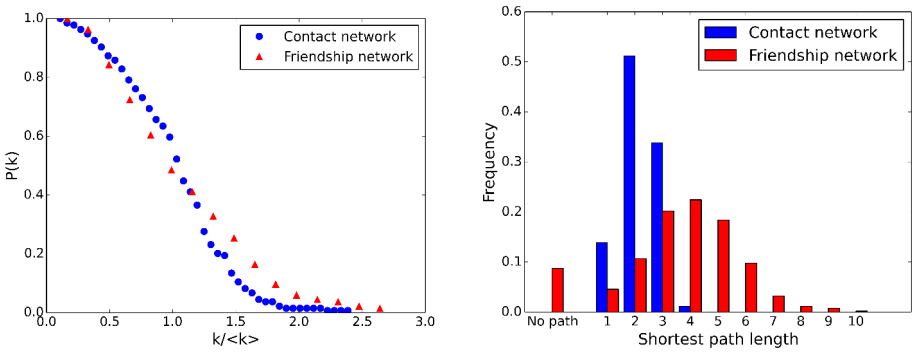
	Contact network	Reported friendships	Contact network	Reported friendships
Number of nodes	327	135	134	134
Number of edges	5818	413	1235	406
Density	0.11	0.05	0.14	0.05
Average degree	35.6 (0.14)	6.1 (0.32)	18.4 (0.18)	6.1 (0.32)
Average clustering	0.50 (0.08)	0.53 (0.29)	0.55 (0.11)	0.54 (0.29)
Average SPL	2.16 (0.08)	4.06* (0.16)	2.22 (0.10)	4.02* (0.16)
Clique number	23	8	14	8

Table 3.5: Comparison of properties for the contact network and the network of reported friendships. All network properties for the network of friendships are computed on its symmetrized version. The right side of the table is performed after matching the population of the two networks. Matching is done by removing the nodes who did not participate to the friendship survey and the one who was not present in sensor data. *The friendship network is disconnected. In this case, we computed the average on the connected pairs only. Standard deviations are given in parentheses.

(durations, aggregate durations, degree and centrality in the contact network measured by the sensors).

In Table 3.5 we compare the main features of the networks of reported friendships and of contacts. On the left side of the table, we consider the network of contacts aggregated over the whole data collection and the whole friendship network. On the right side of the table, we consider the networks restricted to the nodes which are present in both networks (only 1 student participated to the friendship survey but was never detected in contact with other students). As already observed in the case of contact diaries data, the network of friendships is much less dense than the network obtained with sensor data. This is quite expected as one naturally encounters many persons whom one would not list as friends in a survey. Moreover, the degree distribution is narrower than in the case of face-to-face contacts as shown in Figure 3.9(a), this is explained by the average degree $< k >$ which is much smaller in the friendship network, however, once rescaled by the average degree the two distributions have similar dicrease and similar shape; the nodes appear to be further away from each other in the friendship network as well. Actually, the average shortest path length is around 4 in the friendship network compared to 2 in the sensors network and the distribution of the shortest path lengths in the friendship network is also broader (Figure 3.9(b)), finally the maximal distance between two nodes is 4 in the contact network while it is 10 in the friendship network. The contact network has as well larger cliques than the network of friendships. Regarding the clustering coefficient, it displays large values in both networks.

Despite the very different densities of the two networks, the structure of contact matrices is well preserved in the friendship network as shown in Figure 3.10. Even though the density of links between each pair of classes differ between both cases, the diagonal of each matrix display much larger values and even the groups of classes (Biology, Mathematics and Physics) are highlighted in the friendship case: in fact the similarity between the two matrices is very high



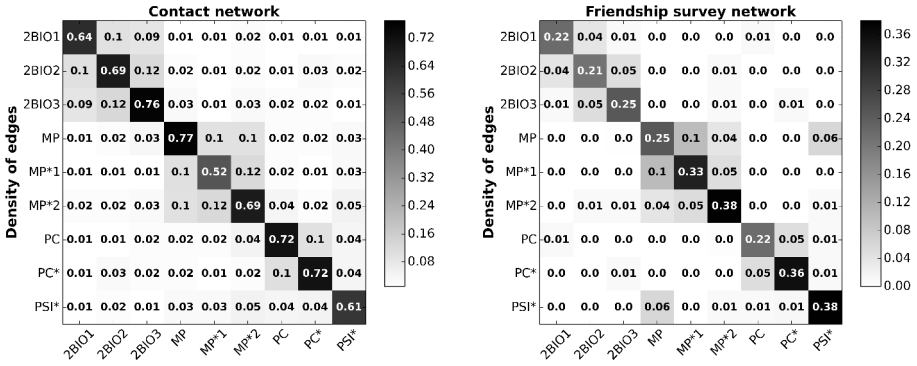
(a) (b)

Figure 3.9: Comparison of the contact network and the network of reported friendships. (a) CCDF of nodes' degrees. (b) Shortest path length distributions. "No path" corresponds to disconnected pairs of nodes.

and equal to 95%. Nevertheless, the major difference is that many values are 0 in Figure 3.10(b), which can be crucial when studying, for example, spreading processes within a population.

More in detail, 86% of declared friendships correspond to actual encounters in the contact network, while only 28% of contact links find a corresponding link in the friendship network. This significant discrepancy might be a priori explained by the fact that one tends to have actual encounters with his/her friends and, in addition one may have contacts with someone without considering him/her as a friend. Moreover, we may think that friendship links correspond to more frequent encounters or longer contacts. In fact, these numbers change if we restrict the contact data to stronger links, i.e., to contacts of larger aggregate duration: if we consider only links with an aggregate duration of more than 1 minute (respectively 3 minutes) we find that 75% (resp. 62%) of the declared friendship links have a corresponding link in the contact network, while 45% (resp. 58%) of the contact network links correspond to friendships.

In Figure 3.11, we show the cumulative distributions of aggregate durations registered by sensors for different types of links in the contact network: we distinguish 3 types (i) pairs of students for which no friendship is declared (887 links), (ii) pairs of students for which only one student reported a friendship (103 links), and (iii) pairs of students who both declared a friendship with each other (245 links). For reference, we also show the cumulative distribution of aggregate durations registered by sensors for all these three types of links (1235 links) between the 134 students common to both networks. The three distributions are broad with very short contacts and very long ones in each case: for instance, pairs of students which are not declared friends have some very long contacts and some students who have both declared a friendship with each other have very short contacts. However, the aggregate durations of contacts of declared friends



(a) (b)

Figure 3.10: Contact matrices of link densities from (a) the global aggregated contact network and (b) the symmetrized network of reported friendships.

Threshold	k_{in}	k_{out}
No threshold	0.42	0.34
40 seconds	0.51	0.44
60 seconds	0.51	0.45
80 seconds	0.53	0.47
100 seconds	0.52	0.47

Table 3.6: Coefficient of correlation between the degree of a node in the contact network and the in- or out-degree of the same node in the friendship network. Each row corresponds to keeping only links with an aggregate duration above a certain threshold in the contact network.

have a larger average and a broader distribution, especially if the friendship was reported by both. In particular, all links in the contact network with aggregate duration larger than a certain threshold (close to 2 hours and a half) correspond to a declared friendship and most contacts over this threshold correspond to a reciprocated friendship. Thus, even if a reported friendship link can correspond to effective contacts of very different durations, the global network of friendships includes the most important contacts in terms of durations.

At the individual level, we look at the correlation of degrees of the two networks. We observe a significant correlation (0.44) between the degree of a node in the contact network and its degree in the symmetrized friendship network (Figure 3.12). Contrary to the case of contact diaries and when considering the directed version of the friendship network, we observe significant correlations between the degree in the contact network and its out- and in-degrees, although slightly higher with the in-degree. As in the contact diaries network, similar correlation values are obtained when considering only contacts larger than a given threshold (Table 3.6).

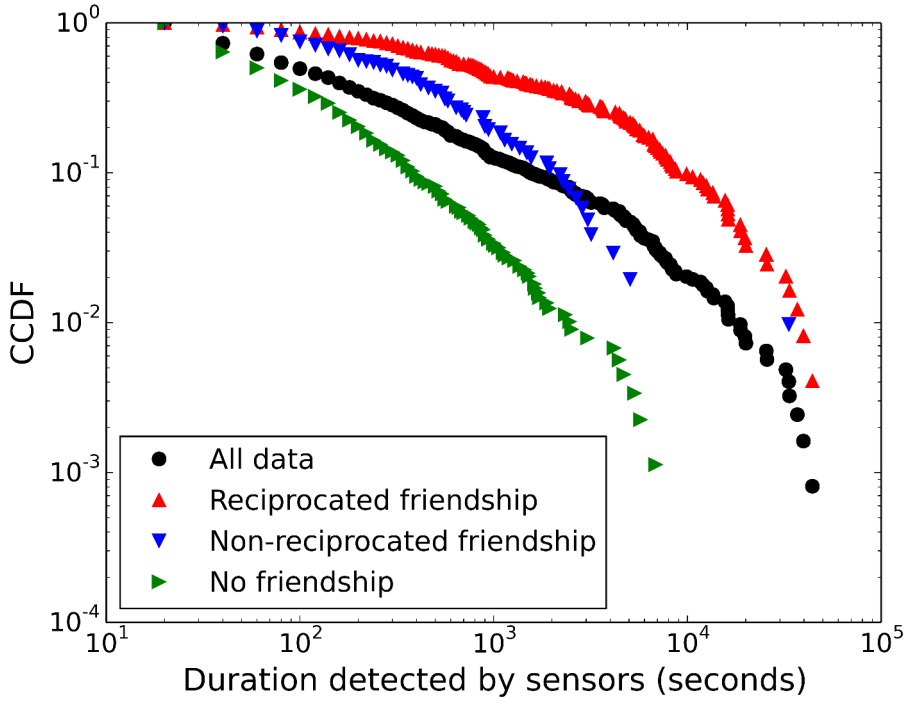


Figure 3.11: Contact network versus friendship network. CCDF of the aggregate durations for different kinds of links in the contact network: (i) all contact links between the 134 nodes belonging to both networks; (ii) links for which both students reported the friendship; (iii) links for which only one node reported a friendship with the other, and (iv) links for which no friendship was reported.

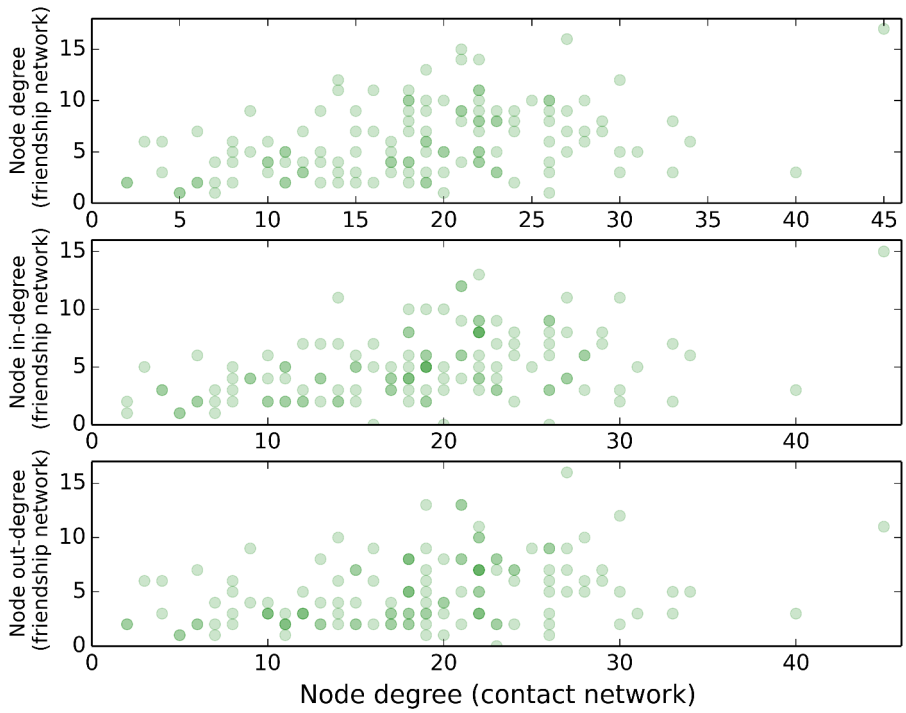


Figure 3.12: Comparison of the degree of individuals in the contact network and the friendship network. Scatterplot of the number of links of each node in both networks. As the friendship network is directed, we consider the degree in its symmetrized version (top), the in-degree (middle) and the out-degree (bottom), vs. the degree in the contact network.

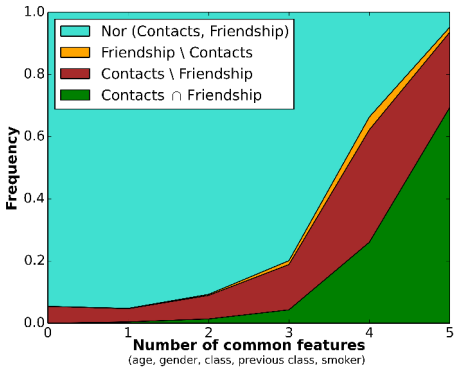
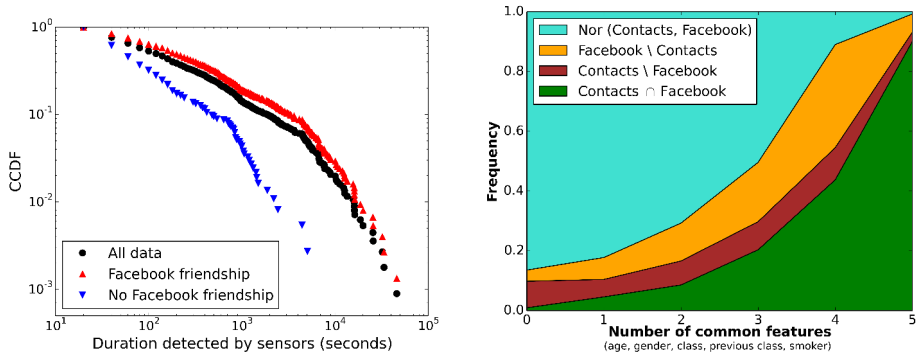


Figure 3.13: Fraction of friendship and contact links as a function of the number of features shared by two students. For instance: among the 1866 pairs of students that share 3 common features, 1491 (80%) have neither contact link nor friendship link, 23 (1%) have a link only in the friendship network, 272 (15%) have a link only in the contact network, 80 (4%) have a link in both networks.

We recall that we have collected five characteristics (metadata) for each student: the gender, the class of the student, his/her class in the previous year, if s/he was repeating the year (called "age") and if s/he was a smoker or not. We want to know if the sharing of a certain number of metadata by two students may inform us on the presence of contact or friendship links between the two. In Figure 3.13, we show the fraction of the 4 types of relations that can exist between two students: (i) contact and friendship links, (ii) only a contact link, (iii) only a friendship link or (iv) neither contact nor friendship link, as a function of the number of shared features (each pair of students can share from 0 to 5 characteristics). If two students share less than 4 features, the largely most probable situation is that they did not have any contact and are not friends either. On the other hand, if they share 4 or 5 features, we find at least one link in 73% of cases. In particular, friendship relations are observed almost only between students sharing 4 or 5 features, especially if they did not have any contact. Contacts among non-declared friends on the other hand can also be found for pairs of students sharing few features, highlighting the more random character of such links.

3.3.2 Face-to-face contacts and Facebook links

In this section, we focus on the comparison of the contact network with Facebook data and perform a similar analysis as for the friendship data. As mentioned before, only 17 students gave us access to their local Facebook network; as a result and strictly speaking, we could not build a network of Facebook relationships but rather work with a list of pairs of students ("known-pairs") for which we know if they have a Facebook link or not. The corresponding data set in-



(a) (b)

Figure 3.14: Contact versus Facebook links. (a) CCDF of the aggregate durations in the contact network for different kinds of links: (i) all contact links for which we know if there is a Facebook link or not (known pairs); (ii) links for which there is a corresponding Facebook friendship; (iii) links for which there is no Facebook link. (b) Fraction of Facebook and contact links as a function of the number of features shared by two students.

cludes 4515 known-pairs involving 156 students, with 1437 Facebook links (and 3078 pairs of students for whom we know they are not friends on Facebook). Moreover, these 156 students have 1118 links in the aggregated contact network. 52% of the Facebook links find a corresponding link in the contact network, and 67% of the 1118 links of the contact network are between Facebook friends.

The 17 students who gave access to their Facebook data are 9 in 2BIO3, 7 in MP and 1 in MP*2. The class repartition for the resulting 156 students is: 17 in 2BIO1, 14 in 2BIO2, 32 in 2BIO3, 26 in MP, 13 in MP*1, 20 in MP*2, 9 in PC, 13 in PC* and 12 in PSI*. As in the other data sets, Wilcoxon tests do not show any significant difference in the distributions of the contact properties of these students with respect to the others.

We report in Figure 3.14(a) the distributions of aggregate durations registered by sensors of links for which we find a Facebook link (750 links) and links for which we know there is no Facebook link (368 links) i.e., we work only with the "known-pairs". Both distributions are broad, as in the case of reported friendships, but the links between students who are not friends on Facebook have a clearly narrower duration distribution, and links with aggregate duration larger than a certain threshold correspond all to contacts between Facebook friends. However if we compare the distribution of the aggregate durations of contacts between students who are Facebook friends with the one for the students with reciprocated reported friendships (shown in Figure 3.11), the first one is less broad (the average duration is smaller and the maximum is also smaller).

Figure 3.14(b) reports the fraction of Facebook and contact links as a func-

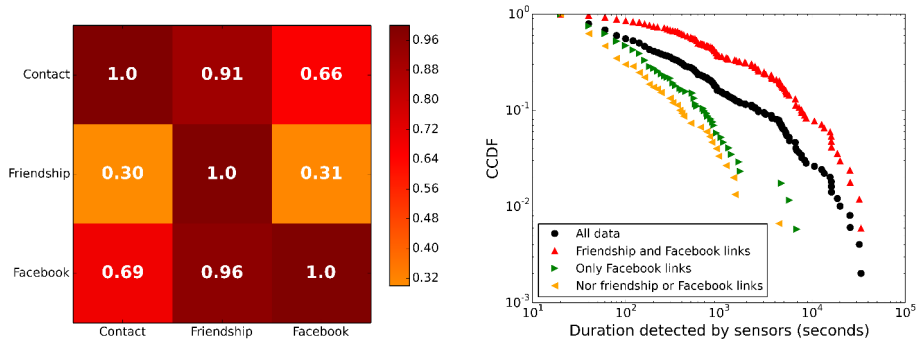
tion of the number of shared features by two students. Comparing this with Figure 3.13, we find that the number of common features has a strong influence on the fraction of such pairs having a link in the contact network or on Facebook, but smaller than in the case of reported friendships, as a non-negligible fraction of pairs of students with none or only one feature in common have a Facebook link. We also find Facebook links without corresponding link in the contact network between pairs of students with small number of shared features and even without any feature in common, contrary to what we found in the case of reported friendships. Finally, above 4 common features, almost all pairs (91%) of students have at least a link in one the two networks. This reveals the more "random" character of Facebook links compared to friendship links given the fact that the existence of a Facebook link gives no intuition on the existence of actual encounters.

3.3.3 Contacts and friendship networks as a multiplex

Instead of seeing them separately, we can combine friendship relations, online friendships and face-to-face contacts to provide a more complete picture of the relationships between students in the high school. This results in a multiplex network with three layers (strictly speaking, we have a multiplex set of nodes and links and not a network, as for the Facebook layer we do not have information about many pairs of nodes) where each pair of students has one, two, three or none of the three possible links. We here perform a simple analysis of this multiplex, in which we consider only students who are part of all three corresponding data sets and only known pairs: they represent 82 nodes, 496 links in the contact network (aggregated over the whole study), 199 reported friendship links and 513 Facebook links.

The conditional probability to find a link in one layer of the multiplex given its existence in another layer is shown as a heat map in Figure 3.15(a). In the case of a reported friendship link, the probabilities that we find a corresponding Facebook link or a contact link are very high (more than 90%). On the contrary, the probabilities that a friendship link was reported by a pair of students given the fact that they have a Facebook link or that they have been in contact are much lower (around 30%). On the other hand, the conditional probabilities between contact and Facebook links are quite similar.

We investigate more in detail the differences between reported friendships and Facebook links. Figure 3.15(b) shows the cumulative distributions of aggregate contact durations detected by the wearable sensors for different sets of pairs of nodes: (i) pairs of students for which no friendship is declared nor Facebook link (150 contact links), (ii) pairs of students for which there exists a Facebook link but no friendship is declared (168 links), (iii) pairs of students who have both a declared friendship and a Facebook link (172 links) and (iv) for reference we also show the case with all the different types of links (496 links). Note that we do not show the case for pairs of students who declared a friendship but do not have a Facebook link because there are only 6 links of this type. All the distributions are broad. However, the distribution for the pairs of



(a) (b)

Figure 3.15: Multiplex Analysis. (a) Conditional probability to find a link in one layer (row index) given its existence in another one (column index); (b) CCDF of the aggregate durations in the contact network for different sets of links. We did not compute the distribution for links for which there is a link in the friendship network but no Facebook link because there was only 6 links of this kind.

students who are both declared friends and friends on Facebook is much broader than the two others. Actually, the distributions of aggregate contact durations for pairs of students who are only friends on Facebook and for pairs who have neither friends on Facebook nor reported friends are very similar. Links of a duration above a certain threshold are observed only between reported friends. With respect to the aggregate duration of registered contacts, having a link on Facebook is therefore not at all equivalent to being reported friends: if it is only a Facebook link (but not a reported friendship), such a link tends to correspond to rather short face-to-face contacts; this reinforces the idea that Facebook links have a more random character with respect to both contacts and friendships. This emphasizes the need of an analysis taking into account the three layers and not only the online friendship one.

3.4 Epidemic risk from different methods of data collection

In the previous sections, we have found that the collection of human relationships data using surveys or contact diaries yields incomplete pictures of the contacts that actually happen between individuals and that are recorded by the use of wearable sensors, at least in our data set. On the one hand, the numbers of students who have filled in contact diaries and surveys were substantially lower than the number of students who accepted to wear the sensors: this corresponds to a sampling of the population of interest. This might be explained by the fact that students in these specific studies ("classes préparatoires") have

already a rather heavy workload without adding the burden of filling in contact diaries or surveys. On the other hand, most short contacts recorded by sensors were not reported in the contact diaries and did not correpond to declared friendships; as a consequence the networks obtained from these methods are much more dilute than the network obtained from sensors: this corresponds to a sampling of ties within the population. Thus, the distance between individuals in the networks obtained with self-reporting methods are overestimated, which could have strong consequences when using such data in data-driven models of dynamical processes (e.g., epidemic spreading). However, above a certain threshold of duration all detected contacts were reported in the diaries and reported as friendships. Moreover, the structural organization of the population in classes was preserved in the networks obtained from diaries and surveys. Thus, one could argue that the reported links carry enough information to feed data-driven models [43].

To start investigating this point, we perform numerical simulations of an SIR spreading model (defined in the first chapter) on the various networks, namely, the contact network built from sensors, the network built from contact diaries in its symmetrized version and the symmetrized friendship network. Specifically, we first compare the simulations performed on the contact diaries network with the aggregated network built with sensor data of the 4th day (i.e., the day we have collected the contact diaries). The network of sensors is weighted so the comparison should be done with results obtained with another weighted network. The network built from diaries is already weighted, however, each reported weight corresponds to a range of duration instead of a specific duration. Moreover, as said before, the reported durations are, on average, overestimated. Overall, there are thus many ways to build a weighted network. For the simulations, we therefore use three different types of weights assignment: (a) we use the reported weights in the following way: each category of duration (less than 5 minutes, between 5 and 15 minutes, between 15 minutes and 1 hour) is replaced by the average of the range, for the category "more than 1 hour" we take the average between one hour and the maximum aggregate duration registered by sensors on the fourth day, note that if the two students reported a contact in different categories of duration we take the highest reported duration, (b) for the reported contacts which have a corresponding link in the sensor network of the fourth day, we take the corresponding weights in this network (first step), for the others we take the weights at random in the list of weights already assigned in the first step, (c) we pick the weights at random from the distribution of weights of the sensor network of the 4th day.

On the other hand, the friendship network does not correspond to a specific day so the results obtained with this one could be compared to the sensor network aggregated over the whole week of study or over any particular day. Moreover, the friendship network is originally unweighted, thus for each comparison we assign weights to the friendship links in a consistent way: (i) for the comparison with the sensor network aggregated over the week of study the weights are picked at random from the list of weights of the aggregated sensor network; (ii) for the comparison with the sensor network of one specific day, the

weights are picked from the distribution of weights corresponding to the sensor network aggregated over this specific day.

Figure 3.16 displays the outcomes of simulations of epidemic spreading on the contact diaries and sensor networks. The simulations performed on the network of contact diaries yield different results depending on the method used to assign the weights. Actually, if the topology remains unchanged in the three cases, the distributions of weights vary from one to another. The epidemic risk is clearly underestimated for methods (b) and (c). Regarding the fraction of epidemics with size above a certain threshold (Figure 3.16(a)), the results obtained with contact diaries and method (a) are in good agreement with the ones obtained for the sensor network. Yet, when looking at the average values of epidemic sizes the use of the contact diaries network clearly underestimate the epidemic risk and the shapes of distributions of epidemic sizes are completely different (Figure 3.16(c) and (d)). We will investigate more closely the importance of the way we assign the weights on edges in the case of the friendship network in the next chapter.

The two different comparisons performed with the friendship network yield similar results (Figures 3.17 and 3.18): the simulations using the friendship network give a very strong underestimation of the epidemic risk with respect to the ones using the contact network built from sensor data.

Overall, both contact diaries and friendship networks underestimate the epidemic risk when used for simulations of epidemic spreading. These discrepancies are quite expected as three factors must be taken into account: the low participation of students in these two cases, the low density of the two networks and the way we assign the weights to the links. However, these comparisons should not be done as an end in itself; actually, the comparison is interesting as it may help us to quantify, understand the biases due to self-reporting data and compensate for these biases. This might give hints on how to use the information contained in incomplete data sets to inform models of epidemic spreading.

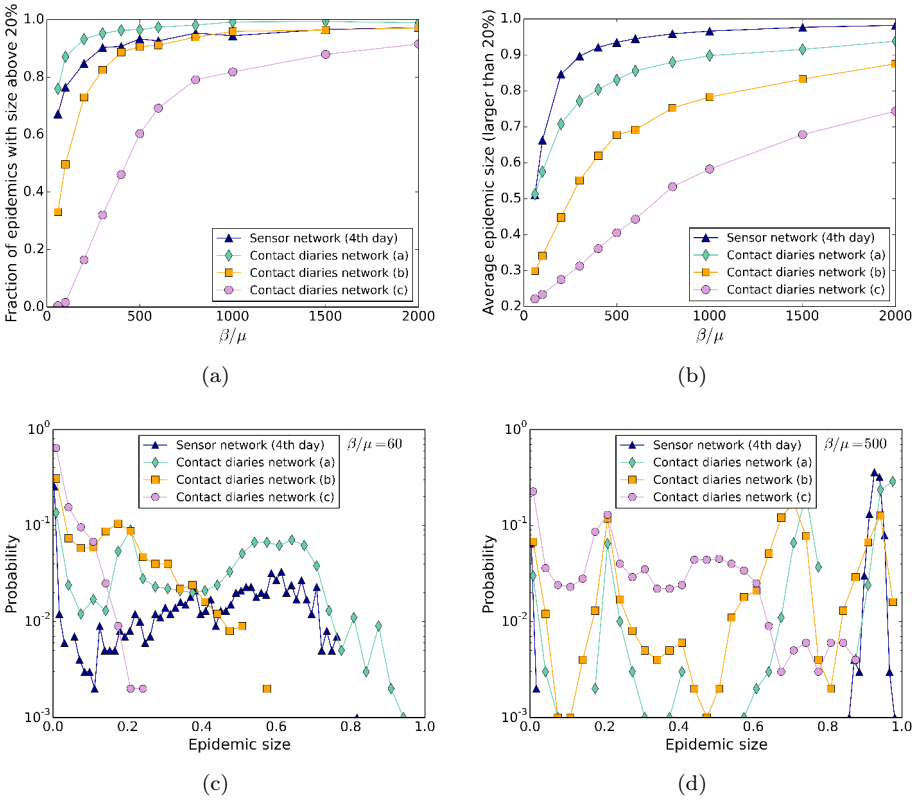


Figure 3.16: Outcome of SIR spreading simulations performed on sensor and contact diaries networks. (a) Fraction of epidemics with size above 20% (at least 20% of recovered individuals at the end of the SIR process) as a function of β/μ. (b) Average size of epidemic with size above 20% as a function of β/μ. (c) and (d) Distributions of epidemic sizes for two different values of β/μ.

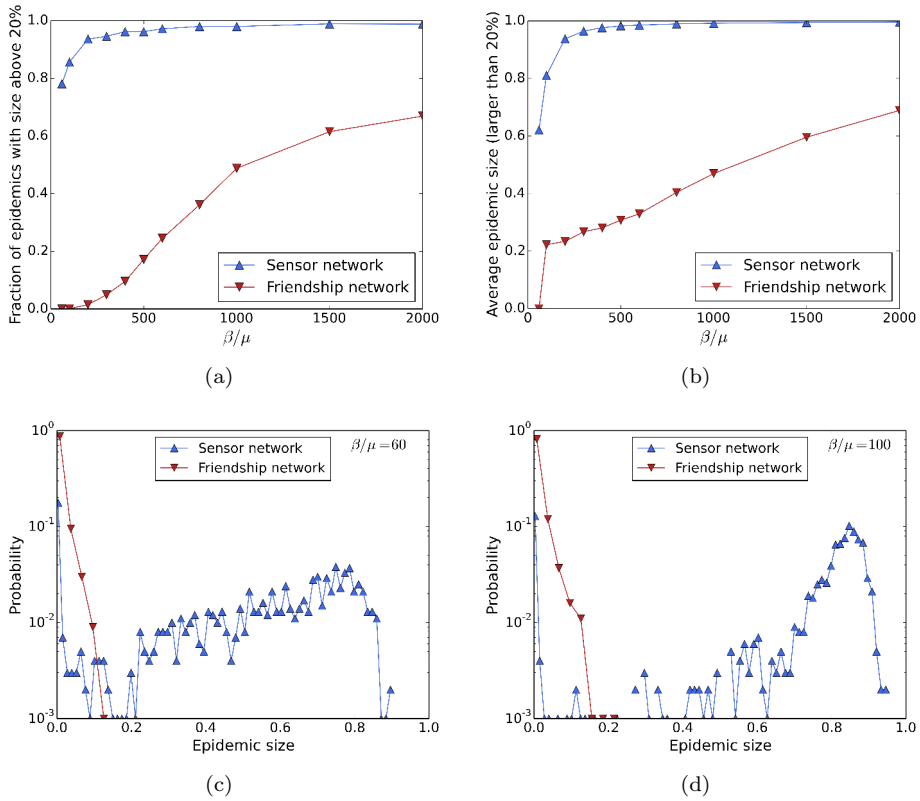
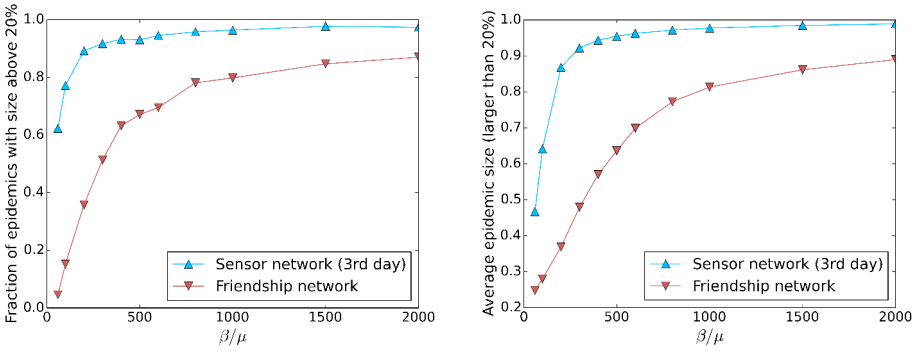


Figure 3.17: Outcome of SIR spreading simulations performed on sensor and friendship networks (first case: we use the sensor network aggregated over the whole study duration and weights of the friendship network are taken at random from the distribution of weights of the sensor network aggregated over the third day). (a) Fraction of epidemics with size above 20% (at least 20% of recovered individuals at the end of the SIR process) as a function of β/μ. (b) Average size of epidemic with size above 20% as a function of β/μ. (c) and (d) Distributions of epidemic sizes for two different values of β/μ.



(a) (b)

Figure 3.18: Outcome of SIR spreading simulations performed on sensor and friendship networks (second case: we use the sensor network aggregated over the third day and weights of the friendship network are taken at random from the distribution of weights of the sensor network aggregated over the whole study duration). (a) Fraction of epidemics with size above 20% (at least 20% of recovered individuals at the end of the SIR process) as a function of β/μ. (b) Average size of epidemic with size above 20% as a function of β/μ.

3.5 Conclusion

In this chapter, we have presented a comparison of several data sets concerning different types of interaction between the same individuals. The fact that these data sets were collected at the same time and in the same population allowed us to compare them and to quantify the overlap and complementarity of data sets of different nature. Collection of contact data and comparison of interactions of different nature are of interest for many purposes, including the information of data-driven models of relevance in epidemiology, as well as the investigation of human behaviour and social relations.

Even though the use of sensors to measure contact patterns has become more widespread in the last years, such deployments are not always feasible. Other methods, in particular based on contact diaries, have been and are still widely used. We have thus, in the same spirit as [13], compared the network of contacts reported in diaries by the students with the contact network measured by the sensors. We have confirmed the results of [13] and obtained some further insights:

1. most students who participated to the data collection by wearing sensors did not fill in the contact diary, probably due to the extra burden at the end of a school's day. However, no significant difference has been found between contact characteristics (in the sensor network) of respondents and non-respondents;

2. most short contacts detected by sensors were not reported in the contact diaries, whereas long contacts have a high probability to be reported and contacts of duration above a certain threshold (about 1 hour) were all reported;

3. among reported contacts, the contacts which were reported by both students were reported in the same category of duration most of the time, similarly to what was observed in [22];

4. the distribution of aggregate durations measured by the sensors were broad for contacts both reported and not reported in the diaries, the distribution was however broader for contacts reported;

5. the contact durations reported by students tended to overestimate the durations measured by sensors. This outcome is in agreement with results from social studies about self-reported diaries biases stating that individuals tend to perceive the time spent in some activities (talk, work, play) differently from the reality [44] and frequently to overestimate it [45];

6. despite the low sampling of individuals and the lower density, the overall structure of the sensor network was well preserved in the contact diaries network with similar contact matrices;

7. at the individual level, the degree of an individual in the sensor network is more correlated with his/her in-degree in the contact diaries network than his/her out-degree. This interesting result seems to indicate that a more accurate picture of the contacts of an individual is given by the other individuals reporting a contact with him/her.

Since this data set in prone to biases (the low participation of individuals (1), the lower density compared to the network obtained from sensor data (2) and the overestimation of contact durations(5)), its use in the context of the design of data-driven models describing human contacts must be done considering these biases. However we retrieve most of the long contacts (2) as well as some crucial characteristics of the network such as the structure in classes(6). Thus the data set of reported contacts might give enough information to feed data-driven models.

Friendship relations represent another type of data for which surveys are commonly used; the resulting data sets are a priori less prone to biases due to imperfect memory, and might also be easier to collect in a given population than contact diaries. Surveys might indeed be given and collected on a less constrained time frame than contact diaries, as friendships evolve on longer time scales than contacts. Similarly, online social networks might in some cases be easier to collect automatically. However, the precise relation between reported friendships or online friendships is not well known and need further investigation. In the case of reported friendships, the major findings were that (i) the friendship survey suffer from low participation rate as well as in the case of contact diaries, (ii) the longest contacts corresponded to reported friendships and

most friendship relations lead to actual face-to-face encounters, but (iii) many short contacts did not correspond to reported friendships, resulting in a friendship network with much lower density than the contact network. Nevertheless, the structure in classes of the contact network is well preserved as seen from the similar structures of the contact matrices. On the other hand, Facebook friendships yield a different picture than the reported friendship links. First, sampling issues prevent us from building a network, strictly speaking, of Facebook relationships. The probability that a contact is observed between two individuals, knowing that they are linked on Facebook, is also much smaller than the same probability conditioned on a reported friendship. In addition, the distribution of aggregate contact durations between individuals who are linked on Facebook but did not report a friendship to each other is much narrower than the one obtained for individuals who reported a friendship: overall, Facebook links seem to have a more "casual" character than friendship links, in agreement with the intuition that Facebook links are easier to establish than real friendships and give no intuition on the strength of friendships. Except the lack of information on contact durations, the network of reported friendships gives information of quality similar to contact diaries with respect to the sensor-measured contact network. And similarly to the case of contact diaries, feeding data-driven models with friendship data would lead to the same biases, especially on the estimation of the epidemic risk. Regarding the case of Facebook, this particular data set contains too little information to be used for the design of models describing human contacts.

Some limitations of this analysis are worth discussing. First, the participation rate to surveys and diaries was substantially lower than for the collection of face-to-face contacts data using wearable sensors. Similarly the networks obtained from contact diaries and friendship surveys are much less dense than the network obtained from sensors, yielding large discrepancies in the properties of the resulting networks. Moreover, little is known about how such samplings affect the characteristics contact diaries and friendship networks and should be investigated further. Actually, our study is limited to only one combined data set in one specific environment and few studies [13, 22, 31] compare data coming from different sources. It would be of great interest to gather more combined data sets in which relationships between the same individuals are collected with different methods.

In the next chapter, we start to explore more in detail the issue discussed above (Section 3.4), i.e., the comparison of simulations of spreading processes using sensor data and friendship survey data. We have seen in 3.4 that the use of friendship data leads to an underestimation of the epidemic risk. We will investigate whether this underestimation might be seen as biases due to a sampling process performed on the sensor network. Note that we will not try to check if the friendship network is precisely a sampling of the contact network, but if it is possible to reproduce the outcomes of simulations of epidemic spreading obtained with the friendship network by using a sampled network of the sensor network. An equivalence between survey and sampling procedures might indeed give hints on how to compensate for the biases due to the incompleteness

of contact data deduced from self-reporting procedures.

Chapter 4

Equivalence between friendship network and a non-uniform sampling of contact network

It is a well known fact that sampling procedures affect networks properties. Depending on the procedure of sampling and on the sample size, the networks can be affected in different ways, many works have studied the impact of sampling on networks' characteristics such as average degree, degree distribution, clustering or assortativity properties [46–52]. On the other hand, few studies have investigated how the outcome of simulations of dynamical processes in data-driven models is affected if incomplete or sampled data are used [43, 53–55]. As most data sets are in fact incomplete samples of the target network, researchers have moreover tackled the issue of inferring network statistics from incomplete data [56–59]. Since many networks are the support of dynamical processes, it is also crucial to develop methods to obtain estimates of the outcome of such processes in the case of incomplete or sampled data [43, 54].

In this perspective, understanding if the differences in outcomes between contact and friendship data may be seen as biases due to a sampling process might then give hints on how to compensate for such biases and how to use the information contained in the friendship network to obtain accurate prediction on the epidemic risk, even in the absence of data on the actual contact network, in the spirit of [43].

We have found in Section 3.4 that the use of the friendship network yields an underestimation of the epidemic risk with respect to the use of the face-to-face contacts network obtained from wearable sensors (we recall that similar results were obtained for contact diaries). In this chapter, we ask if using the reported friendship network is equivalent, in the context of simulations of epidemic

spreading, to a specific sampling of the contact network. To make progresses in this direction, we consider several sampling procedures: some are used as reference while others are attempts to mimic the friendship survey procedure. First, we use simple sampling methods, already described in other contexts, such as uniform random sampling of nodes or edges. Then we present a non-uniform sampling method we have designed: it favors sampling of the most important contacts of each sampled individual (links with large weight). The resulting networks are used for SIR simulations of epidemic spreading. We show that the outcomes of simulations performed on these sampled networks are equivalent to the one obtained when the friendship network is considered.

Then, we apply this specific method of sampling to other data sets describing face-to-face contacts and study how changing its parameters (number of nodes sampled, density of sampled network) impacts the outcome of spreading simulations. We also investigate how the choice of a method for assigning weights on edges of unweighted networks affects the results of SIR simulations performed on the resulting weighted networks.

This chapter covers the results reported in the following paper: *Epidemic risk from friendship network data: an equivalence with a non-uniform sampling of contact networks*, published in Scientific Reports in April 2016 [60].

4.1 Methodology

In this chapter, we consider the combined data set reporting face-to-face contacts and friendship relations between high school students and compare the outcomes obtained from numerical simulations of the SIR spreading model on the corresponding networks with the results of simulations performed on networks obtained by sampling the contact network using several sampling methods (described in the next paragraph). To quantify the epidemic risk and compare outcomes of these simulations, we measure the distributions of epidemic sizes (i.e., the final fraction of recovered nodes), the fraction of epidemics with size larger than 20% and the average size of these epidemics (the cut-off of 20% is chosen arbitrarily to distinguish between small and large epidemics; changing the value of this threshold does not alter our results).

The contact network, measured using wearable sensors, has $N = 327$ nodes and $E = 5818$ weighted edges. Each edge (i, j) with weight W_{ij} corresponds to the fact that individuals i and j have been in contact for a total time W_{ij} during the deployment, which lasted one week. The friendship network, obtained through a survey, has $N_F = 135$ nodes and $E_F = 413$ unweighted and undirected edges. All the nodes of the friendship network (but one) belong to the contact network.

It is important to note that we consider here a static version of the contact network, while the data of sensors provides temporal evolution of the network. The rationale behind this choice is twofold. First, the friendship survey data does not contain temporal information. If using friendship network can be seen as a sampling of contact networks, it is thus necessarily a static sample. Second,

when modeling the propagation of infectious diseases with realistic timescales of several days, it has been shown in [8] that a static weighted contact network contains enough information to obtain a good estimate of the process outcome. Clearly, when dealing with faster processes, the temporal evolution of the network becomes relevant; in that case, studies such as [43] have shown how to build realistic surrogate timelines of contacts on weighted networks, using the robustness of the distributions of the durations of single contact events and of the intervals between successive contacts measured in different contexts.

4.1.1 Sampling methods

Many different sampling procedures of network data have been considered in previous works, and their impact on the network's statistical properties have been studied [46–52]. Sampling of the network used as substrate for transmission events is also known to affect the result of simulations of epidemic spread [43, 53, 54]. In particular, population sampling has a strong impact, even in the case of uniform sampling [43]. We therefore consider various sampling procedures on the contact network: some methods are choosen for reference as they are standard methods, other methods are attempts to truly mimic the friendship survey procedure. As a first step, we consider methods of sampling which are tuned to obtain the same number of nodes as in the friendship network:

- We first consider as reference the Subgraph method ("SubFr"): we consider the 134 nodes of the Friendship network present in the contact network and take the subgraph induced by these nodes on the contact network. This would correspond to a population sampling of the contact network, with the sampled population corresponding to the respondents of the friendship survey, hence different from a random uniform choice.

- In the MSZ method, which is not strictly speaking a sampling, we consider a randomized version of the friendship network using the algorithm of rewiring described in [42]. In the case, the structures and correlations of the friendship network are destroyed by the reshuffling of edges.

- In the Random Node method ("RN"), we choose $N_F = 135$ nodes uniformly at random from the contact network and we take the subgraph induced by these nodes on the contact network. This corresponds to a population sampling with uniformly random choice of the sampled nodes.

- In the Degree-based Random Node method ("DRN"), we choose $N_F = 135$ nodes with probability proportional to their degree in the contact network and we take the subgraph induced by these nodes on the contact network. This corresponds to a non-uniform population sampling where the choice is oriented towards the most "important" nodes of the network.

- We also consider the Egocentric sampling method ("EGO"): we select a node at random and include this node and all its neighbors in the sample. We repeat this step until we reach the desired number of nodes, N_F. If

this number is exceeded by including the chosen node and its neighbors, we randomly choose a set of its neighbors so that the right number of nodes is exactly reached. Then, we take the subgraph induced by this sample of nodes on the original network. This would correspond, in the friendship survey, to a case where respondents report all of their links.

As these methods do not allow us to control the number of edges in the sampled network, we also consider several additional sampling methods, in which we can tune this number and set it equal to E_F.

- In the Random Edge method ("RE"), we first choose edges at random from the contact network until we reach the desired number of nodes N_F; as the number of chosen edges is still lower than E_F we then choose edges at random from the contact network with the condition that both their extremities are in the set of nodes obtained in the first step, until the desired number of edges is reached. This method is the simplest method than can be used to choose the desired numbers of nodes and edges.

- In the Weight-based Random Edge method ("WRE"), we applied a method of sampling similar to the RE method, except the edges are chosen proportionally to their weight in the contact network. We found in the previous chapter that the longest contacts correspond to friendships: in this method, we select edges in agreement with this idea.

- In the Refined Random Node method ("RNref"), we add the following step to the RN method: after the subgraph is obtained, we remove edges at random to get the desired number of edges in the final sampled network. This method is another very simple one in which we can control the numbers of nodes and edges.

- We propose a new Refined Egocentric method ("EGOref"), inspired by the result of [14] that the longest contacts corresponded to reported friendships, while many short contacts did not. Here, we select N nodes called *egos* at random from the contact network. For each *ego* i we select some of its edges as follows: each edge $i-j$ is selected with a probability equal to $\min\left(p * \frac{w_{ij}}{s_i}, 1\right)$, with w_{ij} the weight of the edge between i and j, $s_i = \sum_\ell w_{i\ell}$ the strength of the *ego* node i and p is the parameter of the model. We then keep only the *egos* and the selected edges linking them and we remove the other edges (between egos and non-egos) and nodes (non-egos). With this method, we end up with the desired number of nodes by setting $N = N_F$, and a number of edges that depends on the parameter p. Figure 4.1 summarizes this process. This method really tries to mimic the survey procedure: the egos correspond to the respondents to the survey, they report edges with egos and non-egos depending on their weight (the higher is the weight, the higher is the probability to report the weight as a friendship), then the non-egos i.e., the non-respondents are removed from the network yielding a friendship network among only respondents to the survey.

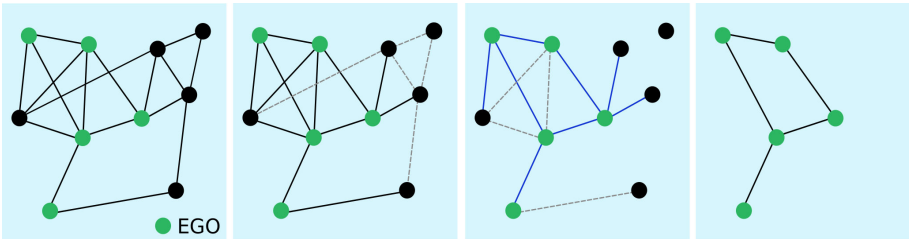


Figure 4.1: Sketch of the EGOref sampling process. We first select a certain number of nodes as egos (in green), who represent the respondents to the surveys. Links between non-respondents cannot be observed (dashed grey links in the second panel). Each ego then "chooses" to report some of its links, with probability depending on their weights (links shown in blue in the third panel, while the non-reported links are shown in dashed grey lines). We then finally keep only the egos and, among the chosen edges, only the ones joining egos (last panel).

- While the *egos* are chosen uniformly at random among the nodes of the contact network in the EGOref method, we also consider a heterogenous EGOref method ("EGOref-het"), in which the distribution of *egos* in the various high school classes corresponds to the one of the friendship network (*egos* still being chosen at random within each class).

4.1.2 How are simulations performed ?

As mentioned in the previous chapter, the friendship network is unweighted. However, for the purposes of comparison with the contact network, it is necessary to have a weighted network: actually we might start with comparing only the structure of networks using equal weights for all edges for the friendship network as well as for the sampled networks, but it is a well-known fact that weights and especially the heterogeneity of weights is important in the context of spreading phenomena. Thus, we assign to each edge (i, j) of the friendship network a weight w_{ij}: there are various ways to assign weights to edges; we have chosen three of them as they are the most intuitive ways. The first method is to choose weights at random from the distribution of weights of the contact network. In the second one, we recall that 86% of declared friendships correspond to links in the contact network, thus for these friendship ties we assign the weight found in the contact network and for the others 14% we choose weights at random from the distribution of the 86% already assigned weights. In the third one, we choose the weights as in the previous method and then we reshuffle randomly the weights among the links.

In the case of networks obtained from the sampling methods described above, we have as well multiple choices to assign the weights, for instance: we can keep the weight of sampled edges (as the sampled networks are subgraphs of the

contact network, all edges have a corresponding weight in the contact network) or we can choose the weights at random from the distribution of weights of the contact network.

The use of different methods of weight assignment will cause discrepancies in the outcome of simulations of epidemic spreading. First, the randomization of the assignment of weights might break some correlation between the strength and the degree of a node. Then, some of the sampling methods choose edges proportionally to their weights, as a consequence, the distribution of the empirical weights of the sampled network is not the same as the one obtained in the contact network. We recall that the purpose of this chapter is to understand if the friendship survey procedure is equivalent to a sampling procedure. Hence, the most suitable method for comparison is to assign weights chosen at random from the distribution of weights of the contact network in friendship and sampled networks, which is known to be a robust feature of human contact patterns [2, 34]. We come back more in detail to the differences caused by the choice of either of these methods further in this chapter.

4.2 Properties of sampled networks and outcome of SIR simulations

4.2.1 Simple sampling methods

As a first step, we study sampled networks obtained from simplest methods of sampling and compare the results obtained with these networks with the contact and friendship networks. Table 4.1 shows the characteristics of the empirical contact and friendship networks compared to the networks obtained by the simplest sampling techniques (SubFr, MSZ, RN, DRN, EGO, RE). The contact network has 327 nodes, while the friendship and sampled networks have 135 nodes. The density is twice higher in the contact network than in the friendship network; however, the subgraph induced by the nodes of the friendship network (SubFr) has in fact a slightly higher density than the contact network as well as the DRN sampled networks, that is not surprising as the DRN sampling chooses preferentially nodes with high degree. As the RN method samples uniformly the nodes of the contact network, the resulting density is on average equal to the density of the contact network. On the other hand, the density of the EGO sampled networks is even higher. Finally, RE and MSZ methods yield by construction the same density as the friendship network.

The clustering coefficient displays interesting features: despite a much lower density, the friendship network has a higher clustering coefficient than the contact network. Networks obtained through the SubFr, RN, DRN and EGO methods have as well rather large clustering coefficients, while the RE sampling yields much lower values. Similarly, the MSZ reshuffling of the friendship network makes the clustering coefficient drop as it breaks some correlation of the friendship network.

Figure 4.2 shows the outcome of the spreading simulations performed on the

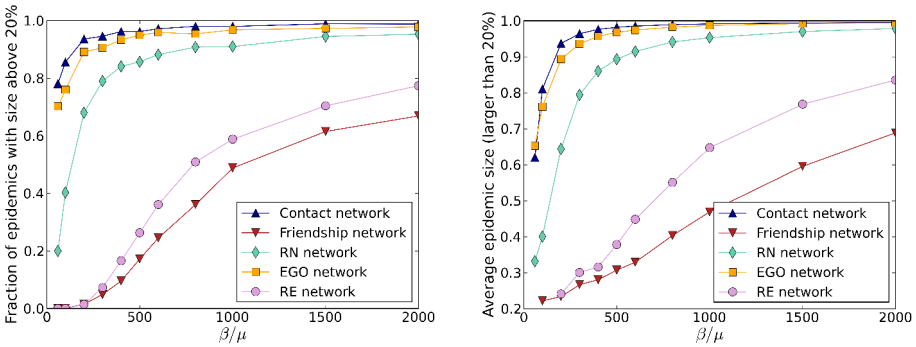
	Number of nodes	Number of edges	Density	Average degree	Average clustering	Avg shortest path*
Contact network	327	5818	0.11	35.6	0.50	2.15
Friendship network	135	413	0.05	6.1	0.53	4.06
SubFr	134	1235	0.14	18.4	0.55	2.22
RN	135	987	0.11	14.6	0.50	2.37
DRN	135	1218	0.13	18	0.50	2.18
EGO	135	1679	0.19	24.9	0.57	2.04
MSZ	135	413	0.05	6.1	0.07	2.92
RE	135	413	0.05	6.1	0.16	3.19

Table 4.1: Features of the empirical networks and of the networks obtained with the simplest sampling methods: SubFr, RN, DRN, EGO preserving the number of nodes and MSZ and RE preserving both the number of nodes and edges of the friendship network. *The average shortest path length is computed on the largest connected component of the network.

two empirical networks and on sampled networks obtained by the RN, RE and EGO sampling methods: it displays the fraction of epidemics with size above 20% and the average size of epidemics among the ones with size above 20%, as a function of the spreading parameter β/μ.

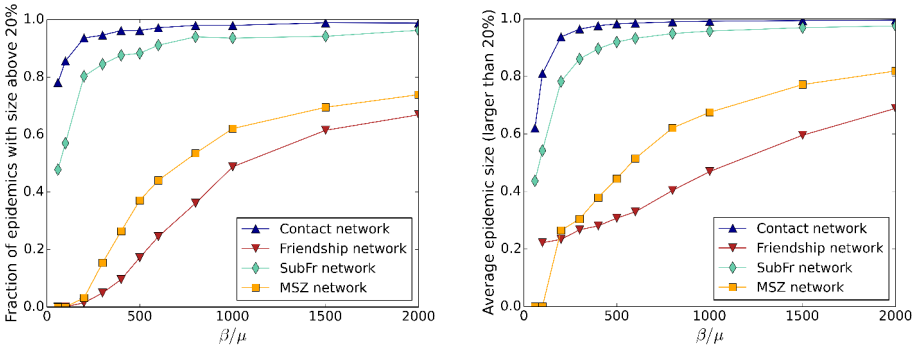
As found in Section 3.4, simulations using the friendship network give a very strong underestimation of the epidemic risk with respect to the ones using the contact network. The simulations performed on the sampled networks RN and EGO yield a much larger estimation of the epidemic risk than when using the friendship network but smaller than in the case of the contact network. Moreover, the estimation of the epidemic risk increases with the density of the sampled network, as expected since the availability of transmission paths increases with density. In the case of the RE sampling, the resulting networks have the same density as the friendship network, however simulations using RE sampled networks yield larger epidemic sizes than when using the friendship network: the main difference between these networks is that the friendship network has a much larger clustering coefficient. This is in agreement with the results of Smieszek et al. [61] stating that high clustering values tend to hinder propagation processes, at fixed density.

We show in Figure 4.3 the results of simulations performed on the SubFr network as well as on the randomized version of the friendship network using the MSZ method. The SubFr network corresponds to a population sampling of the contact network, and leads to a limited underestimation of the epidemic risk with respect to the whole contact network as the density is larger than in the whole contact network. Contrary to the RN case, the SubFr procedure is not a uniform sampling method. Indeed, there is a slight bias towards nodes with high degree: the 134 nodes of SubFr have an average degree of 37.9 in the contact network while 135 nodes chosen at random have an average degree of 35.6 in the contact network (equal to the average degree of the 327 nodes), explaining the discrepancies between the results of simulations performed on the resulting networks of these two methods of sampling. The MSZ network has the same number of nodes and edges than the frienship network, hence the same density, but a much smaller clustering, due to the randomization, and gives thus a higher epidemic risk, in agreement with [61].



(a) (b)

Figure 4.2: Outcome of SIR spreading simulations performed on empirical and sampled networks. (a) Fraction of epidemics with size above 20% (at least 20% of recovered individuals at the end of the SIR process) as a function of β/μ. (b) Average size of epidemic with size above 20% as a function of the parameter of spreading β/μ.



(a) (b)

Figure 4.3: Outcome of SIR spreading simulations performed on empirical, sampled and reshuffled networks. We compare here the case of the SubFr sampling and of the randomized friendship network (MSZ) with the empirical contact and friendship networks. (a) Fraction of epidemics with size above 20% (at least 20% of recovered individuals at the end of the SIR process) as a function of β/μ. (b) Average size of epidemics with size above 20% as a function of β/μ.

All methods used in this section yield intermediate results between results obtained with the contact network and results obtained with the friendship network. As expected, the sizes of simulated epidemics depend mostly on two characteristics of the network used for simulations: the density and the clustering coefficient, indeed the higher is the density, the higher is the epidemic risk, on the contrary, at fixed density, the smaller is the clustering, the higher is the epidemic risk.

4.2.2 More refined methods of sampling

We now turn to more sophisticated methods of sampling where the numbers of nodes and edges can be tuned to correspond to the friendship network and in particular we investigate the case of the EGOref sampling method. In the resulting EGOref sampled networks, the number of edges depends on the parameter p, at fixed number of sampled nodes N_F. Figure 4.4 shows the average number of edges in the sampled network as a function of p. For $p > \max_{i,j}(s_i/W_{ij})$, this number reaches the average number of edges in the subgraph induced by N_F nodes chosen at random on the contact network, which is equal to the number of edges obtained through the RN sampling process. As our goal is to obtain a sampled subgraph of the contact network that is similar to the friendship network, we tune p in order to obtain sampled networks with an averaged number of edges close to $E_F = 413$, the number of edges in the friendship network. This value is obtained for $p \approx 31.3$ (Figure 4.4). At this value, we point out that the obtained network is always connected.

In Table 4.2, we report the properties of the networks sampled using the RE, WRE, RNref and EGOref methods. These methods allow to choose the number of edges and we choose it to be equal to E_F, the number of edges in the friendship network. The clustering coefficient is systematically smaller in sampled networks than in the friendship network, but a better agreement is found for the WRE and EGOref networks. The average shortest path length is also smaller in sampled networks than in the friendship network, except for the sampled network using WRE method; again the WRE and EGOref sampled networks are the ones with the average shortest path length closest to the case of the friendship network. Further refinements of the EGOref method might yield a clustering closer to the one of the friendship network, at the cost however of an increase in the method's complexity and number of parameters.

	Number of nodes	Number of edges	Density	Average degree	Average clustering	Avg shortest path*
Friendship network	135	413	0.05	6.1	0.532	4.06
RE	135	413	0.05	6.1	0.157	3.19
WRE	135	413	0.05	6.1	0.371	4.22
RNref	135	413	0.05	6.1	0.197	3.21
EGOref	135	413	0.05	6.1	0.355	3.90

Table 4.2: Features of the friendship network and of the sampled networks obtained by RE, RNref and EGOref method, all of them preserving the number of nodes and edges of the friendship network. *The average shortest path length is computed on the largest connected component of the network.

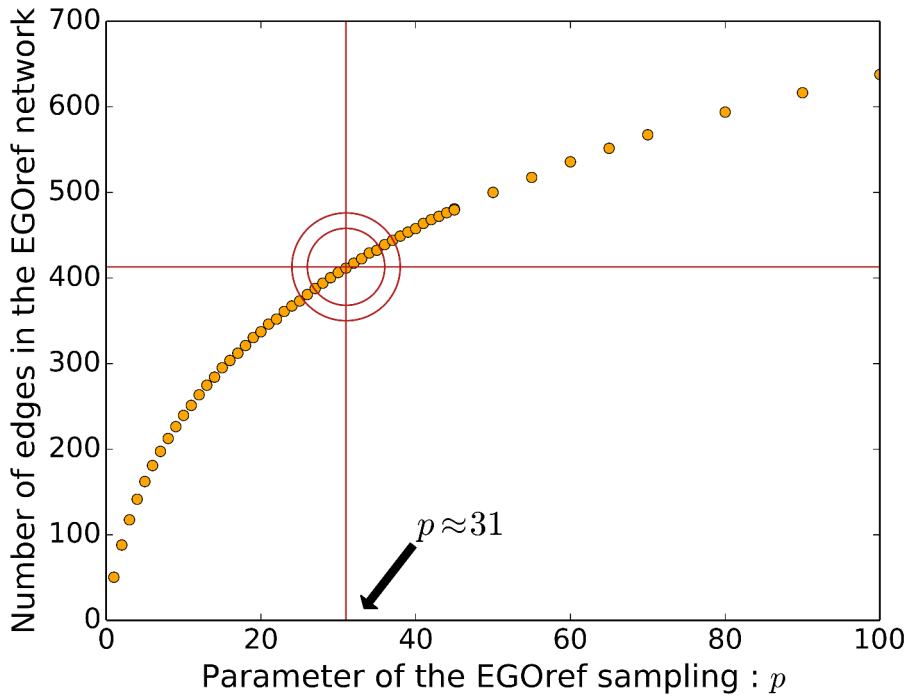
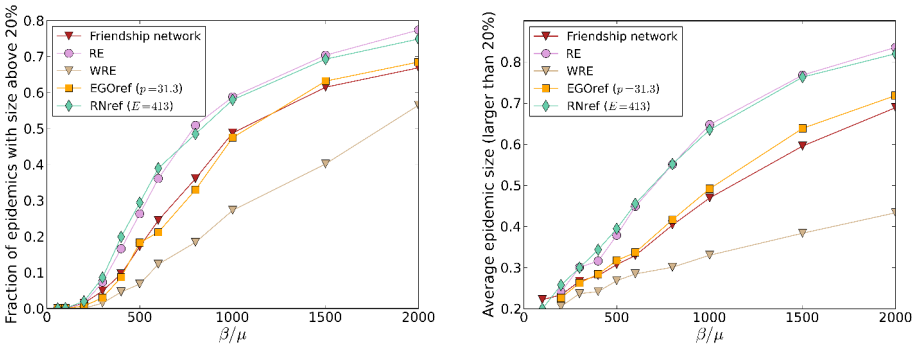


Figure 4.4: Average number of edges in the sampled network obtained by the EGOref method as a function of the parameter p with $N_F = 135$. The horizontal red line represents the number of edges E_F in the friendship network.

Figure 4.5 shows the outcomes of epidemic spreading simulations performed on the friendship network and on the contact networks sampled using the EGOref, RNref, RE and WRE methods. A very good agreement with the epidemic risk estimated from the friendship network is obtained for the EGOref sampling, while the RE and RNref sampled networks yield large epidemic sizes with higher probability and larger average epidemic sizes, even if they have the same density. In the case of the WRE sampling, the epidemic sizes are smaller than the friendship network: this can be explained by the larger average shortest path length (as the nodes are on average further than in the friendship network, the spreading is slowed down) and the rather large clustering coefficient. Figure 4.6 displays the whole distributions of epidemic sizes for simulations performed on the friendship and EGOref networks, for 4 values of the spreading parameter β/μ. A good agreement in the shape of the distributions is observed, although the maximal size of epidemics is systematically higher in the EGOref sampled contact networks than in the friendship network, especially at large β/μ.

As a further refinement, we consider the case of the EGOref-het sampling method. In that case, the nodes are not sampled uniformly but we select a certain number of nodes in each class (uniformly in each class) such that the



(a) (b)

Figure 4.5: Outcome of SIR spreading simulations performed on friendship and sampled networks. (a) Fraction of epidemics with size above 20% (at least 20% of recovered individuals at the end of the SIR process) as a function of the parameter of spreading β/μ. (b) Average size of epidemic with size above 20% as a function of the spreading parameter β/μ.

repartition of nodes in classes is the same as in the friendship network. The number of edges is fixed at the same value as before i.e., 413 edges; in that case, this number is obtained for $p \approx 22$. The resulting average shortest path length is 4.03 and the average clustering is 0.334, still smaller than in the friendship network and in fact, very close to the clustering in the EGOref sampled networks. We show in Figure 4.6 that the outcomes of the spreading simulations are similar to the EGOref case, with a slightly better agreement with the results of simulations using the friendship network, in particular for large epidemics. This refinement of the model of sampling appears not so useful as it does not bring strong improvements to the results of simulations.

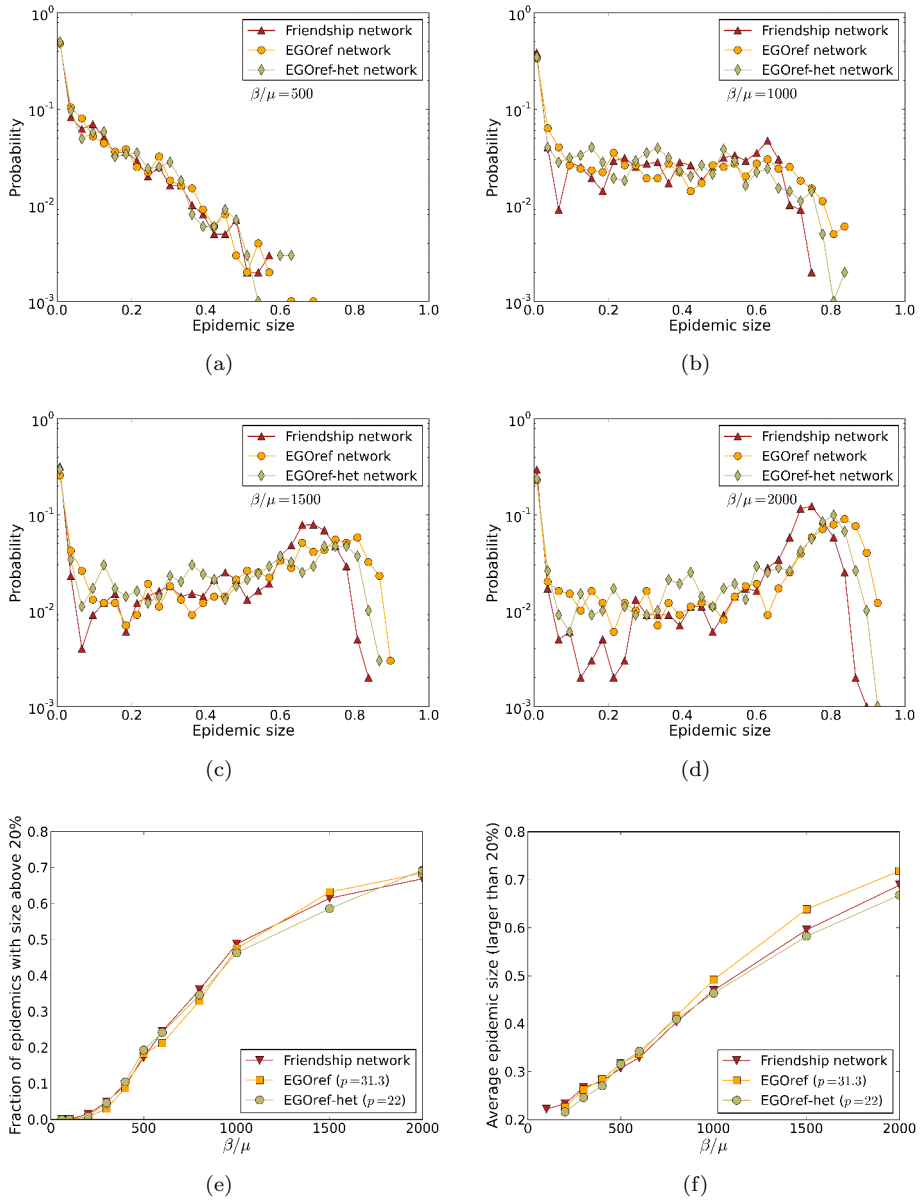


Figure 4.6: Outcome of SIR spreading simulations performed on friendship and both EGOref sampled networks. (a)-(d) Distributions of epidemic sizes for different values of β/μ. (e) Fraction of epidemics with size above 20% (at least 20% of recovered individuals at the end of the SIR process) as a function of the parameter of spreading β/μ. (f) Average size of epidemic with size above 20% as a function of the spreading parameter β/μ.

4.3 Sampling model exploration

In this section, we investigate how simulations of spreading processes performed on networks obtained from the contact network using the EGOref sampling method depend on the method's parameters p and N. We consider $135 \leq N \leq 327$ and $15 \leq p \leq 500$ (as for $p > 500$ the number of links is almost equal to its maximum possible value). We first observe (Figure 4.7) that, at fixed N, the density of the sampled network is fully determined by the value of p. Changing p is thus equivalent to tuning the resulting network's density. For $p = 500$ and $N = 327$, we almost recover the whole contact network.

In Figure 4.8, we show the average epidemic size obtained on the sampled networks as a function of p and N for different values of the spreading parameter β/μ: this size increases with both p and N. Increasing p (which would correspond to the probability to report a link as a friendship in the case of the friendship survey) at fixed N (which would correspond the number of participants to the survey) or the contrary is not enough to obtain a correct estimation of the epidemic risk: both have to be increased in order to obtain the same value as when using the whole contact network, shown by the continuous line. The dashed lines show the values of p and N necessary to obtain an estimation of the epidemic size within 5%, 10% or 20% of this reference value.

We also note that the average epidemic size obtained for the largest values of p and N is actually larger than the reference, although the corresponding EGOref network is structurally almost the same as the contact network. This discrepancy stems from the fact that the edge weights are placed differently in both cases: weights are indeed assigned at random on the edges of the network obtained by the EGOref sampling procedure. In the next section, we investigate more in detail the reasons for this discrepancy.

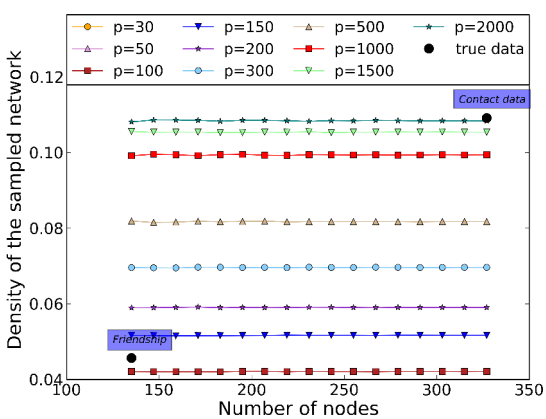
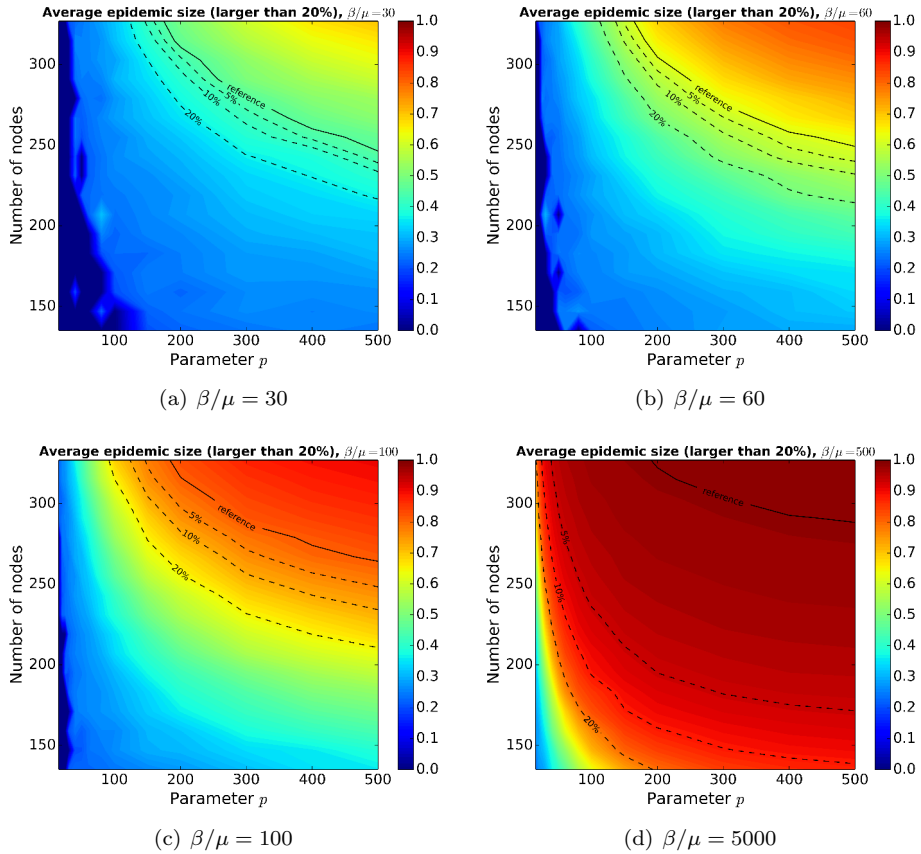


Figure 4.7: Average density of the resulting EGOref sampled networks as a function of the number of nodes for different values of p.



(a) $\beta/\mu = 30$

(b) $\beta/\mu = 60$

(c) $\beta/\mu = 100$

(d) $\beta/\mu = 5000$

Figure 4.8: Color maps of the average epidemic size for epidemics with size above 20% for several values of β/μ. When no epidemics has size above 20%, the value is zero. The three dashed lines represent the value of average epidemic size at 5%, 10%, 20% of the reference value (solid line), which corresponds to the average epidemic size of the SIR spreading simulations performed on the contact network.

4.4 Impact of weight assignment

We have adressed in Section 4.1.2 the need for weight assignment to edges of unweighted network to properly compare the outcomes of simulations of epidemic spreading; the results will be however different depending on the choice of the weight assignment method. We investigate here the impact of several possible ways to assign weights to the networks used for the simulation of spreading processes, namely the empirical contact network measured by the wearable sensors, the network of reported friendships, and the networks obtained through various sampling procedures from the contact network. Indeed:

- the measured contact network is weighted, as the sensors give access to the duration of contacts. We can therefore consider the weighted network with its true weights ("Original weights") or, to assess the impact of correlations between weights and structure, reshuffle them at random among the network's links ("Reshuffled weights").

- any sampling procedure produces a subgraph of the contact network. As a consequence, the weights of the edges can be either taken directly from the contact network ("Original weights"), under the hypothesis that the sampling procedure keeps information about the weights. On the other hand, the opposite hypothesis, that sampling only informs about the existence of a link, and not on its importance, leads us to another weight assignment procedure, namely a random weight assignment from the distribution of weights of the contact network ("Random weights"). This is the procedure considered in the previous sections, as it is the most parsimonious and realistic in terms of availability of information.

- the friendship network is not weighted. In order to use it in the simulations of SIR process, we can assign weights to links in different ways:

  - we can choose the weights randomly from the distribution of weights in the contact network ("Random weights"), again this method is the one used in the previous sections for the same reason, or,

  - for each edge of the friendship network present in the contact network (86% of the links in the friendship network find a corresponding link in the contact network), we can use the corresponding weight and, for the remaining 14%, we can take the weights at random from the distribution of weights obtained from the first step (we call this assignment procedure "Original weights"), or,

  - we assign the weights as in the previous method and then reshuffle randomly the weights among the links of the friendship network ("Reshuffled weights").

4.4.1 Contact network and sampling procedures independent from weights

Most of the methods of sampling used on the contact network (RN, DRN, RNref, RE, EGO and SubFr) sample edges independently from their weights, thus the distribution of weights of the contact network is preserved in the sample neworks. In this sense, the "Random weights" procedure performed on these sampled networks is equivalent to the "Reshuffled weights" performed on the contact network. Indeed, reshuffling the weights does not change the distribution of weights.

Even though the distribution of weights is unchanged by the reshuffling procedure, the simulations performed on the contact network with "Reshuffled weights" yield larger epidemic sizes than with original weights (Figure 4.9). Similarly, in the case of sampled networks, we show in Figure 4.10 that the simulations performed with the use of the "Random weights" assignment procedure leads systematically to larger epidemic sizes than the use of the "Original weights" assignment procedure. Given the fact that the distribution of weights is unchanged in both procedures, this discrepancy is a sign of some correlations between weights and structure of the network that hinders the propagation of simulated epidemics.

We investigate this potential correlation in Figure 4.11: it displays the ratio s_k/k as a function of k for the contact network and the RN, EGO and RE sampled networks (similar results are obtained for DRN, RNref and SubFr sampled networks), where s_k is the average strength of nodes of degree k. The two different ways of assigning the weigths yield different behaviors: (i) when weights are shuffled or assigned at random, s_k/k is independent of k, (ii) on the contrary, when the original weights are used, a distinct trend is observed, with smaller strengths at large k than for the random/reshuffled weights. In the latter case, the hubs (i.e., nodes with high degree) have smaller spreading power than expected by random chance, and the epidemic spread is hindered, leading to smaller epidemic risk.

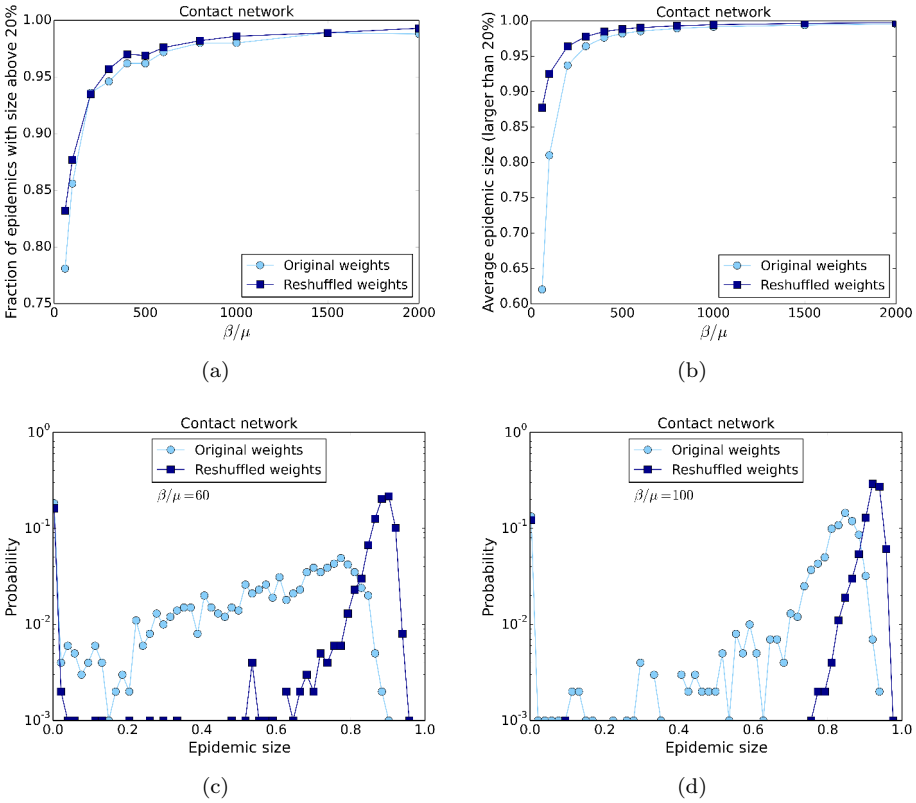


Figure 4.9: Outcome of SIR spreading simulations performed on contact network with "Original weights" and "Reshuffled weights". (a) Fraction of epidemics with size above 20% (at least 20% of recovered individuals at the end of the SIR process) as a function of β/μ. (b) Average size of epidemics with size above 20% as a function of β/μ. (c) and (d) Distributions of epidemic sizes for two different values of β/μ.

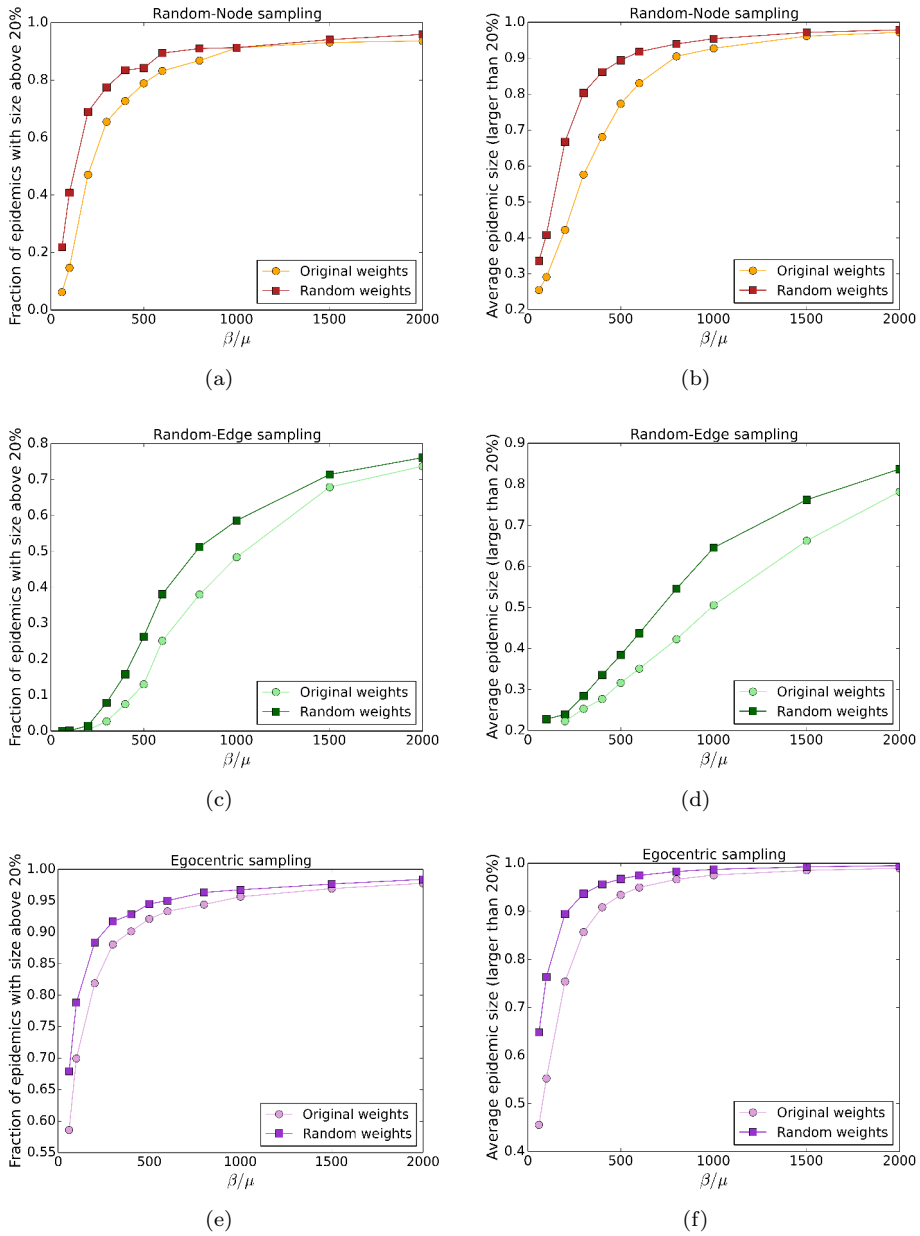


Figure 4.10: Outcome of SIR spreading simulations performed on contact networks sampled with the RN, RE and EGO methods, and with weights assigned either through the "Original weights" or "Random weights" procedures. (a),(c),(e) Fraction of epidemics with size above 20% (at least 20% of recovered individuals at the end of the SIR process) as a function of β/μ. (b),(d),(f) Average size of epidemics with size larger than 20% as a function of β/μ.

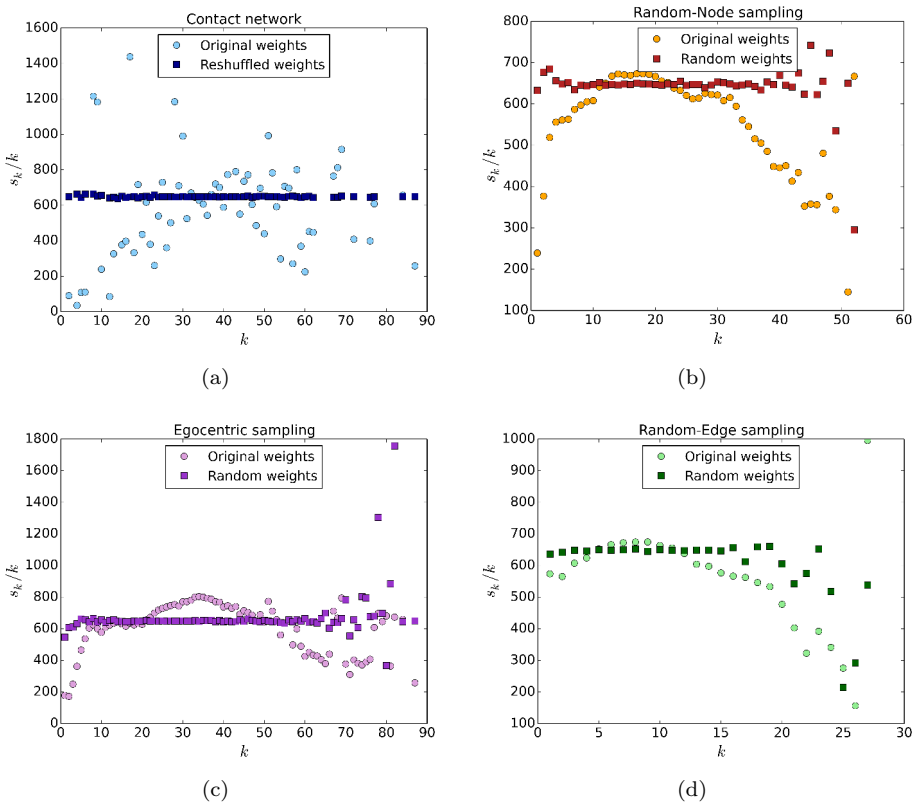
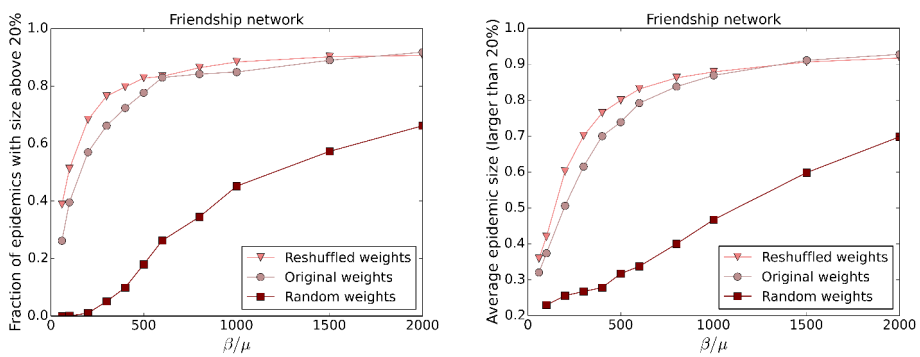


Figure 4.11: Comparison of average strength s_k of nodes with degree k, divided by the degree k, as a function of k between the "Original weights" and the "Reshuffled/Random weights" cases for (a) the contact network, (b) the RN-sampled network, (c) the EGO-sampled network, (d) the RE-sampled network.

4.4.2 Friendship network

Figure 4.12 compares the outcomes of SIR simulations on the friendship network with weights assigned in the three different ways described above. The size of epidemics is a little higher in the case of "Reshuffled weights" than in the case of "Original weights": this is due to the same mechanism as for the contact network, i.e., to correlations between weight and structure that are destroyed by the reshuffling.

Simulations on the network with "Random weights" lead on the other hand to a much smaller epidemic risk. In Figure 4.13 we show the distributions of weights of all the links of the contact network (in blue, 5818 links) and of friendship links that are also present in the contact network (in red, 348 links). As already found in the previous chapter, the distributions are not the same and the links present in both networks tend to correspond to larger cumulative durations. The blue distribution is the one used in the "Random weights" case, while the red one is used in"Original weights" and "Reshuffled weights". The average weight is larger in two latter cases and this of course favours the spread.



(a) (b)

Figure 4.12: Outcome of SIR spreading simulations performed on the friendship network with "Original weights", "Reshuffled weights" and "Random weights". (a) Fraction of epidemics with size above 20% (at least 20% of recovered individuals at the end of the SIR process) as a function of β/μ. (b) Average size of epidemic with size above 20% as a function of β/μ.

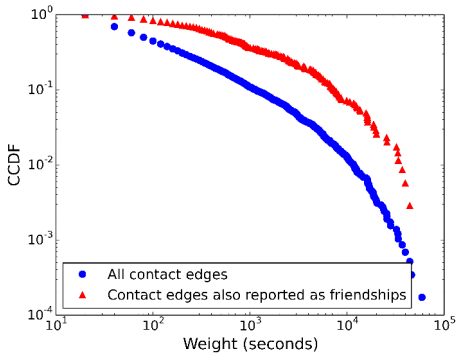


Figure 4.13: CCDF of weights for two different kinds of edges in the contact network: (i) all edges, (ii) edges corresponding to a reported friendship (i.e., present in both friendship and contact networks).

4.4.3 WRE sampling procedure

The case of the WRE sampling procedure is different from the other sampling techniques as it selects edges with a probability depending on their weights. Thus, the distribution of "Original weights" of the edges selected throughout the sampling is different from the distribution of weights of edges of the contact network used in the "Random weights" assignment procedure (Figure 4.14): the distribution of "Original weights" is shifted towards larger weights.

Figure 4.15 shows the outcomes of simulations performed on the WRE sampled networks with both weight assignment techniques. Contrary to sampling techniques independent from weights, the epidemic sizes are larger in the case of "Original weights" than with "Random weights" and the difference between the two cases is also larger. In the case of "Random weights" there is no correlations between weights and structure, while in the case of "Original weights" the correlations between weights and structure are preserved but they are counterbalanced by the larger weights in the sampled networks such that epidemic sizes are larger.

A very good agreement is found between the results of SIR simulations performed on WRE sampled networks with "Original weights" assignment and the results obtained with the friendship network (with "Random weights" assignment), yet less good than between EGOref sampled networks and friendship network (Figure C.1). However this agreement is misleading as it might be the result of compensation between two effects in this specific case: (i) in the friendship network, there is no correlations weights-structure (accerelerates the propagation of simulated epidemics) and the distribution of weights of the contact network is preserved (ii) in the WRE sampled networks, the correlations weights-structure are preserved (hinders the propagation) but weights are larger (risk of infection from a Infectious node to a Susceptible node is higher). Moreover, it assumes that we are able to measure the weights, while in the EGOref

case we can use a distribution of weights from other data sets.

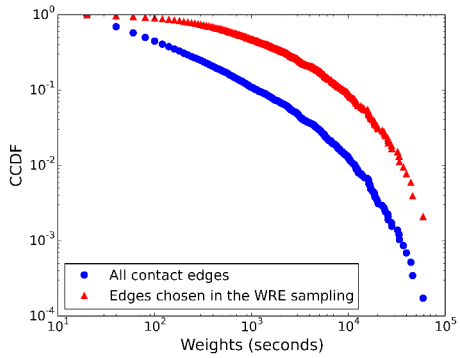
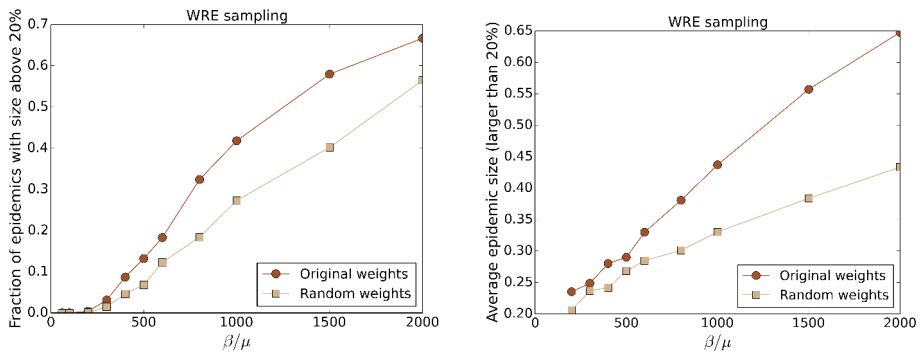


Figure 4.14: CCDF of weights for two different kinds of edges in the contact network: (i) all edges, (ii) edges which were chosen in the WRE sampling.



(a) (b)

Figure 4.15: Outcome of SIR spreading simulations performed on the WRE sampled networks with "Original weights" and "Random weights". (a) Fraction of epidemics with size above 20% (at least 20% of recovered individuals at the end of the SIR process) as a function of β/μ. (b) Average size of epidemic with size above 20% as a function of β/μ.

4.4.4 EGOref sampling procedure

In the case of the EGOref sampling procedure, the interesting part is that it combines the two effects discussed above: there is a competition between the fact that the random assignment of weights does not take into account the correlations weights-structure and the fact that the EGOref sampling procedure chooses preferentially edges with large weights. Indeed, the difference between the outcomes of simulations performed on EGOref-sampled networks using the "Original weights" and "Random weights" assignment procedures depends on the parameters of the sampling procedure p and N, as shown in Figure 4.16.

For small p, the average epidemic size is higher in the case of "Original weights" whereas for large p, it is higher in the "Random weights" case. This can be explained by the two following competing effects:

- at small p, relatively few edges are selected in the contact network, and each ego selects preferentially links with large weights. The resulting distribution of original weights is thus biased towards large weights, and the weights are on average larger than when using weights taken at random from the overall distribution of weights of the contact network. This tends to favour the spread and thus leads to a larger epidemic risk for "Original weights" than for "Random weights".

- at large p, the probability to select an edge is large even for links with small weights. As a result, the distribution of weights of sampled links becomes close to the global distribution of weights in the contact network used in the "Random weights" cases. The correlations between weights and structure present in the contact network can then play a role and act in the same way as for the RN, RE and EGO sampling methods: a random assignment of weights destroys the correlations and favours the spread.

Finally, we note that the value of N (Figures 4.16(b) and (d)) does not change the sign of the difference between the epidemic risk obtained by the two weight assignment procedures, but only its amplitude.

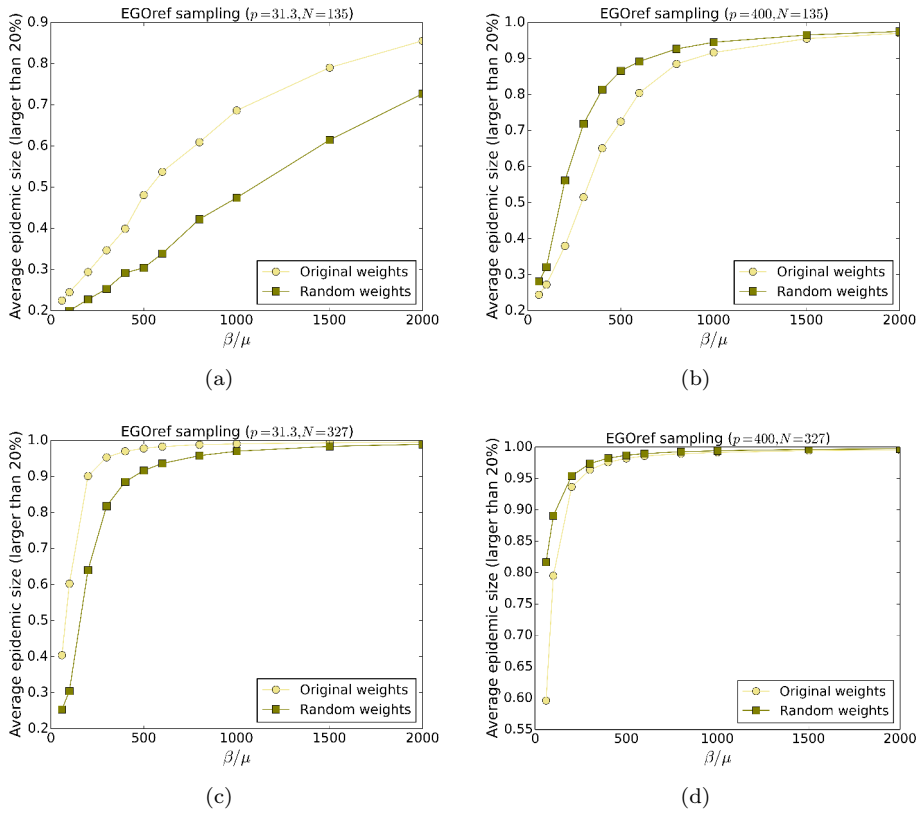


Figure 4.16: Average size of epidemic (with size above 20%) for SIR spreading simulations performed on EGOref-sampled network with "Original weights" and "Random weights" for four different couples of p and N: (a) $p = 31.3, N = 135$, (b) $p = 400, N = 135$, (c) $p = 31.3, N = 327$, (d) $p = 400, N = 327$.

4.5 The EGOref sampling in other contexts

The previous sections have shown the efficiency of the EGOref sampling in reproducing the results of SIR simulations performed on the friendship network of our data set. However, the efficiency of this method of sampling is attested in only one case: the current lack of data sets combining contact data and friendship data prevents us from testing this efficiency in other cases. Given this fact, we can still study the consequences of the EGOref sampling on other data sets decribing face-to-face contacts and compare to the impact of other types of sampling.

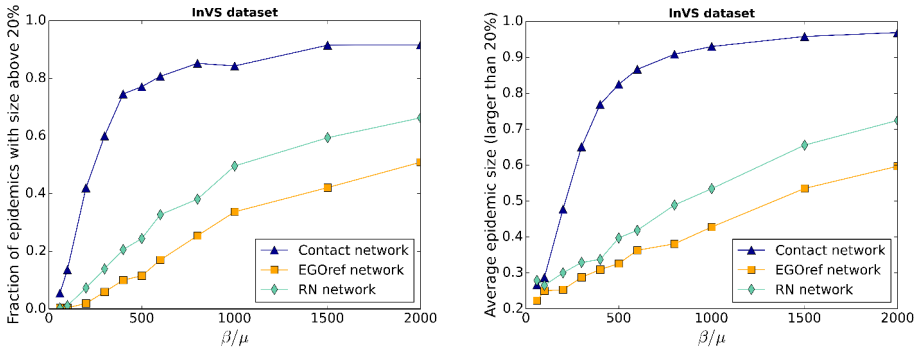
We consider here two data sets describing contacts in (i) offices (InVS) and (ii) a conference (SFHH) already mentioned in Section 1.3.2. The InVS data contains the contacts measured in offices during two weeks between 92 individuals, while the SFHH data describes contacts between 403 individuals during the two days of a conference.

As a first step, we perform the RN and EGOref sampling methods on both InVS and SFHH data sets. For this, we use the same parameters of sampling used in the case of high school (nodes are sampled at 41% and for the EGOref case we choose $p = 31.3$) and the same method of weight assignment ("Random weights" procedure). In Table 4.3, we report the number of nodes and edges in each empirical and sampled networks: in both cases, the RN sampling preserves the density of the contact network and the density is much smaller in EGOref sampled networks. We show in Figures 4.17-4.20, the outcome of simulations performed on these networks. The simulations performed on sampled networks yield much smaller epidemic sizes than with the use of the contact network, moreover the epidemic sizes are smaller with the EGOref sampling than with the RN sampling which is just a population sampling: the ranking between the two methods is the same as for the high school case. However, it is impossible to compare the results with a potential friendship network.

Finally, we show in Figures 4.21-4.22, the equivalent of Figure 4.8 in the case of InVS and SFHH data sets, highlighting the combined effects of population sampling and of the absence of links with small weights in the sampled network.

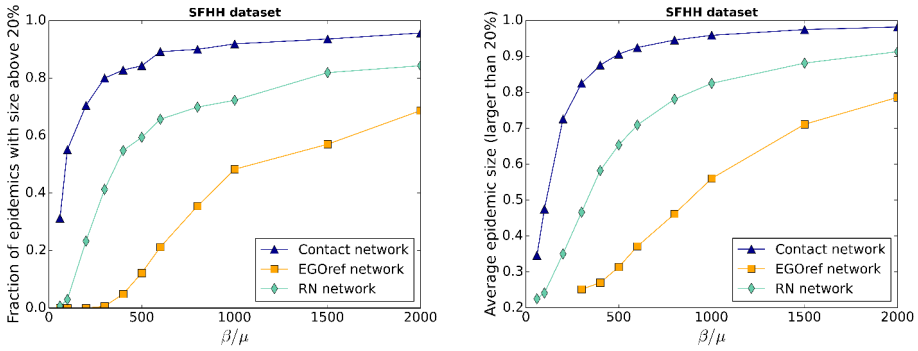
	Number of nodes	Number of edges	Density
InVS Contact network	92	755	0.18
InVS EGOref network	37	88	0.13
InVS RN network	37	119	0.18
SFHH Contact network	403	9565	0.12
SFHH EGOref network	165	766	0.06
SFHH RN network	165	1598	0.12

Table 4.3: Number of nodes and edges in the empirical and sampled data sets. The sampled networks consider the same fraction of nodes as for the data set used in high school case.



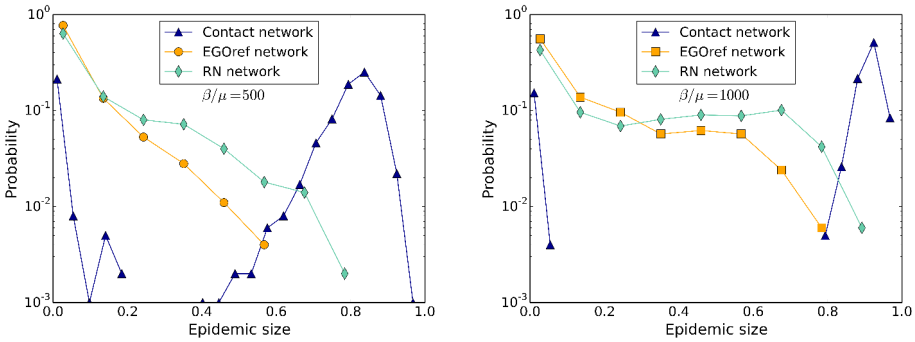
(a) (b)

Figure 4.17: Outcome of SIR spreading simulations performed on empirical and sampled contact networks (InVS data set). We compare here the simulations on the original contact network with a sampled network using the EGOref sampling procedure with $p = 31.3$ and $N = 165$ nodes (corresponding to a sampling fraction equal to the case of of Lycée Thiers) and with the RN case (still with $N = 165$ nodes). (a) Fraction of epidemics with size above 20% (at least 20% of recovered individuals at the end of the SIR process) as a function of β/μ. (b) Average size of epidemics with size above 20% as a function of β/μ.



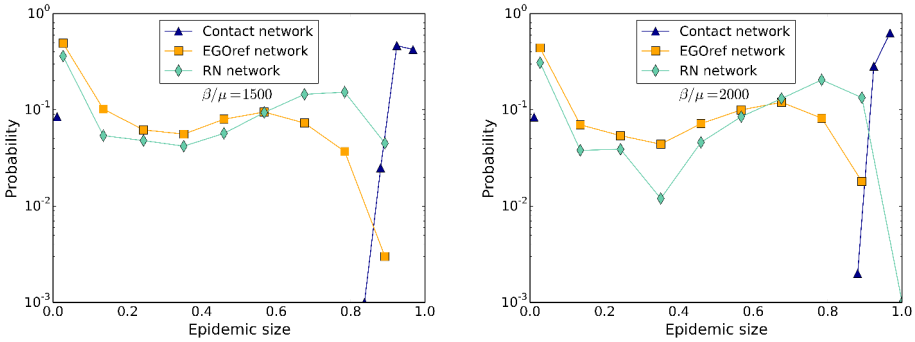
(a) (b)

Figure 4.18: Outcome of SIR spreading simulations performed on empirical and sampled contact networks (SFHH data set). We compare here the simulations on the original contact network with a sampled network using the EGOref sampling procedure with $p = 31.3$ and $N = 37$ nodes (corresponding to a sampling fraction equal to the case of of Lycée Thiers) and with the RN case (still with $N = 37$ nodes). (a) Fraction of epidemics with size above 20% (at least 20% of recovered individuals at the end of the SIR process) as a function of β/μ. (b) Average size of epidemics with size above 20% as a function of β/μ.



(a) $\beta/\mu = 500$

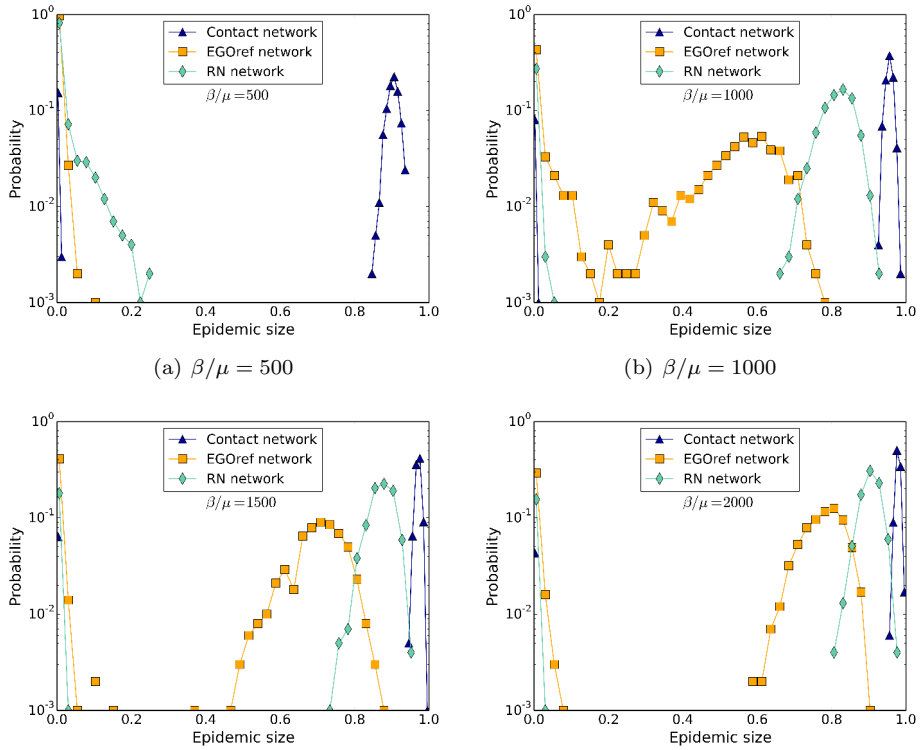
(b) $\beta/\mu = 1000$



(c) $\beta/\mu = 1500$

(d) $\beta/\mu = 2000$

Figure 4.19: Distributions of epidemic sizes of SIR spreading simulations (InVS data set). We compare the distributions of epidemic sizes for simulations performed on the original contact network and on the sampled network using the EGOref sampling procedure with $p = 31.3$ and $N = 37$ nodes (corresponding to a sampling fraction equal to the case of Lycée Thiers) and with the RN case (still with $N = 37$ nodes), for different values of β/μ.



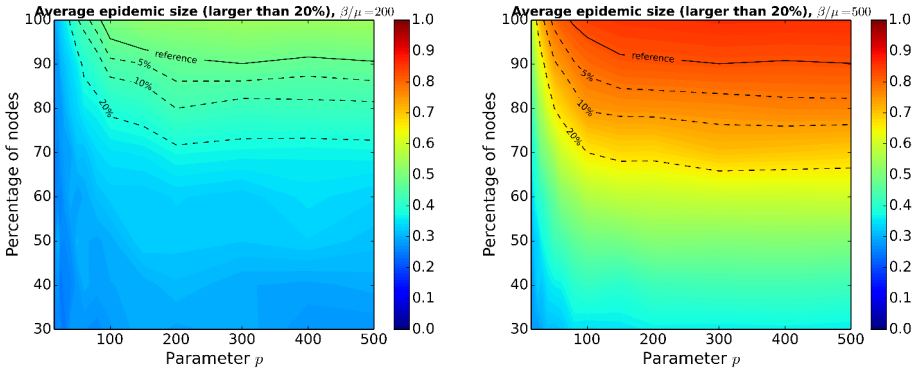
(a) $\beta/\mu = 500$

(b) $\beta/\mu = 1000$

(c) $\beta/\mu = 1500$

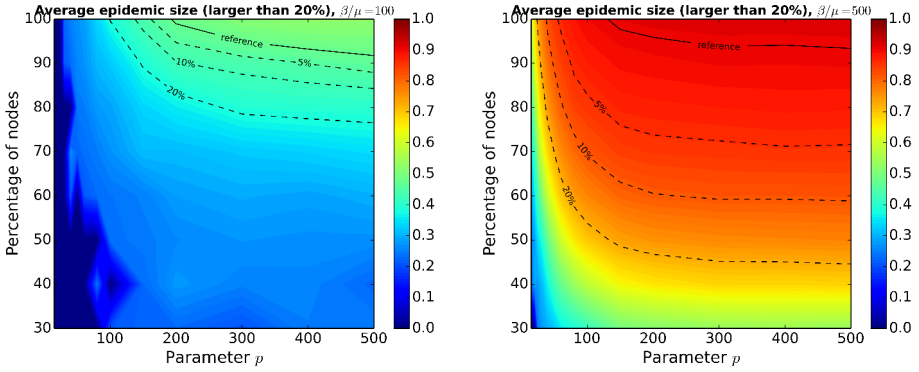
(d) $\beta/\mu = 2000$

Figure 4.20: Distributions of epidemic sizes of SIR spreading simulations (SFHH data set). We compare the distributions of epidemic sizes for simulations performed on the original contact network and on the sampled network using the EGOref sampling procedure with $p = 31.3$ and $N = 165$ nodes (corresponding to a sampling fraction equal to the case of Lycée Thiers) and with the RN case (still with $N = 165$ nodes), for different values of β/μ.



(a) $\beta/\mu = 200$ (b) $\beta/\mu = 500$

Figure 4.21: Color maps of the average epidemic size for epidemics with size above 20% for several values of β/μ (InVS data set). When no epidemics has size above 20%, the value is zero. The three dashed lines represent the value of average epidemic size at 5%, 10%, 20% of the reference value (solid line), which corresponds to the average epidemic size of the SIR spreading simulations performed on the contact network.



(a) $\beta/\mu = 100$ (b) $\beta/\mu = 500$

Figure 4.22: Color maps of the average epidemic size for epidemics with size above 20% for several values of β/μ (SFHH data set). When no epidemics has size above 20%, the value is zero. The three dashed lines represent the value of average epidemic size at 5%, 10%, 20% of the reference value (solid line), which corresponds to the average epidemic size of the SIR spreading simulations performed on the contact network.

4.6 Conclusion and outlook

In the context of the spread of infectious diseases, contact networks are considered as a relevant proxy of transmission possibilities [26]. However, gathering objective data on contacts is not always feasible and other types of data, such as friendship networks, could be easier to obtain than contact networks in some situations. Moreover, surveys asking individuals about their friends might suffer less from memory biases than contact diaries. In Section 3.4, we have found, considering a data set combining contacts and friendship relations, than the use of the friendship network for simulations of epidemic spreading leads to an underestimation of the epidemic risk with respect to the use of the contact network obtained from sensors. In this chapter, we have investigated if this underestimation can be seen as biases resulting from a sampling procedure performed on the contact network. The rationale leading to this question comes from the quantitative comparison between the friendship and contact networks performed in the previous chapter: friendship and contact networks are indeed different, the friendship network has much less nodes than the contact network, and many short contacts occur between individuals who are not friends; however, the longest contacts, which play an important role in potential propagation events, effectively correspond to friendship links, and the overall structure of the networks, in terms of interactions between different classes, are similar. Understanding if and how friendship surveys could be used in models of spreading processes would thus be interesting. Most importantly, framing the relation between friendship networks and actual contact networks as a sampling procedure might also help design and evaluate procedures to compensate for the resulting biases in the estimation of epidemic risks.

To make progresses in this direction, we have considered several ways of sampling the contact network and investigated the similarities of the resulting sampled networks with the friendship network with respect to the outcomes of simulations of epidemic spread. The friendship network has much less nodes than the contact network, so that, we have first considered a uniform sampling of nodes. Assuming that the individuals who have answered the friendship survey are more "important" in the contact network, we have considered the DRN method which samples nodes with probability proportional to their degrees in the contact network. These two methods simply sample nodes and consider the subgraph induced by these nodes on the contact network, while the EGO method is closer in its mechanism to a survey procedure as it includes nodes chosen at random and all their contacts. The resulting networks obtained with these three methods have much higher densities than the friendship network; in particular the density of the DRN sampled networks is equivalent to the density obtained by taking the subgraph induced by the respondent nodes to the friendship network on the contact network (SubFr method). The epidemic risk obtained with simulations performed on these sampled networks is much higher than in the case of the friendship network. We have therefore considered sampling methods in which the density could be tuned to be equal to the density of the friendship network. In the RE and RNref methods, we select edges

at random, which leads to a rather small clustering coefficient in the sampled networks with respect to the friendship network and thus yields larger epidemic sizes. These methods seem indeed too naive to recover the small scale structures and correlations of the friendship network. As the friendship links present in the contact network tend to correspond to edges with large weights, we have then considered the WRE method which samples edges with a probability proportional to their weights in the contact network. The resulting networks have a rather large clustering coefficient but still smaller than in the friendship network, however the average shortest path length is higher in the WRE sampled networks, i.e., the nodes are further apart than in the friendship network; as a consequence the epidemic risk is smaller with this method of sampling than with the friendship network.

Finally, we designed the EGOref sampling method as a way to mimic a survey procedure in which individuals (egos nodes) report on their friendships under the assumptions that all reported friendships correspond to contacts, and that the probability that a contact corresponds to a friendship is larger for longer contacts. Note that this method is not aimed at reproducing exactly the friendship network (in which some links in fact do not correspond to contacts), but its goal is to produce a sampled network whose properties are close enough to the friendship network to lead to similar outcomes when used in simulations of spreading processes. This method of sampling depends mainly on a single parameter p which determines the number of edges an ego would report as friendships: actually, at fixed number of egos, the density of the sampled networks is fully determined by this parameter. In particular, we can choose the number of egos equal to the number of nodes in the friendship network and tune the parameter p to have the same density in the sampled networks than in the friendship network. At fixed density, this method allows us to recover a higher coefficient of clustering than the other sampling methods. Most importantly, simulations of spreading processes performed on the resulting sampled network yield an estimation of the epidemic risk in very good agreement with simulations using the friendship network, for a wide range of spreading parameters. Note that the EGOref method samples egos uniformly at random while the EGOref-het method chooses egos in the various classes to correspond to the numbers of indivuals in each class in the friendship network. The value of p required to have the desired number of edges is smaller with this variation of the EGOref method but the outcomes of simulations do not show significant improvements with respect to the EGOref method. Moreover, the clustering coefficient is still lower than in the friendship network. More involved sampling procedures allowing to control the clustering might be sought, but would result in more complex and less intuitive sampling rules.

Some limitations of our approach are worth discussing. First, the way we choose to assign the weights to edges in the sampled networks can be discussed. On the one hand, the distribution of weights is known to be very robust in human contact networks, even in very different contexts such as schools or hospitals [10, 12], so that it seems natural to use the empirical distribution of weights, which can be taken from publicly available datasets, to assign weights

to links obtained through surveys. Here the distribution of weights used is the one of the contact network obtained from sensors. On the other hand, a random assignment of weights destroys correlations between the weights and the structure of the network that can have an impact on the outcome of spreading simulations, a scenario verified in our case. We have considered such a random assignment as our goal is to compare the friendship network to a sampled contact network: when only friendship data is available we do not have information on the weights so a natural way to perform simulations of epidemic spreading processes is indeed to assign weights at random to the friendship links. More accurate ways to assign weights in order to mimic the correlations linking the strengths and degrees of nodes in the contact network would be of great interest, but might depend on the context and would require supplementary information. Moreover, here we have considered static networks while real contact networks evolve over time. As mentioned above, this is based on a twofold rationale: first the data obtained from friendship surveys does not contain temporal information, the comparison should be done with a sampling of a static version of the contact network; then for slow propagation timescales corresponding e.g. to flu-like illnesses, the precise dynamic of the contacts does not represent a crucial information [8]. In the case of faster processes where the temporal evolution of the network is important, one should create surrogate timescales on the links similarly to [43]. Finally the main limitation comes from the fact that our study is limited to one specific population. This is due to the current lack of datasets combining both contacts and friendship relations. Further investigations in different contexts would be of great interest.

The sampling method we have proposed here depends on two parameters: the number of egos N and p. In the context of a survey, these two numbers allow to tune the number of respondents and the amount of contacts reported. This makes its application possible in various contexts; actually, we have started to study the impact of this method of sampling on various data sets, yet the comparison with a corresponding friendship network is impossible as these data sets do not combine both contacts and friendship data. Following the work of [43], we have also started preliminary investigations on reconstruction strategies to obtain better estimates of the epidemic risk from sampled network, with respect to the results of simulations performed on the contact network. However, in [43] Génois et al. have chosen a uniform method of sampling in which selected nodes (and their edges) are removed from the contact network which is equivalent to the RN method described in this chapter. This method preserves the density of the contact network which allows to rebuild a surrogate contact network by adding the desired number of nodes (to obtain the total number of nodes in the target population) and a certain number of edges chosen to preserve the overall density. Moreover, in data sets where nodes are naturally separated in different groups (e.g., classes for high school), it is also possible to preserve the density at the scale of these groups when it is necessary; indeed in the case of Lycée Thiers data set the density is much higher inside classes than between different classes which is known to have an importance in the context of epidemic spreading. The case of the EGOref method of sampling is different as the

sampling of edges is not uniform and in the case described in this chapter, at small p the density is much lower than the density of the contact network. Thus a method of reconstruction equivalent to the one described in [43] turns out not to be to obtain a good estimation of the epidemic risk with the reconstructed networks. In Appendix (Figures C.3-C.7), we show the results of simulations of epidemic spreading obtained before and after reconstruction (the method of reconstruction is equivalent to the one of [43]) of the sampled networks. At small p and after reconstruction, the epidemic risk is still much smaller than in the case of the contact network. Moreover, the epidemic risk depends only on p and the number of egos N chosen in the sampling phase does not have any impact on the epidemic risk simulated after reconstruction. In other words, if we aim to obtain a good estimation of the epidemic risk, we have to increase the value of p or, in the case of an empirical survey, increase the number of contacts reported. Other reconstruction strategies should also be sought. Future work will focus on designing and evaluating new methods of reconstruction to estimate the epidemic risk from incompletely and non-uniformly sampled data.

Chapter 5

Conclusion

Measuring human interactions and in particular face-to-face contacts between individuals has been a challenge of crucial importance in scientific domains such as social sciences or epidemiology. As such, various methods of data collection have been developed in the last decades by the research community [2, 3]. Traditional methods consist mostly in the use of surveys or diaries while the emergence of new technologies have made possible the measurement of contacts with high precision. Each of these methods has advantages and limitations.

On the one hand, methods based on surveys can help to gather various types of information. Indeed, well-studied questionnaires can ask not only on the existence of contacts but also an estimated duration of contacts as well as the context of such contacts (e.g., home, work, travel). In some contexts where it is requested (e.g., when studying the propagation of sexually transmitted diseases), questionnaires can also ask whether contacts involved physical contact or not. Moreover, surveys can collect information about other types of relationships such as friendship which could lead to actual contacts. Surveys however have limitations [4, 5]. First of all, they are costly and it is difficult to recruit participants as individuals can be discouraged by the burden of filling questionnaires. Most importantly, the self-reporting character of these methods yields different types of biases that are difficult to estimate. In retrospective diaries, participants might not recall all their contacts, especially the shortest ones, and make incorrect estimations of contact durations. In surveys asking about friendships for instance, personal feelings and perceptions can lead to other biases if the questionnaires are not precise enough.

On the other hand, methods based on wearable sensors allow to measure face-to-face contacts in an objective way and thus avoid the biases inherent in self-reporting methods. Sensors are able to detect contacts, even very short ones, with a high spatio-temporal resolution and gives access to temporal networks that evolve over time. Moreover the decrease in the related costs makes nowadays large-scale deployments feasible. The main limitation of such procedures comes from the fact that they do not register contacts with individuals not participating to the deployment i.e., are limited to the study of contacts within

a closed population. Sampling issues can also arise if not all the members of the target population agree to wear the sensors.

In this thesis, we have analyzed a data set combining data about face-to-face contacts detected with sensors (collected three years in a row), contact diaries, friendship relations and Facebook relationships. The analysis of the contact network obtained from sensors has given the following conclusions. First, the distributions of typical durations are heterogeneous and the distribution of degrees is narrow, in agreement with results obtained in other environments. The network was highly structured in classes and no gender homophily was observed contrary to the case of a primary school [33]. The longitudinal study at two different timescales has shown the robustness of contact patterns.

The comparison of data collected with different methods in the same context has been performed, to our knowledge, in only two other studies [22, 31]. The comparison we made on the combined data set has yielded the following results. The contact diaries suffer from low participation, moreover most short contacts were not reported and we have observed an overestimation of contact durations reported by participants. However, the longest contacts were all reported and the structure in classes observed in the network obtained from sensors was preserved in the network obtained from diaries. In the case of the friendship survey, the participation rate was also very low. Most shorts contacts do not correspond to friendships while the longest contacts all correspond to friendships and most friendships lead to actual encounters. As for contact diaries, the structure in classes was preserved in the friendship network. The two networks obtained with these methods of data collection are more dilute than the sensor network and thus are considered as incomplete data sets. Finally, the comparison of Facebook data with networks obtained from sensors and friendship survey lead to the conclusion that the existence of a Facebook link does not give any information on the existence of face-to-face contacts or friendships.

Then, we have investigated the use of the incomplete data sets in the simulations of epidemic spreading processes. The simulations yield strong underestimations of the epidemic risk with respect to the results obtained with the contact network of sensors. In order to understand if this underestimation can be seen as biases resulting from a sampling process performed on the contact network, we have designed a non-uniform method of sampling created to mimic the friendship survey procedure. This could indeed give hints on the possibilities of using incomplete data to obtain an accurate estimation of the epidemic risk [43]. This method of sampling has given results in very good agreement with the results obtained when the friendship network is used. The interest of our method of sampling is that it depends only on two parameters which makes its utilisation in various contexts. We have started to tackle this perspective using the sampling procedure on other data sets, however the comparison of networks sampled from the contact network with networks obtained through other methods of data collection was impossible given the current lack of combined data sets. The next step was the reconstruction of a surrogate contact network from sampled networks, in order to use it in simulations of epidemic spreading, and the comparison of the results with the outcomes of such simulations

obtained with the actual contact network. The first attempts of reconstruction using intuitive strategies such as the one used in [43] have failed to obtain a good estimate of the epidemic risk as it is still much smaller with respect to the epidemic risk measured with the contact network, mainly because of the non-uniform sampling of edges. We should not forget that most data sets are in fact incomplete samples of the network of interest. The design of more refined reconstruction strategies would lead on a mid-term objective to estimate the epidemic risk from incomplete data. Finally, even if the precise epidemic risk is not well estimated, a complementary research direction will be to investigate if it is still possible to evaluate mitigations strategies using incomplete data or reconstructed surrogate data [43].

Appendices

Appendix A

Introduction

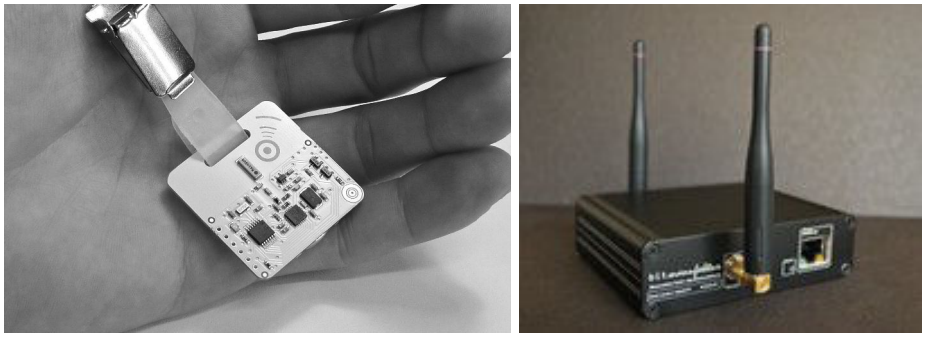


Figure A.1: RFID sensors and antennas used for deployments. Photo credit: SocioPatterns

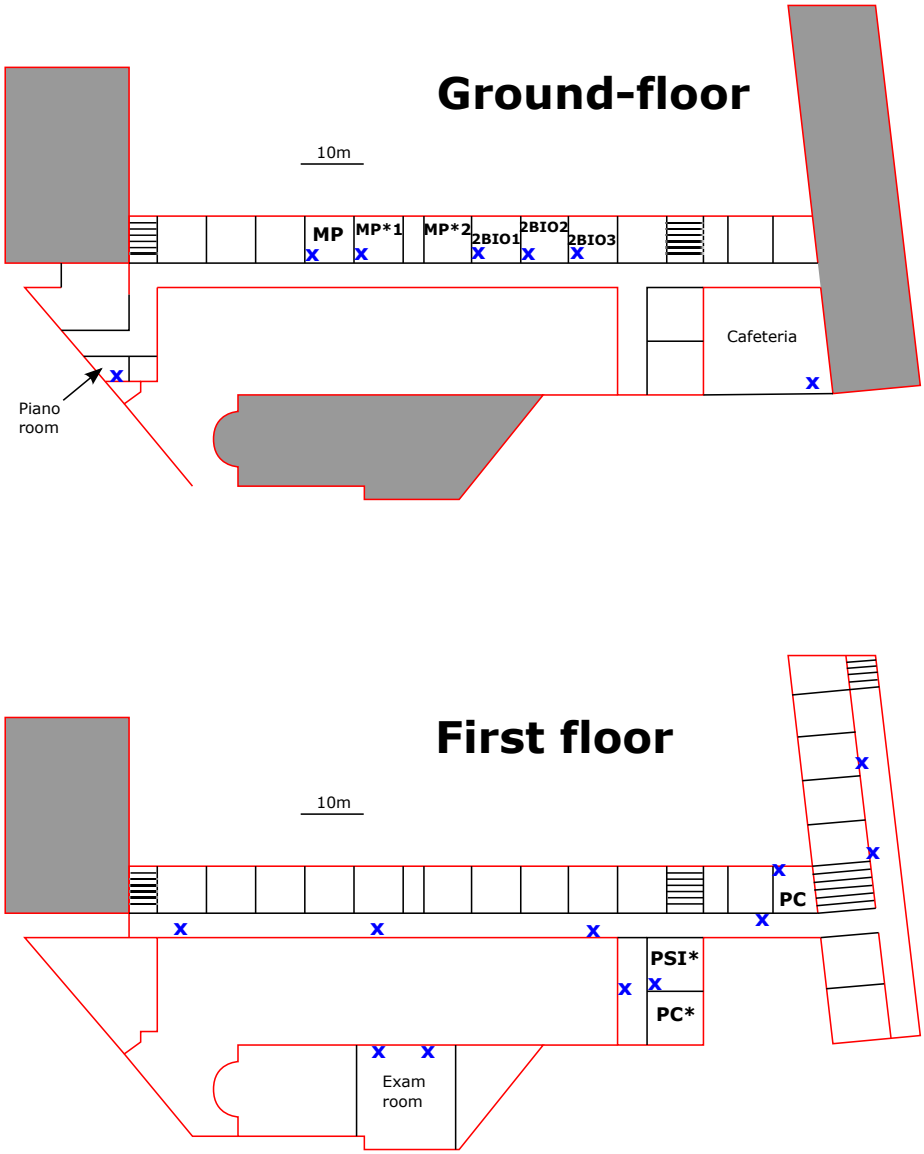


Figure A.2: Locations of antennas (blue crosses) used in the deployment in Lycée Thiers in 2013. Each class has most of its classes in one specific classroom (written in black).

Appendix B

Analysis of face-to-face proximity data

	Number of contacts	Cumulative duration of contacts			N	E
Day	Number (% of total)	Seconds (% of total)	Minutes	Hours		
Tuesday	4,234 (39.7)	211,980 (37.1)	3,533	59	121	975
Wednesday	1,826(17.1)	93,120 (16.3)	1,552	26	113	582
Thursday	2,477 (23.2)	146,520 (25.7)	2,442	41	115	597
Friday	2,140 (20.0)	119,060 (20.9)	1,984	33	112	609
Total	10,677	570,680	9,511	159	126	1710

Table B.1: Number and duration of contacts in the different days of the 2011 data collection that lasted 4 days.

	Number of contacts	Cumulative duration of contacts			N	E
Day	Number (% of total)	Seconds (% of total)	Minutes	Hours		
1st Monday	4,191 (21.2)	199,140 (22.1)	3,319	55	156	758
1st Tuesday	3,170 (16.0)	132,720 (14.7)	2,212	37	158	664
Wednesday	1,547 (7.8)	64,540 (6.4)	965	16	145	486
Thursday	2,641 (13.4)	106,920 (11.9)	1,782	30	146	550
Friday	3,184 (16.1)	154,360 (17.1)	2,573	43	151	659
2nd Monday	2,988 (15.1)	156,360 (17.4)	2,606	43	153	566
2nd Tuesday	2,053 (10.4)	93,540 (10.4)	1,559	26	151	483
Total	19,774	900,940	15,016	250	180	2220

Table B.2: Number and duration of contacts in the different days of the 2012 data collection that lasted 7 days.

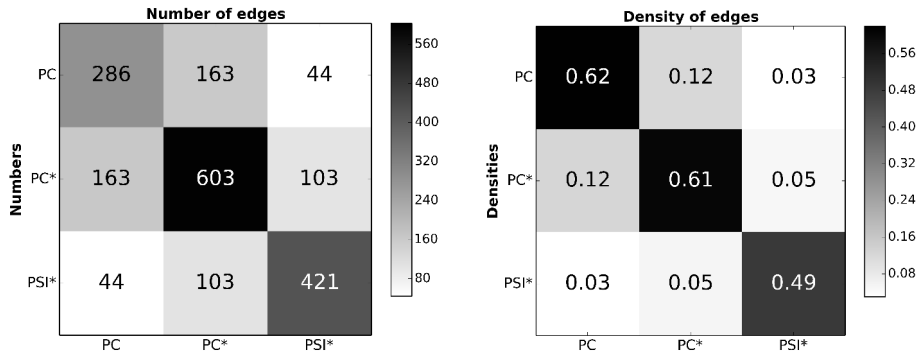


Figure B.1: 2011 dataset: Contact matrices of edge numbers and densities. Left: the matrix entry at row X and column Y gives E_{XY}, i.e., the number of pairs of individuals of classes X and Y who have been in contact at least once during the study. Right: the matrix entry at row X and column Y gives ρ_{XY}, i.e., E_{XY} normalized by the maximal possible number of pairs of individuals of classes X and Y.

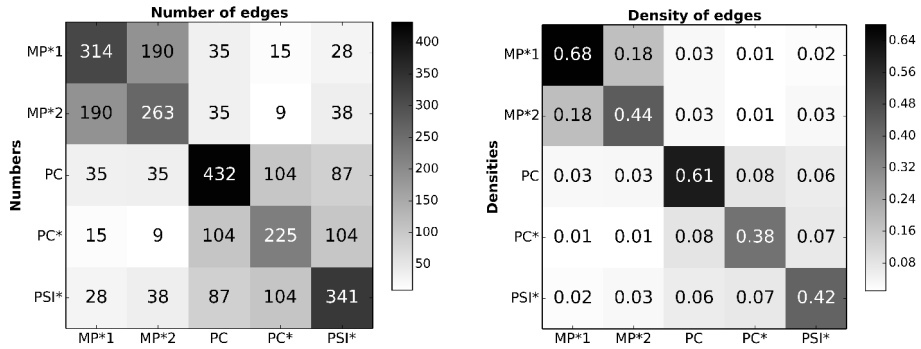


Figure B.2: 2011 dataset: Contact matrices of edge numbers and densities. Left: the matrix entry at row X and column Y gives E_{XY}, i.e., the number of pairs of individuals of classes X and Y who have been in contact at least once during the study. Right: the matrix entry at row X and column Y gives ρ_{XY}, i.e., E_{XY} normalized by the maximal possible number of pairs of individuals of classes X and Y.

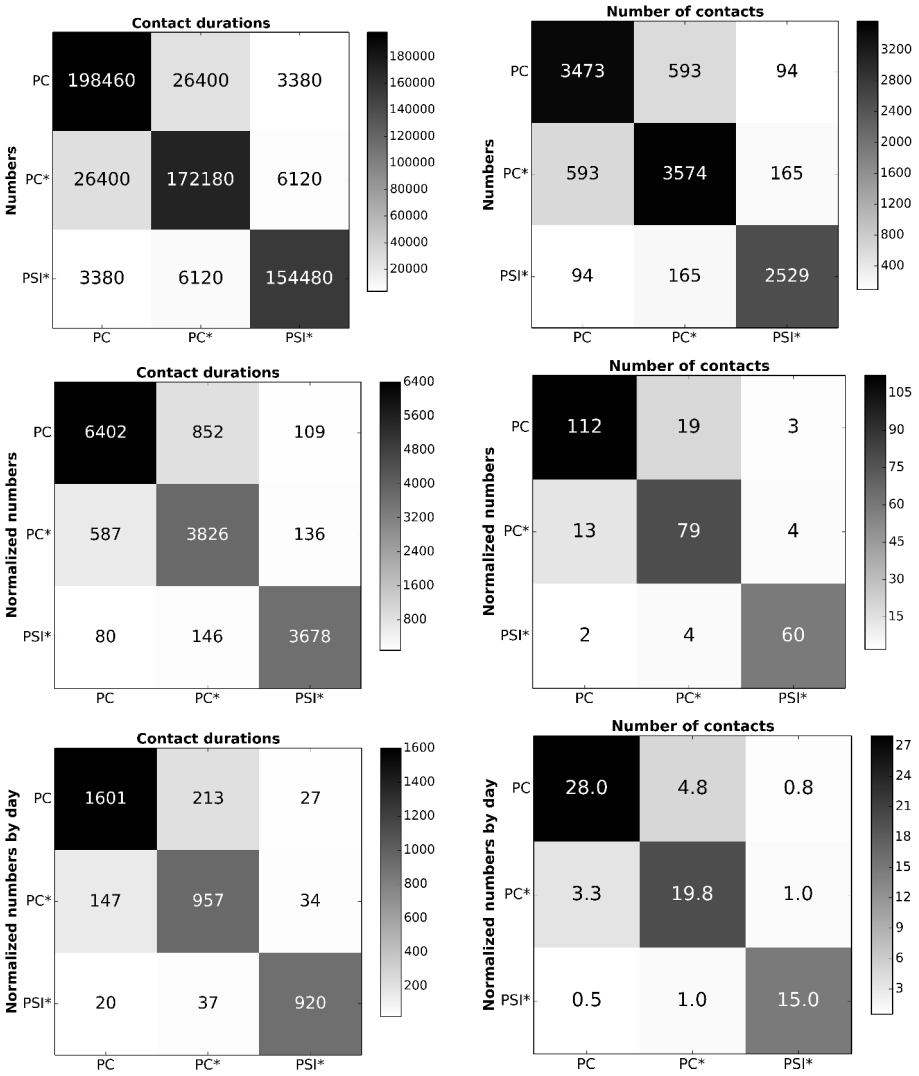


Figure B.3: 2011 dataset: Contact matrices giving the cumulated durations in seconds (first column) and the numbers (second column) of contacts between classes during the whole study. In the first row, the matrix entry at row X and column Y gives the total duration (resp. number) of all contacts between all individuals of class X with all individuals of class Y. In the second row, the matrix entry at row X and column Y gives the average duration (resp. number) of contacts of an individual of class X with all individuals of class Y. In the third row, we normalize each matrix element of the second column matrices by the duration of the study, in days, to obtain at row X and column Y the average daily duration (resp. number) of contacts of an individual of class X with individuals of class Y.

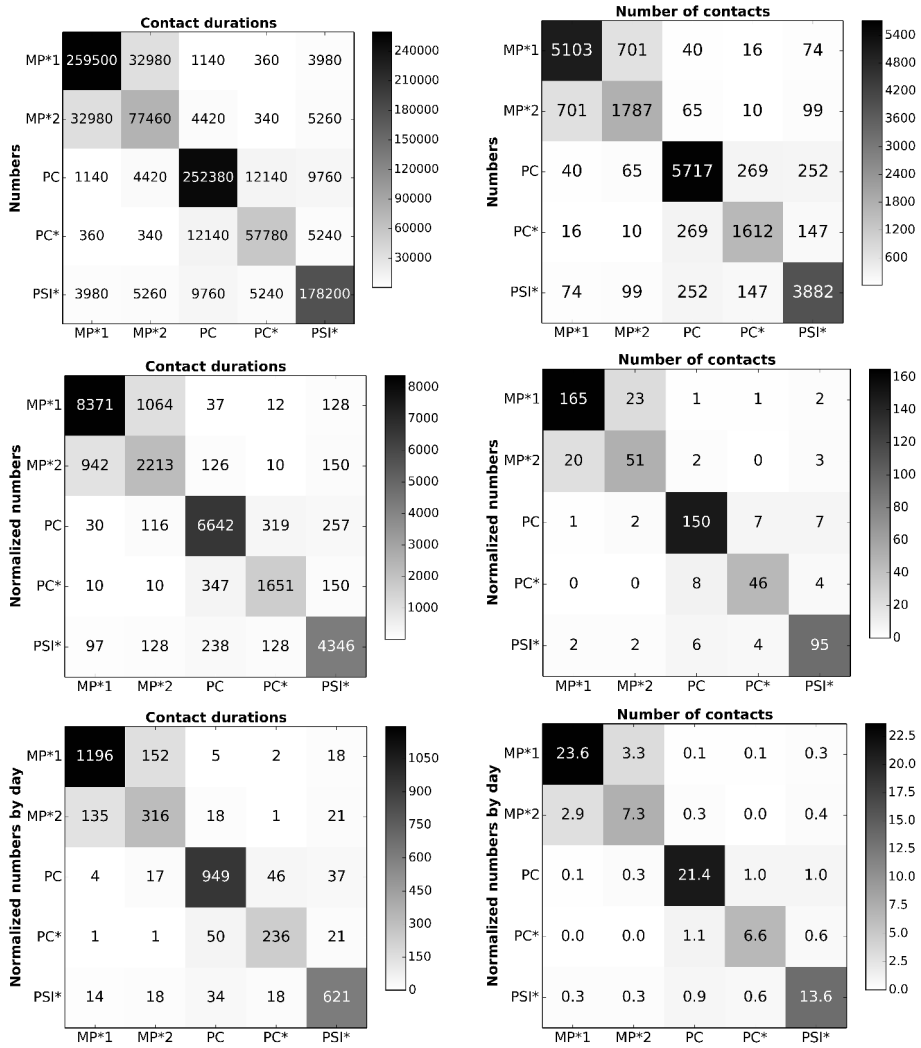
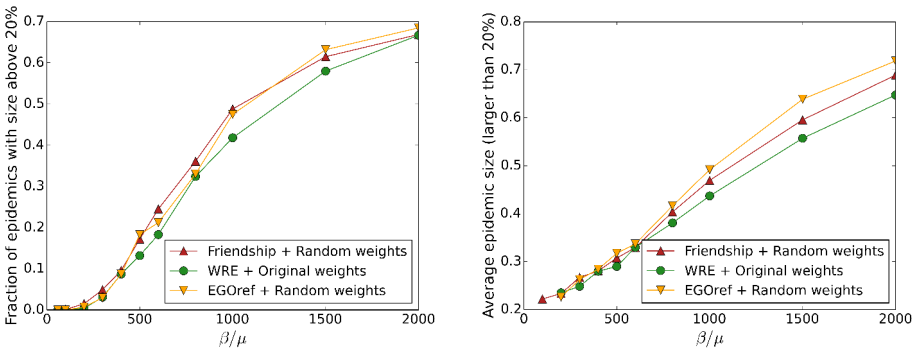


Figure B.4: 2012 dataset: Contact matrices giving the cumulated durations in seconds (first column) and the numbers (second column) of contacts between classes during the whole study. In the first row, the matrix entry at row X and column Y gives the total duration (resp. number) of all contacts between all individuals of class X with all individuals of class Y. In the second row, the matrix entry at row X and column Y gives the average duration (resp. number) of contacts of an individual of class X with all individuals of class Y. In the third row, we normalize each matrix element of the second column matrices by the duration of the study, in days, to obtain at row X and column Y the average daily duration (resp. number) of contacts of an individual of class X with individuals of class Y.

Appendix C

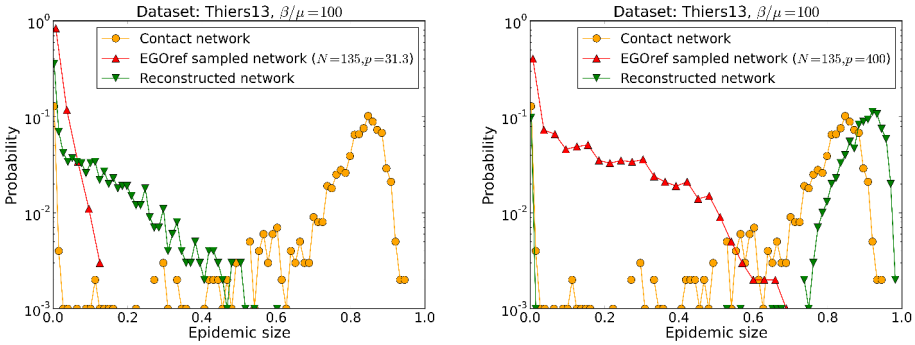
Equivalence between friendship network and a non-uniform sampling of contact network



(a) (b)

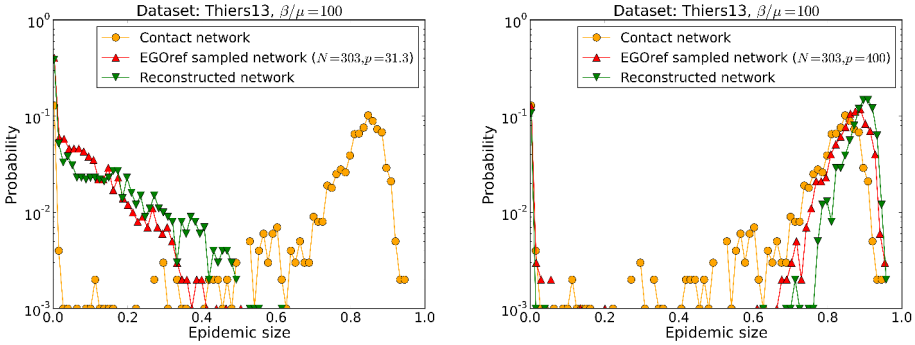
Figure C.1: Outcome of SIR spreading simulations performed on the WRE sampled networks with "Original weights" and on friendship and EGOref sampled networks with "Random weights". (a) Fraction of epidemics with size above 20% (at least 20% of recovered individuals at the end of the SIR process) as a function of β/μ. (b) Average size of epidemic with size above 20% as a function of β/μ.

Here we describe the reconstruction procedure. This method is applied to the EGOref sampled networks and preserves the density of the sampled networks. We assume that we know the total number of individuals in the target population (i.e., number of nodes in the contact network) and their potential repartition in groups (classes, office departments...). From the EGOref sampled networks, we build the matrix of link densities between different groups. Then we add in the sampled networks the number of nodes missing in each group. Finally, we add edges randomly but preserving the matrix of link densities measured before the addition of nodes. We show in Figure C.2 the results of simulations obtained when using the contact network, the EGOref sampled networks and the reconstructed network for the dataset Thiers13. One can see that changing the value of N in the EGOref sampling procedure does not have a strong impact on the results obtained with the reconstructed network. On the contrary, the value of p has a strong influence on the final results. On the one hand, when p is large, the EGOref sampling method leads to sampled networks with rather large density and the results obtained after reconstruction are in good agreement with the results obtained with the empirical contact network. In fact, when p is large enough the EGOref method of sampling is quite equivalent to the RN method of sampling for which we know that the reconstruction of a surrogate contact network yields very good results with respect to the original contact network [43]. On the other hand, when p is small, the results are better after reconstruction: the discrepancy between sampled and contact networks is reduced after reconstruction but the epidemic risk is still significantly underestimated given the very small density.



(a) $N = 135,\ p = 31.3$

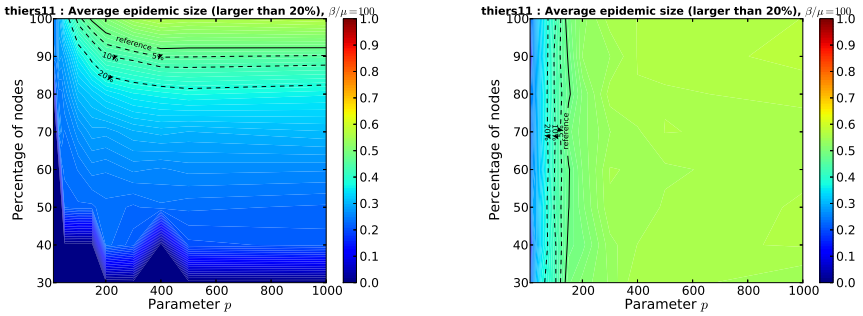
(b) $N = 135,\ p = 400$



(c) $N = 303,\ p = 31.3$

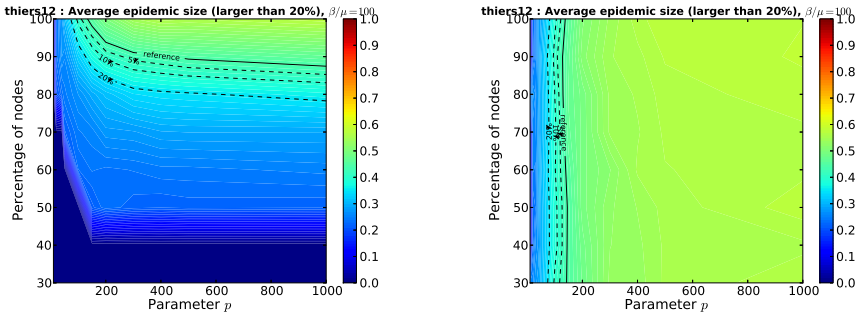
(d) $N = 303,\ p = 400$

Figure C.2: Thiers13 dataset: Distributions of epidemic sizes for the empirical contact network, EGOref sampled networks and reconstructed networks. The distributions are shown for two different values of the parameter p and two values of N for the EGOref sampling method.



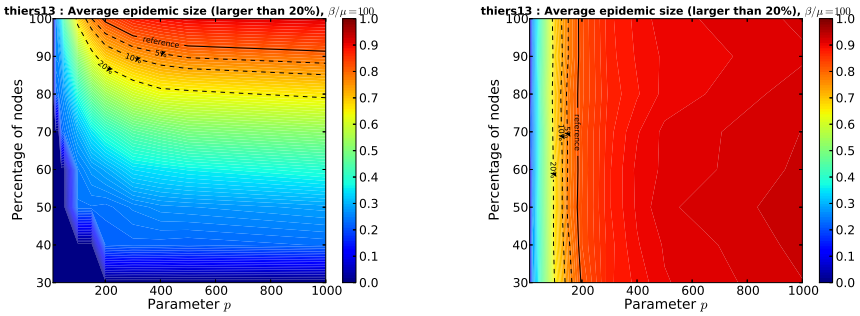
(a) (b)

Figure C.3: Outcome of SIR spreading simulations performed on networks sampled with the EGOref method, before and after reconstruction (Lycée Thiers 2011 dataset).



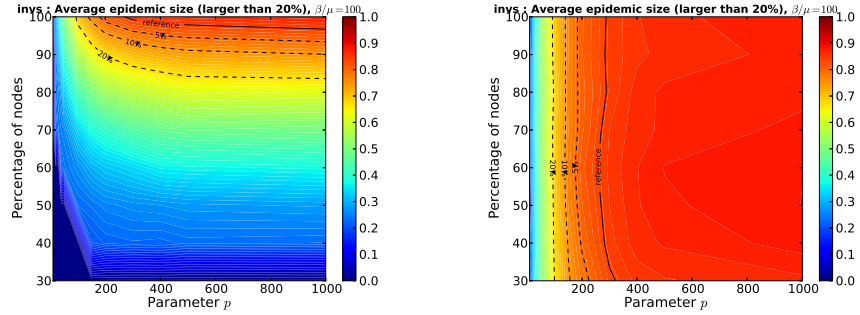
(a) (b)

Figure C.4: Outcome of SIR spreading simulations performed on networks sampled with the EGOref method, before and after reconstruction (Lycée Thiers 2012 dataset).



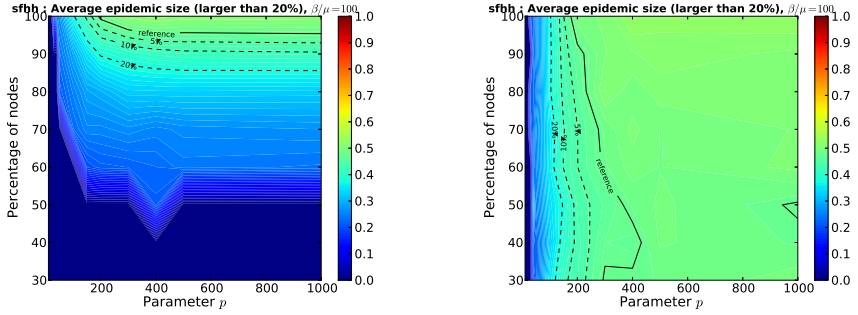
(a) (b)

Figure C.5: Outcome of SIR spreading simulations performed on networks sampled with the EGOref method, before and after reconstruction (Lycée Thiers 2013 dataset).



(a) (b)

Figure C.6: Outcome of SIR spreading simulations performed on networks sampled with the EGOref method, before and after reconstruction (InVS dataset).



(a) (b)

Figure C.7: Outcome of SIR spreading simulations performed on networks sampled with the EGOref method, before and after reconstruction (SFHH dataset).

List of publications

- J. Fournet and A. Barrat. *Contact patterns among high school students*, PLoS ONE, 9(9):e107878, 2014.
- M. Génois., C. L. Vestergaard, J. Fournet, A. Panisson, I. Bonmarin, and A. Barrat. *Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers*, Network Science, 3:326–347, 2015.
- R. Mastrandrea, J. Fournet and A. Barrat. *Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys*, PLoS ONE, 10(9):e136497, 2015.
- J. Fournet and A. Barrat. *Epidemic risk from friendship network data: an equivalence with a non-uniform sampling of contact networks*, Scientific Reports, 6:24593, 2016.

Bibliography

[1] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424:175–308, February 2006.

[2] A Barrat and C Cattuto. *Social Phenomena (eds B Gonçalves and N Perra)*, chapter 3, pages 37–57. 2015.

[3] K Eames, S Bansal, S Frost, and S Riley. Six challenges in measuring contact networks for use in modelling. *Epidemics*, 10:72–77, 2015.

[4] HR Bernard, P Killworth, D Kronenfeld, and L Sailer. The problem of informant accuracy: The validity of retrospective data. *Annual review of anthropology*, pages 495–517, 1984.

[5] DL Paulhus and S Vazire. The self-report method. *Handbook of research methods in personality psychology*, pages 224–239, 2007.

[6] The sociopatterns website. `http://www.sociopatterns.org`.

[7] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE*, 5(7):1–9, 07 2010.

[8] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Vittoria Colizza, Lorenzo Isella, Corinne Régis, Jean-François Pinton, Nagham Khanafer, Wouter Van den Broeck, and Philippe Vanhems. Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees. *BMC Medicine*, 9(1):1–15, 2011.

[9] Lorenzo Isella, Mariateresa Romano, Alain Barrat, Ciro Cattuto, Vittoria Colizza, Wouter Van den Broeck, Francesco Gesualdo, Elisabetta Pandolfi, Lucilla Ravà, Caterina Rizzo, and Alberto Eugenio Tozzi. Close encounters in a pediatric ward: Measuring face-to-face proximity and mixing patterns with wearable sensors. *PLoS ONE*, 6(2):1–10, 02 2011.

[10] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne

Régis, Bruno Lina, and Philippe Vanhems. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE*, 6(8):1–13, 08 2011.

[11] Lorenzo Isella, Juliette Stehlé, Alain Barrat, Ciro Cattuto, Jean-François Pinton, and Wouter Van den Broeck. What's in a crowd? analysis of face-to-face behavioral networks. *Journal of Theoretical Biology*, 271(1):166–180, 2011.

[12] Philippe Vanhems, Alain Barrat, Ciro Cattuto, Jean-François Pinton, Nagham Khanafer, Corinne Régis, Byeul-a Kim, Brigitte Comte, and Nicolas Voirin. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS ONE*, 8(9):e73970, 09 2013.

[13] T. Smieszek, V.C. Barclay, I. Seeni, J.J. Rainey, H. Gao, A. Uzicanin, and M. Salathé. How should social mixing be measured: comparing web-based survey and sensor-based methods. *BMC Infectious Diseases*, 14(1):1–13, 2014.

[14] R. Mastrandrea, J. Fournet, and A. Barrat. Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLoS ONE*, 10(9):1–26, 09 2015.

[15] J.M. Read, W.J. Edmunds, S. Riley, J. Lessler, and D.A.T. Cummings. Close encounters of the infectious kind: methods to measure social mixing behaviour. *Epidemiol Infect.*, 140:2117–2130, 12 2012.

[16] W. J. Edmunds, C. J. O'callaghan, and D. J. Nokes. Who mixes with whom? a method to determine the contact patterns of adults that may lead to the spread of airborne infections. *Proceedings of the Royal Society of London B: Biological Sciences*, 264(1384):949–957, 1997.

[17] Joël Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, Janneke Heijne, Malgorzata Sadkowska-Todys, Magdalena Rosinska, and W. John Edmunds. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med*, 5(3):1–1, 03 2008.

[18] Jonathan M Read, Ken T.D Eames, and W. John Edmunds. Dynamic social networks and the implications for the spread of infectious disease. *Journal of The Royal Society Interface*, 5(26):1001–1007, 2008.

[19] Emilio Zagheni, Francesco C. Billari, Piero Manfredi, Alessia Melegaro, Joël Mossong, and WJ Edmunds. Using time-use data to parameterize models for the spread of close-contact infectious diseases. *Am J Epidemiol*, 168(9):1082–1090, 2008.

[20] R.T. Mikolajczyk, M.K. Akmatov, S. Rastin, and M. Kretzschmar. Social contacts of school children and the transmission of respiratory-spread pathogens. *Epidemiology Infect.*, 136:813–822, 6 2008.

[21] A. J. K. Conlan, K. T. D. Eames, J. A. Gage, J. C. von Kirchbach, J. V. Ross, R. A. Saenz, and J. R. Gog. Measuring social networks in british primary schools through scientific engagement. *Proceedings of the Royal Society of London B: Biological Sciences*, 278(1711):1467–1475, 2011.

[22] T. Smieszek, E.U. Burri, R. Scherzinger, and R.W. Scholz. Collecting close-contact social mixing data with contact diaries: reporting errors and biases. *Epidemiol Infect.*, 140:744–752, 04 2012.

[23] Gail E. Potter, Mark S. Handcock, Ira M. Longini, and M. Elizabeth Halloran. Estimating within-school contact networks to understand influenza transmission. *Ann. Appl. Stat.*, 6(1):1–26, 03 2012.

[24] Leon Danon, Jonathan M. Read, Thomas A. House, Matthew C. Vernon, and Matt J. Keeling. Social encounter networks: characterizing great britain. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1765), 2013.

[25] A. Pentland. Honest signals: how they shape our world. *Cambridge, MA: MIT Press*, 2008.

[26] Marcel Salathé, Maria Kazandjieva, Jung Woo Lee, Philip Levis, Marcus W. Feldman, and James H. Jones. A high-resolution human contact network for infectious disease transmission. *PNAS*, 107(51), 2010.

[27] Thomas Hornbeck, David Naylor, Alberto M. Segre, Geb Thomas, Ted Herman, and Philip M. Polgreen. Using sensor networks to study the effect of peripatetic healthcare workers on the spread of hospital-associated infections. *The Journal of Infectious Diseases*, 206(10):1549–1557, 2012.

[28] Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. Measuring large-scale social networks with high resolution. *PLoS ONE*, 9(4):1–24, 04 2014.

[29] Victoria C. Barclay, Timo Smieszek, Jianping He, Guohong Cao, Jeanette J. Rainey, Hongjiang Gao, Amra Uzicanin, and Marcel Salathé. Positive network assortativity of influenza vaccination at a high school: Implications for outbreak risk and herd immunity. *PLoS ONE*, 9(2):1–11, 02 2014.

[30] Hasan Guclu, Jonathan Read, Charles J. Vukotich, Jr, David D. Galloway, Hongjiang Gao, Jeanette J. Rainey, Amra Uzicanin, Shanta M. Zimmer, and Derek A. T. Cummings. Social Contact Networks and Mixing among Students in K-12 Schools in Pittsburgh, PA. *PLoS ONE*, 11(3):1–19, 03 2016.

[31] Molly Leecaster, Damon J. A. Toth, Warren B. P. Pettey, Jeanette J. Rainey, Hongjiang Gao, Amra Uzicanin, and Matthew Samore. Estimates of social contact in a middle school based on self-report and wireless sensor data. *PLoS ONE*, 11(4):1–21, 04 2016.

[32] M. Génois, C. L. Vestergaard, J. Fournet, A. Panisson, I. Bonmarin, and A. Barrat. Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. *Network Science*, 3:326–347, 9 2015.

[33] Juliette Stehlé, François Charbonnier, Tristan Picard, Ciro Cattuto, and Alain Barrat. Gender homophily from spatial behavior in a primary school: A sociometric study. *Social Networks*, 35(4):604 – 613, 2013.

[34] J. Fournet and A. Barrat. Contact patterns among high school students. *PLoS ONE*, 9(9):1–17, 09 2014.

[35] Mark Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[36] Albert-László Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, 2005.

[37] Albert-László Barabási. Bursts: The hidden pattern behind everything we do. *Dutton Adult*, 2010.

[38] Fabrizio Iozzi, Francesco Trusiano, Matteo Chinazzi, Francesco C. Billari, Emilio Zagheni, Stefano Merler, Marco Ajelli, Emanuele Del Fava, and Piero Manfredi. Little italy: An agent-based approach to the estimation of contact patterns- fitting predicted matrices to serological data. *PLoS Comput Biol*, 6(12):1–10, 12 2010.

[39] M McPherson, L Smith-Lovin, and JM Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[40] Sally Blower and Myong-Hyun Go. The importance of including dynamic social networks when modeling epidemics of airborne infections: does increasing complexity increase accuracy? *BMC Medicine*, 9(1):1–3, 2011.

[41] Anna Machens, Francesco Gesualdo, Caterina Rizzo, Alberto E. Tozzi, Alain Barrat, and Ciro Cattuto. An infectious disease model on empirical networks of human contact: bridging the gap between dynamic network data and contact matrices. *BMC Infectious Diseases*, 13(1):1–15, 2013.

[42] Sergei Maslov, Kim Sneppen, and Alexei Zaliznyak. Detection of topological patterns in complex networks: correlation profile of the internet. *Physica A*, 333:529–540, 2004.

[43] Mathieu Génois, Christian Vestergaard, Ciro Cattuto, and Alain Barrat. Compensating for population sampling in simulations of epidemic spread on temporal contact networks. *Nature Communications*, 8:8860, 2015.

[44] Fred Collopy. Biases in retrospective self-reports of time use: An empirical study of computer users. *Management Science*, 42(5):758–767, 1996.

[45] GP Hyett. Validation of diary records of telephone calling behavior the recall method in social surveys. *Univ. of London Institute of Education*, pages 136–138, 1979.

[46] Mark Granovetter. Network sampling: Some first steps. *American Journal of Sociology*, 81(6):1287–1303, 1976.

[47] O Frank. Sampling and estimation in large social networks. *Social Networks*, 1(91), 1979.

[48] D Achlioptas, A Clauset, D Kempe, and C Moore. On the bias of traceroute sampling: Or, power-law degree distributions in regular graphs . In *Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing*, 2005.

[49] Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong. Statistical properties of sampled networks. *Phys. Rev. E*, 73:016102, Jan 2006.

[50] Gueorgi Kossinets. Effects of missing data in social networks. *Social Networks*, 28(3):247–268, 2006.

[51] Jukka-Pekka Onnela and Nicholas A. Christakis. Spreading paths in partially observed social networks. *Phys. Rev. E*, 85:036106, Mar 2012.

[52] Neli Blagus, Lovro Subelj, and Marko Bajec. Empirical comparison of network sampling techniques. *Preprint arxiv*, abs/1506.02449, 2015.

[53] A. C. Ghani, C. A. Donnelly, and G. P. Garnett. Sampling biases and missing data in explorations of sexual partner networks for the spread of sexually transmitted diseases. *Statistics in Medicine*, 17(18):2079–2097, 1998.

[54] A. C. Ghani and G. P. Garnett. Measuring sexual partner networks for transmission of sexually transmitted diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 161(2):227–238, 1998.

[55] Georgiy Bobashev, Robert J. Morris, and D. Michael Goedecke. Sampling for global epidemic models and the topology of an international airport network. *PLoS ONE*, 3(9):1–8, 09 2008.

[56] J Leskovec and C Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006.

[57] Fabien Viger, Alain Barrat, Luca Dall'Asta, Cun-Hui Zhang, and Eric D. Kolaczyk. What is the real size of a sampled network? the case of the internet. *Phys. Rev. E*, 75:056111, May 2007.

[58] Catherine A. Bliss, Christopher M. Danforth, and Peter Sheridan Dodds. Estimation of global network statistics from incomplete data. *PLoS ONE*, 9(10):1–18, 10 2014.

[59] Yaonan Zhang, Eric D. Kolaczyk, and Bruce D. Spencer. Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks. *Ann. Appl. Stat.*, 9(1):166–199, 03 2015.

[60] J Fournet and A Barrat. Epidemic risk from friendship network data: an equivalence with a non-uniform sampling of contact networks. *Scientific Reports*, 6(24593), 2016.

[61] Timo Smieszek, Lena Fiebig, and Roland W. Scholz. Models of epidemics: when contact repetition and clustering should be included. *Theoretical Biology and Medical Modelling*, 6(1):1–15, 2009.

[62] A. Barrat, M. Barthélemy, and A Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[63] Luis E. C. Rocha, Anna E. Thorson, Renaud Lambiotte, and Fredrik Liljeros. Respondent-driven sampling bias induced by community structure and response rates in social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, pages n/a–n/a, 2016.

[64] IM Jr. Longini, JS Koopman, AS Monto, and JP Fox. Estimating household and community transmission parameters for influenza. *Am J Epidemiol*, 115(5):736–751, 05 1982.

[65] Cécile Viboud, Pierre-Yves Boëlle, Simon Cauchemez, Audrey Lavenu, Alain-Jacques Valleron, Antoine Flahault, and Fabrice Carrat. Risk factors of influenza transmission in households. *International Congress Series*, 1263:291 – 294, 2004. Options for the Control of Influenza V. Proceedings of the International Conference on Options for the Control of Influenza V.

[66] J. Stehlé, A. Barrat, and G. Bianconi. Dynamical and bursty interactions in social networks. *Physical Review E*, 81 : 035101 (R), 2010.

[67] K. Zhao, J. Stehlé, A. Barrat, and G. Bianconi. Social network dynamics of face-to-face interactions. *Physical Review E*, 83 : 056109 (R), 2011.

[68] Michele Starnini, Andrea Baronchelli, and Romualdo Pastor-Satorras. Modeling human dynamics of face-to-face interaction networks. *Phys. Rev. Lett.*, 110:168701, Apr 2013.

[69] Márton Karsai, Nicola Perra, and Alessandro Vespignani. Time varying networks and the weakness of strong ties. *Scientific Reports*, 4(4001), 2014.

[70] Valerio Gemmetto, Alain Barrat, and Ciro Cattuto. Mitigation of infectious disease at school: targeted class closure vs school closure. *BMC infectious diseases*, 14(1):695, December 2014.

[71] Laura Fumanelli, Marco Ajelli, Piero Manfredi, Alessandro Vespignani, and Stefano Merler. Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. *PLoS Comput Biol*, 8(9):1–10, 09 2012.

[72] Alessia Melegaro, Mark Jit, Nigel Gay, Emilio Zagheni, and W. John Edmunds. What types of contacts are important for the spread of infections? using contact survey data to explore european mixing patterns. *Epidemics*, 3:143–151, 2011.

[73] Nele Goeyvaerts, Niel Hens, Benson Ogunjimi, Marc Aerts, Ziv Shkedy, Pierre Van Damme, and Philippe Beutels. Estimating infectious disease parameters from data on social contacts and serological status. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):255–277, 2010.

[74] H. Bernard, R. Fischer, R.T. Mikolajczyk, M. Kretzschmar, and M. Wildner. Nurses' contacts and potential for infectious disease transmission. *Emerging Infectious Disease*, 15(9):1438–1444, 2009.

[75] Leon Danon, Thomas A. House, Jonathan M. Read, and Matt J. Keeling. Social encounter networks: collective properties and disease transmission. *Journal of The Royal Society Interface*, 9(76):2826–2833, 2012.

[76] Stephen Eubank, Hasan Guclu, V. S. Anil Kumar, Madhav V. Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429:180–184, 2004.

[77] Damon J. A. Toth, Molly Leecaster, Warren B. P. Pettey, Adi V. Gundlapalli, Hongjiang Gao, Jeanette J. Rainey, Amra Uzicanin, and Matthew H. Samore. The role of heterogeneity in contact timing and duration in network models of influenza spread in schools. *Journal of The Royal Society Interface*, 12(108), 2015.

[78] Kim Van Kerckhove, Niel Hens, W.J. Edmunds, and Ken T.D. Eames. The impact of illness on social networks: Implications for transmission and control of influenza. *Am J Epidemiol*, 178(11):1655–1662, 10 2013.

[79] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. *Science*, 323(5915):721–723, 2009.

[80] Alain Barrat, Ciro Cattuto, Martin Szomszor, Wouter Van den Broeck, and Harith Alani. Social dynamics in conferences: Analysis of data from the live social semantics application. In *Proceedings of the 9th International Semantic Web Conference (ISWC '10)*, 2010.

[81] A. Barrat, C. Cattuto, V. Colizza, F. Gesualdo, L. Isella, E. Pandolfi, J.-F. Pinton, L. Rava, C. Rizzo, M. Romano, J. Stehlé, A.E. Tozzi, and W. Broeck. Empirical temporal networks of face-to-face human interactions. *The European Physical Journal Special Topics*, 222(6):1295–1309, 2013.

[82] A. Barrat, C. Cattuto, A. E. Tozzi, P. Vanhems, and N. Voirin. Measuring contact patterns with wearable sensors: methods, data characteristics and applications to data-driven simulations of infectious diseases. *Clinical Microbiology and Infection*, 20(1):10–16, 2014.

[83] F Battiston, V Nicosia, and V Latora. Structural measures for multiplex networks. 2013.

[84] Sonja Filiposka, Andrej Gajduk, Tamara Dimitrova, and Ljupco Kocarev. Bridging online and offline social networks: Multiplex analysis. *CoRR*, abs/1605.01901, 2016.

[85] Lorenzo Coviello, Massimo Franceshetti, Manuel Garcia-Herranz, and Iyad Rahwan. Limits of friendship networks in predicting epidemic risk. *Preprint arxiv*, abs/1509.08368, 2015.

[86] Rossana Mastrandrea and Alain Barrat. How to estimate epidemic risk from incomplete contact diaries data? *Plos Computational Biology*, 2016.