

Human proximity networks: analysis, modeling and dynamical phenomena

Juliette Stehlé

▶ **To cite this version:**

Juliette Stehlé. Human proximity networks: analysis, modeling and dynamical phenomena. Physics and Society [physics.soc-ph]. Aix-Marseille Université, 2012. English. NNT : . tel-00777540

HAL Id: tel-00777540
https://theses.hal.science/tel-00777540

Submitted on 17 Jan 2013

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

École doctorale ED352
Physique et Sciences de la matière

Thèse de Doctorat présentée par

Juliette STEHLÉ

pour obtenir le grade de

Docteur d'Aix-Marseille Université

Specialité : PHYSIQUE THÉORIQUE ET MATHÉMATIQUES

Soutenue le 17 décembre 2012

Réseaux de proximité humaine : analyse, modélisation et processus dynamiques

Directeur de thèse : Alain BARRAT

Thèse préparée au Centre de Physique Théorique

Jury :

Rapporteurs :	Jean-Pierre NADAL	- ENS & EHESS
	Alessandro VESPIGNANI	- Northeastern University
Directeur de thèse :	Alain BARRAT	- CPT (Aix-Marseille Université)
Membres du jury :	Juliette ROUCHIER	- GREQAM (Aix-Marseille Université)
	Ciro CATTUTO	- ISI Foundation
	Eric GAUTIER	- CREST (GENES)
	Frédéric van WIJLAND	- MSC (Université Paris Diderot)

Remerciements

Nous y voilà, le doctorat arrive à son terme, les rapports ont été reçus et la soutenance approche à grands pas. La route n'a pourtant pas été facile, ni très linéaire. Partiellement du fait des aléas de la thèse, des résultats inattendus mais observés – et réciproquement, des résultats attendus mais inobservés – rendant certaines journées radieuses et d'autres plus moroses. Partiellement du fait de la navigation chaotique entre les univers éloignés de la physique théorique et de l'économie appliquée. Partiellement du fait des contraintes géographiques et temporelles entre les deux agendas, entre Paris et Marseille. Et même s'il a fallu du travail pour en arriver là, j'ai tout de même eu une chance incroyable d'avoir été aussi bien entourée. Il est donc temps d'adresser tous mes remerciements à tous ceux qui m'ont fait confiance et m'ont entourée.

En tout premier, je tiens tout particulièrement à remercier mon directeur de thèse de m'avoir entraînée dans cette aventure autour du projet SocioPatterns. Il m'a offert cette formidable occasion de travailler sur un sujet novateur, à la confluence entre de nombreuses disciplines, avec des collaborations riches, dans un cadre de travail stimulant et dans de parfaites conditions matérielles. Je lui suis également infiniment reconnaissante de la confiance qu'il m'a accordée en me laissant une grande liberté tant du point de vue de la recherche que de l'organisation de mon travail, en acceptant ma décision de partir à Paris à l'ENSAE tout en continuant la thèse. J'ai particulièrement apprécié sa disponibilité, en me demandant toujours comment il faisait pour arriver systématiquement à répondre à mes mails en moins de 24 heures, à écouter mes difficultés et à en discuter presque immédiatement, à me faire rapidement des retour sur des chapitres entiers du manuscrit ou des présentations, même au milieu de l'été. Je remercie également les autres coordinateurs de SocioPatterns, Wouter, Jean-François et encore plus particulièrement Ciro. Bien qu'il n'ait pas été officiellement étiqueté comme co-directeur de thèse, les nombreux échanges scientifiques que nous avons eus, les encouragements et les précieux conseils qu'il a pu me prodiguer en font un personnage central dans l'histoire de cette thèse et dans mon apprentissage de la recherche. De plus, son enthousiasme, son énergie et sa curiosité insatiable dans tous les domaines ont été une profonde source d'inspiration pour moi. Je lui suis également reconnaissante du chaleureux accueil lors de mes séjours à Turin. Ma famille et mes amis devaient avoir du mal à croire que je n'étais pas partie en vacances lorsque je leur racontais mes visites touristiques, les soirées et *aperitivi*, l'excursion en montagne pour voir les chamois et le Cervin. Pourtant, ces séjours en Italie étaient véritablement des périodes de production scientifique intenses et qui alimentaient mes journées longtemps après mon retour en France.

Mes remerciements vont également à Nicolas et Philippe des Hospices Civils de Lyon, à Lorenzo, Vittoria et Marco de l'ISI, à Ginestra et Kun de Boston pour toutes ces collaborations qui ont fait la richesse de ces années passées sur cette thèse. Travailler avec de telles personnes a été un véritable plaisir. Je remercie aussi Titan et François pour s'être laissés embarquer dans le projet de deuxième année d'Ensae

sur l'homophilie. Ce projet que j'avais construit pour faire d'une pierre deux coups avec la thèse et qui a finalement débouché sur de beaux résultats, était pourtant plus improbable que la majorité des autres sujets de stat'app. Je leur suis redevable de m'avoir fait confiance, d'autant plus que nous ne nous connaissions pas bien à cette époque.

J'ai également beaucoup apprécié l'environnement du CPT à Marseille et je tiens à remercier Marc Knecht, Thierry Martin et Serge Lazzarini de m'avoir accueillie dans de bonnes conditions. Je suis également infiniment reconnaissante envers l'administration actuelle et passée de l'ENSAE, et tout particulièrement Sylviane Gastaldo et Elise Coudin, pour leurs encouragements et l'inestimable soutien moral et matériel qui m'ont permis de concilier mes travaux de thèse avec le parcours à l'ENSAE. Moins visibles mais non moins essentielles, je remercie les mains et âmes plus discrètes du CPT et de l'Ensae, en particulier Brigitte, Véronique, Marie-Hélène, Magatte, Vincent et Juliette, toujours prêts à aider les chercheurs ou étudiants et sans qui la science aurait du mal à avancer. Enfin, je tiens à remercier Pablo Jensen, qui a joué un rôle essentiel tant pour mon choix de thèse et de directeur de thèse que pour mon initiation à l'interdisciplinarité vers les sciences sociales.

Je tiens également à exprimer ma gratitude envers mes deux rapporteurs Alessandro Vespignani et Jean-Pierre Nadal pour le temps qu'ils ont consacré à la lecture détaillée du manuscrit et à la rédaction de leurs rapports. Avec les modalités actuelles de la fabrication du savoir scientifique, qui réduit toujours davantage la part de temps de recherche pure, je reconnais la valeur de ces heures de lecture critique du manuscrit. Je remercie également Frédéric van Wijland, Eric Gautier et Juliette Rouchier, les autres membres de mon jury d'avoir accepté de consacrer de leur précieux temps à considérer mes travaux.

Enfin, je voudrais remercier mes collègues de bureau Valerio, Emmanuele, Roberto, Anna, Arnab et Sarah qui m'ont quotidiennement supportée, mes deux relectrices Anna et Kristen, qui auraient pu trouver meilleure lecture pendant l'été, mes autres amis et collègues de Marseille et de Turin avec qui j'ai pu échanger, scientifiquement et moins scientifiquement. Si comme le disaient les latins, *mens sana in corpore sano*, mon équilibre pendant ces années doit beaucoup à Roberto, pour l'organisation des parties de foot du labo, à Antonino qui faisait souvent fi de son hypocondrie pour m'accompagner régulièrement courir dans les calanques, à Valerio pour m'avoir emmenée grimper la grande candelle, et à Jules pour la natation. Je remercie également le même Jules, Annalisa, Clément, Amandine, Edward, Kristen, Simone et Tapio, qui m'ont offert l'hospitalité lors de mes séjours dans la cité phocéenne après mon installation à Paris, et mon colocataire parisien, Joël. Je suis reconnaissante envers le PMMH et le LASIM, et leurs figures emblématiques Nawal et Denis, pour m'avoir accueillie quelques laborieuses journées lors de l'écriture du manuscrit et offert un environnement de travail bien meilleur que mon appartement parisien. Je remercie aussi le LSQ et la division études Sociales de l'Insee, mes deux nouvelles maisons, pour leur formidable accueil.

Enfin, les derniers remerciements mais certainement pas les moindres sont pour mes amis et ma famille pour toujours avoir été à mes côtés.

Synthèse en français

Introduction

Nous vivons dans une époque où l'informatique a pris une importance majeure dans notre société. Non seulement les ordinateurs ont modifié nos modes de travail et de vie, mais ils ont également largement contribué à l'émergence de la science dite des *systèmes complexes* grâce à la construction d'immenses bases de données et à la démultiplication des puissances de calcul. L'étude des réseaux complexes en est une sous-branche. Le paradigme de la complexité est d'étudier comment des relations individuelles, microscopiques, peuvent créer des phénomènes et structures macroscopiques qui ne répondent pas à une conception globale préalable. Dans le domaine des réseaux, la littérature s'est considérablement accrue depuis la fin des années 1990, à la suite de deux articles fondateurs [Watts 1998, Barabási 1999]. Ils mettent tous deux l'accent sur l'existence de structures semblables entre des réseaux très divers, et que l'on peut interpréter comme la trace de mécanismes simples, génériques au niveau microscopique. Un sujet d'étude important au sein de cette littérature concerne les processus dynamiques évoluant sur des réseaux. La topologie de ces réseaux est particulièrement importante pour le comportement général de ces processus. Par exemple, la sur-représentation de nœuds avec un nombre de voisins très élevé peut conduire à la disparition de transitions de phases. C'est par exemple le cas pour un modèle de diffusion très utilisé en épidémiologie, appelé modèle SIR, et dont l'existence d'une transition de phase rendaient les mesures de vaccinations particulièrement efficaces. Plus récemment, la problématique de la dynamique des réseaux en elle-même s'est invitée à la table. En effet, on étudiait généralement des réseaux statiques, comme si les objets qu'ils représentent avaient des relations immuables, alors que généralement, ces relations évoluent dans le temps. Par exemple, considérer un réseau de transport aérien comme statique revient à ignorer les changements saisonniers de routes, les créations d'aéroports, la fermeture d'autres, sous l'effet de la vétusté des équipements ou bien d'autres événements comme les éruptions volcaniques, les intempéries. Lorsque l'on étudie des processus dynamiques, cette dynamique du réseau lui-même peut revêtir une importance particulière, notamment lorsque les échelles d'évolution sont comparables. C'est dans cette perspective que je me suis intéressée lors de mon doctorat au réseau dynamique des proximités physiques.

Mon travail de thèse s'articule autour des mesures collectées dans le cadre de la collaboration SocioPatterns. Cette dernière avait mis en place peu avant le début de mon doctorat une infrastructure permettant d'enregistrer avec une très grande résolution temporelle la proximité humaine face-à-face entre individus volontaires sur des échelles de temps de l'ordre de plusieurs jours. Le dispositif repose sur l'utilisation de badges – dits RFID pour Radio Frequency Identification Devices – qui communiquent entre eux en émission et réception, ainsi qu'avec des antennes, par des signaux radio de faible puissance. Cette puissance peut être calibrée de manière

à ce que le rayon d'émission du signal entre les badges soit de l'ordre de un à deux mètres, et que le signal soit écranté par le corps humain. Ainsi, lorsque ces badges sont portés sur la poitrine des personnes volontaires, un signal n'est enregistré entre deux personnes que si elles sont proches l'une de l'autre, dans le rayon d'émission des badges, et si elles se font *plus ou moins* face. En effet, si une partie du corps humain se trouve entre les deux badges, par exemple si une personne tourne le dos à l'autre, alors le signal sera écranté. La fréquence d'émission et réception des badges est de quelques secondes. Ensuite, ces badges transmettent avec une autre amplitude l'information à des antennes branchées sur un réseau physique et celles-ci envoient le flux d'informations à un serveur central. On enregistre donc avec une résolution temporelle la séquence d'échange de signaux entre les badges. Cette séquence est ensuite agrégée par fenêtres de 20 secondes. On dit que deux personnes sont en contact sur la fenêtre de temps t si leurs badges se sont transmis au moins une fois des informations pendant l'intervalle de temps de 20 secondes correspondant. La durée d'un contact correspond au nombre de fenêtres de 20 secondes consécutives pendant lesquelles les personnes sont en contact. La durée a donc une valeur minimale de 20 secondes et incrémente par pas de 20 secondes. Compte tenu de l'infrastructure actuelle et plus précisément du fonctionnement des antennes, ce type de déploiement est restreint à des enceintes géographiquement peu étalées, dans un seul bâtiment, et n'est pas possible en plein air ou à l'échelle d'une ville.

L'infrastructure a été déployée dans des environnements très divers, tels que lors de conférences scientifiques, comme expérience interactive dans un musée scientifique, dans une école primaire ou encore dans une pépinière d'entreprises. La durée des déploiements varie entre 2 jours et 6 semaines, et le nombre de participants de 80 à presque 12000 personnes dans le cas du musée. Le taux de participation dépasse généralement les 80% lorsque toute la population d'étude peut être équipée d'un badge, et atteint fréquemment plus de 95%, ce qui est tout à fait remarquable en comparaison des enquêtes traditionnelles qui demandent un investissement plus important de la part des participants.

Analyse statistiques des données de proximité

La première étape de la thèse est de caractériser de manière descriptive les données. Le point le plus remarquable est certainement la dynamique très hétérogène des durées des contacts, des durées entre contacts et des durées des groupes (la durée d'un groupe de taille p, de manière analogue à la durée d'un contact, est égale au nombre de fenêtres de 20 secondes pendant lesquelles un groupe de p personnes sont en contact). Cette hétérogénéité se traduit quantitativement par des distributions larges, c'est-à-dire qui décroissent en loi de puissance $P(x) \sim x^{-\alpha}$. Cela signifie d'une part que les durées des contacts sont généralement courtes, mais parfois très longues, et d'autre part que l'on ne peut pas identifier de durée caractéristique, comme on peut le faire avec des distributions gaussiennes par exemple. Ce résultat a été observé quel que soit le contexte, et qui plus est, les distributions sont très

similaires entre les différents contextes. Une telle observation avait été faite concernant les temps d'inactivité dans les correspondances humaines. Plus précisément, la littérature scientifique montrait que la distribution des temps écoulés entre deux envois de mails, deux appels téléphoniques ou des envois de courrier suivaient des lois de puissance [Rybski 2009]. Quant à la proximité humaine, d'autres mesures faites avec des infrastructures différentes tendent à confirmer actuellement le caractère général de ces hétérogénéités [Hui 2005, Salathé 2010].

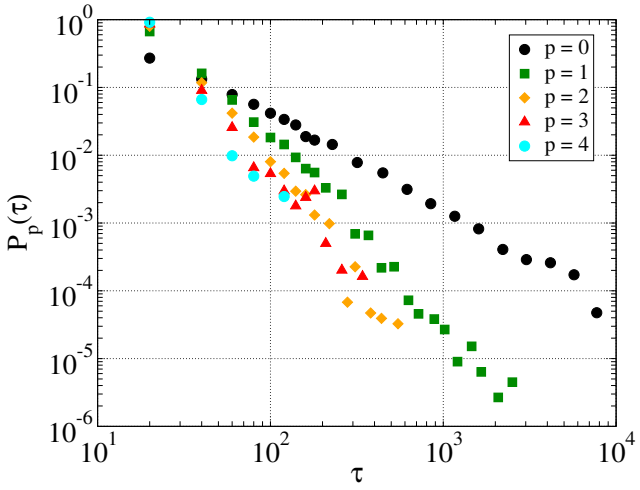


Figure 1: Distribution $P_p(\tau)$ des durées en secondes des groupes de taille $p+1$ lors d'une conférence scientifique (un groupe de taille 1 correspondant à une personne seule, donc sans contact avec aucune autre personne). L'échelle logarithmique en abscisse et ordonnée permet de mettre en évidence le caractère très hétérogène des durées des groupes.

D'autre part, pour analyser la structure des interactions, je me suis également servi du cadre formel des réseaux statiques. Un réseau est une représentation théorique d'interactions relativement homogènes entre entités homogènes. Les entités, dans notre cas les personnes, sont représentées par des nœuds, d'un ou plusieurs types (par exemple homme ou femme), et les interactions sont symbolisées par des liens entre une paire de nœuds. Lors de ma thèse, je me suis intéressée le plus souvent à des réseaux dit *agrégés*,pour lesquels on considère qu'un lien existe entre deux personnes si elles ont eu au moins un contact sur une certaine durée, souvent sur quelques minutes ou une journée. À chaque lien est associé un poids, définit généralement comme la durée cumulée des interactions sur l'échelle de temps sur laquelle on agrège les contacts. Ensuite, la boite à outils traditionnelle des réseaux sert à caractériser la structure des relations, avec des grandeurs définies pour chaque nœud, comme le degré, i.e. le nombre de nœuds avec lesquels le nœud en question est directement relié, et des grandeurs définies de façon globale, comme le diamètre, i.e. la plus grande distance géodésique entre les nœuds d'une composante connexe.

On retrouve dans ces observations des traces caractéristiques du contexte de

l'événement. Par exemple, le diamètre du réseau agrégé à l'échelle d'une journée pour une conférence scientifique est bien plus petit que celui que l'on mesure pour un musée. En effet, un des objectifs principaux d'une conférence est de faire se rencontrer les scientifiques. Il est assez heureux que les interactions qui se font alors n'isolent pas un ou plusieurs groupes de chercheurs. En revanche, un musée n'est pas fait pour qu'une communauté de personnes se rencontre. De plus, un visiteur aura peu de chance de rencontrer les visiteurs qui arrivent même une heure plus tard, créant ainsi un réseau agrégé allongé, dont la structure est liée à la temporalité des visites. Quantitativement, le diamètre passe par des nœuds correspondant à des personnes dont les heures d'arrivée sont chronologiquement assez ordonnées.

Un autre exemple de l'impact du contexte de l'événement sur la structure des réseaux agrégés est donné par les interactions collectées dans une école primaire. Dans ce cas, les enfants interagissent avant tout avec les enfants de leur classe et à moindre échelle avec les enfants de même niveau scolaire, y compris pendant les récréations et le déjeuner, qui sont les périodes d'activité sociale les plus intenses. L'infrastructure pour mesurer les interactions trouve alors une importance particulière pour les enjeux épidémiologiques. En effet, il est assez difficile de quantifier les contacts susceptibles de transmettre une maladie transmissible telle que la grippe dans une population. En général, les mesures que l'on trouve dans la littérature scientifique reposent sur des questionnaires auto-administrés, où l'on demande aux personnes de détailler les contacts qu'ils ont eu au cours d'une journée typique. Les données alors recueillies sont particulièrement sensibles aux biais de mémoire des individus qui se souviennent assez mal des rencontres qu'ils font, surtout lorsque celles-ci durent peu de temps [Smieszek 2011]. Avec les badges RFID, nous pouvons quantifier de manière très détaillée les interactions entre enfants de la même classe et entre différentes classes. En particulier, on observe que les enfants ont rencontré individuellement la majorité des enfants de leur classe en moins deux heures, alors que les pauses et le déjeuner leur permettent d'interagir avec les enfants des autres classes mais de manière assez réduite puisqu'au bout de deux jours, un enfant n'a eu de contacts qu'avec, en moyenne, moins d'un tiers de l'école. Il est à noter que cette proportion tendrait à augmenter encore mais relativement lentement sur une troisième journée. Ces données détaillées sont très importantes pour estimer l'impact de stratégies de fermeture de classes dans le cas d'épidémies, ce qui peut constituer une stratégie de santé publique plus intéressante que la fermeture d'écoles entières, sachant d'autant plus que les écoles sont des lieux privilégiés pour la transmission de maladies au sein de la population entière.

Tester des théories socio-psychologiques

Le contexte est un élément important pour contraindre la structure des interactions humaines, mais d'autres mécanismes entrent également en jeu. La littérature des réseaux sociaux est à ce propos particulièrement riche. Un mécanisme important sur lequel j'ai travaillé lors de ma thèse est l'homophilie, mécanisme qui correspond à la tendance des individus à entretenir des relations sociales avec des individus

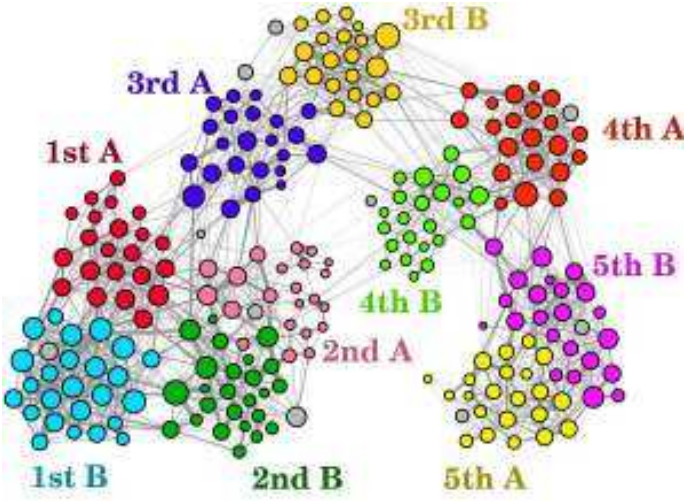


Figure 2: Réseau de contacts dans une école primaire, agrégé à l'échelle d'une journée. Les liens entre les individus ayant interagi moins de deux minutes ont été retirés par soucis de lisibilité. La largeur des liens correspond à la durée cumulée des interactions et les nœuds ayant un degré plus important sont plus gros. Les couleurs correspondent aux classes, les enseignants étant en gris.

qui leur ressemblent, que ce soit sur le plan des opinions, de la catégorie sociale, de l'âge, ou de beaucoup d'autres choses. D'autres mécanismes comportementaux, tels que la fermeture triadique, la réciprocité, le prestige, la balance structurelle contribuent également à la structure. Généralement ces mécanismes sont étudiés sur des réseaux dits sociaux, c'est-à-dire reposant sur les déclarations de relations par les individus. Par exemple, on demande à chaque individu de donner la liste de ses amis, des personnes auprès desquelles il demanderait des conseils, etc. Ces relations déclarées ont donc une signification pour les individus en termes de liens affectifs ou professionnels, ils s'agit de personnes qui *comptent*. Dans ce sens, la proximité physique entre individus est très éloignée de ce que l'on considère comme un lien social et est généralement considérée comme un indicateur assez pauvre des relations sociales. Il est cependant assez évident qu'un lien social se construit autour et se nourrit d'interactions réelles, généralement liées à la proximité physique, au moins à un moment donné de la relation. Il est donc assez légitime de se demander si les mécanismes déterminant la structure des réseaux sociaux sont également présents pour les interactions instantanées, comportementales.

Je me suis intéressée à l'homophilie de genre entre enfants d'une école primaire. On peut définir une homophilie à l'échelle mésoscopique et une homophilie à l'échelle individuelle. Dans le premier cas, la question est de savoir si à l'échelle d'un groupe, d'une classe par exemple, les interactions ont lieu préférentiellement entre personnes du même sexe. Pour tester l'hypothèse selon laquelle la structure du réseau social n'est pas corrélée au sexe des personnes, on construit un modèle statistique simple,

dans lequel la probabilité qu'un lien existe entre deux nœuds est indépendante du sexe des nœuds. Ce modèle permet de déterminer la loi du nombre de liens entre personnes du même sexe, avec relativement peu d'hypothèses. On compare ensuite le nombre observé de liens entre personnes du même sexe avec la distribution de probabilité que l'on obtiendrait si le modèle d'indépendance était vrai. On trouve alors que, de manière significative, les filles ont collectivement tendance à interagir davantage entre elles qu'avec les garçons dans 6 classes sur 10, et inversement, les garçons ont tendance à interagir davantage entre eux qu'avec les filles dans 8 classes sur 10.

Au niveau individuel, on définit l'homophilie d'un individu comme le ratio entre le nombre d'enfants du même sexe que lui avec lesquels il interagit plus de cinq minutes pendant les pauses déjeuner et les récréations sur deux journées, divisé par le nombre total d'enfants avec lesquels il interagit plus de cinq minutes. Ce ratio est nul si l'enfant n'interagit qu'avec des enfants du sexe opposé et vaut 1 si inversement, l'enfant n'interagit qu'avec que des enfants du même sexe. On obtient ainsi un indice d'homophilie pour chaque enfant de l'école, ce qui permet de faire des analyses statistiques plus fines qu'au niveau de la classe. En résumé, les analyses montrent que l'homophilie est plus importante chez les garçons que chez les filles et qu'elle augmente avec l'âge. Ces résultats sont en accord avec la littérature dans le domaine. Nous pouvons donc conclure sur le fait que la mesure de la proximité physique donne des résultats similaires à l'étude des relations sociales concernant l'homophilie de genre. Par ailleurs, on montre que la proportion d'interactions de très courtes durées (moins de trois minutes) faites avec le sexe opposé augmente avec l'âge pour les filles alors qu'elle diminue pour les garçons. Des évolutions sont donc très différentes entre les deux sexes concernant les relations de courtes durées, ce qui n'avait jamais été mentionné auparavant, certainement parce qu'il est très difficile de quantifier les relations avec des *personnes qui comptent moins ou peu* avec les méthodes traditionnelles en sciences sociales.

Un autre aspect que j'ai étudié dans ma thèse concerne le lien entre la proximité physique et les réseaux sociaux sur internet, tels que Facebook, Flickr, Delicious et LastFM. Pour ce travail, nous nous sommes appuyés sur une articulation entre la plate-forme Live-Social Semantics qui permet d'étudier le comportement d'internautes sur le web 2.0 et l'infrastructure de mesure de la proximité physique par les badges radio. Une mesure groupée lors d'une conférence a permis de collecter des informations sur les deux aspects qui nous intéressent. En comparant la structure des réseaux en ligne et celle du réseau agrégé sur la durée du déploiement, nous avons pu mettre en évidence que les contacts entre participants sont plus fréquents et durent plus longtemps entre personnes qui partagent un lien virtuel (par exemple une amitié sur Facebook) qu'entre ceux qui n'en partagent pas. Ces personnes liées virtuellement se comportent de manière plus semblable que si elles ne l'étaient pas. Ces résultats permettent d'affirmer que les relations de proximité physique lors d'une conférence contiennent de l'information sur l'existence de liens virtuels, ce que nous quantifions en terme de prédiction de liens virtuels. Si des résultats antérieurs montraient que les relations virtuelles existaient davan-

tage entre personnes géographiquement proches (au mieux à l'échelle d'une ville) [Liben-Nowell 2005, Takhteyev 2012], c'est à notre connaissance la première étude faite à l'échelle de la proximité physique de l'ordre du mètre.

Modéliser la dynamique des rencontres

Comme évoqué précédemment, les distributions des temps de contacts et entre plusieurs contacts sont très similaires d'un contexte à l'autre. On aurait pu s'attendre intuitivement à des distributions d'allures bien différentes entre les données collectées dans un musée et celles des conférences scientifiques. De même, il est relativement surprenant que les durées des contacts entre enfants dans une école primaire soient distribuées de manière analogue. C'est pourquoi, avec Ginestra Bianconi et Kun Zhao de la Northeastern University et mon directeur de thèse, nous avons travaillé à un modèle proposant une piste d'explication à cette observation. Le modèle que nous avons construit est dit *individu-centré* puisque nous définissons au niveau individuel des règles d'interaction et d'évolution. Ces règles peuvent reproduire au niveau collectif la phénoménologie observée dans les données empiriques. Une autre manière de construire un modèle reproduisant cette phénoménologie aurait été de donner comme ingrédient d'entrée la distribution que l'on souhaite obtenir. Plus précisément, si l'on souhaite obtenir une distribution des temps de contacts qui suive une loi quelconque, alors on peut reproduire cette phénoménologie en construisant des temps de contacts comme des variables aléatoires tirées dans cette loi. Cette méthode a par exemple été retenue par Rocha et al. qui s'intéressaient à l'influence de la distribution des temps de contact pour la propagation de maladies contagieuses [Rocha 2012]. Un tel modèle ne cherche pas à expliquer mais seulement à reproduire la phénoménologie. Notre modèle, en revanche, par ses règles d'interactions au niveau individuel a vocation à donner une interprétation plausible à cette phénoménologie. Il s'inscrit dans la même veine que le modèle de Barabàsi de 2005 de files d'attente dans les communications humaines [Barabási 2005] et plus généralement des modèles provenant de la littérature des réseaux complexes [Boccaletti 2006].

Le modèle est le suivant. On considère un ensemble de N agents qui peuvent être soit isolés, soit interagir sous forme de groupes. Il vise à décrire une population dans une enceinte relativement petite (il n'y a pas d'effet de distance entre les individus) et sur une échelle de temps de quelques jours (les effets de réseaux sociaux à proprement parler sont négligés, n'importe qui interagit avec n'importe qui d'autre). Les groupes représenteraient le type de petits groupes de discussion que l'on observe dans les données empiriques. Un agent peut, au cours de l'évolution dynamique du modèle, rester seul, quitter un groupe ou bien en rejoindre un. Les groupes ne peuvent pas fusionner entre eux, ni se scinder ; ils grossissent ou diminuent uniquement de manière incrémentale par l'introduction d'un nouveau membre ou bien par la sortie d'un autre. Plus précisément, chaque agent i est décrit par une variable d'état p_i qui varie dans le temps et qui vaut 0 si l'agent est isolé, ou le nombre de personnes du groupe dans lequel il est, excepté lui-même. La dynamique est discrète. À chaque

pas de temps t, un agent i est sélectionné complètement aléatoirement.

- Si cet agent est isolé, i.e. $p_i = 0$, alors avec une probabilité $b_0 f(t, t_i)$ il choisit un autre agent isolé j pour former une paire. Cette probabilité dépend à la fois du temps t mais aussi du temps t_i de la dernière fois que l'agent i a changé d'état. L'autre agent isolé j est choisit parmi les agents isolés avec une probabilité $Pi(t, t_j)$, dépendant également du temps t et du dernier changement d'état de j.

- Si l'agent est dans un groupe, alors avec une probabilité $b_1 f(t, t_i)$, son état change. Dans ce cas, avec une probabilité λ, l'agent quitte le groupe et devient isolé, et sa variable p_i devient nulle alors que celles des autres membres du groupe diminue d'une unité. Sinon, l'agent i introduit dans le groupe un agent isolé j, choisit parmi les agents isolés avec une probabilité $\Pi(t, t_j)$. La variable de chacun devient alors égale à la taille du groupe diminué d'une unité.

Le modèle est donc décrit par trois paramètres, b_0, b_1 et λ qui contrôlent la probabilité de rester isolé, de garder la taille d'un groupe constante et la tendance à quitter un groupe plutôt que d'introduire un nouveau membre, et par deux fonctions f et Π.

Sous certaines approximations et pour certaines valeurs des paramètres b_0 b_1 et λ, on peut résoudre analytiquement le modèle et obtenir la distribution $P_p(t)$ du temps passé dans l'état p. Par exemple, dans le cas où f et Π sont constants, alors ces distributions décroissent exponentiellement avec t, alors qu'avec $f(t) = \Pi(t) = (1 + t/N)^{-1}$, on obtient des distributions larges :

$$\begin{cases} P_0(\tau) = (1+\tau)^{-1-b_0 \frac{3\lambda-1}{2\lambda-1}} \\ P_p(\tau) = (1+\tau)^{-1-(p+1)b_1} \quad \text{pour } p \geq 1. \end{cases} \tag{1}$$

Ces distributions sont particulièrement intéressantes puisqu'elles correspondent à notre phénoménologie empirique.

Le modèle se prête particulièrement bien aux simulations numériques. Nous avons pu vérifier le bon accord entre celles-ci et les résultats analytiques, justifiant ainsi les approximations faites. Nous avons également pu explorer grâce aux simulations numériques le comportement du modèle en lui apportant de légères modifications, comme l'introduction d'une hétérogénéité entre les agents (ils n'auraient dans ce cas pas tous les mêmes propensions à rester inactifs, ou à maintenir la taille du groupe stable dans temps) ou bien considérer un nombre d'agents N qui varierait dans le temps. Si la résolution analytique est encore possible, bien que plus ardue, dans le premier cas, nous ne l'avons pas tentée dans le second. Les distributions de durées de vie des groupes et des temps d'inactivité restent distribuées selon des lois larges, malgré ces modifications. Dans le deuxième cas, on peut reproduire une phénoménologie encore plus proche de celle observée empiriquement, avec par exemple une dynamique circadienne caractérisée une activité intense le jour et plus faible la nuit.

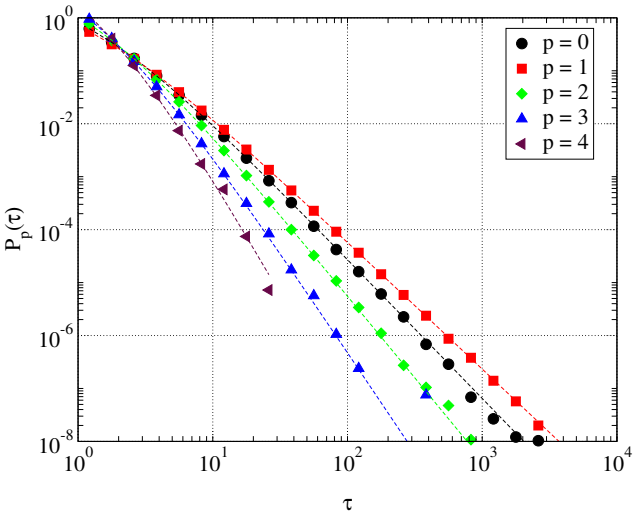


Figure 3: Distribution $P_p(\tau)$ des temps passés dans un état p. Les simulations numériques sont faites avec $N = 1000$, $b_0 = b_1 = 0.7$, $\lambda = 0.8$ et pour une durée de simulation totale de $T = 10^6 N$ pas de temps. Les lignes représentent le résultat analytique.

Même si le modèle ne propose pas une explication complète à cette distribution des durées de contacts entre les différents contextes, il suggère une piste. En effet, la présence de mécanismes de renforcement semble être déterminante. Ceux-ci peuvent être résumés de la manière suivante : plus un agent est seul, moins il a de chances d'engager une interaction ou d'être invité à rejoindre un groupe, et plus longtemps un agent reste dans un groupe, moins il a de chance de le quitter. Ce principe est suffisamment générique pour exister dans des contextes bien différents. Cependant, sans une étude de type cognitive, il est difficile d'aller plus loin dans l'origine de tels mécanismes de renforcement. Une piste pourrait être l'existence d'un compromis coût à changer d'état/bénéfice à ne rien changer, qui, au fil du temps, tendrait à voir le bénéfice l'emporter sur le coût.

Propagation de maladies

La dynamique des contacts a de profondes conséquences pour les processus dynamiques faisant intervenir les contacts entre personnes. Ces processus peuvent, par exemple, décrire la synchronisation d'opinions, les effets de mode, la diffusion d'information ou la transmission de maladies contagieuses. C'est à ce dernier type de processus dynamique que je me suis intéressée. La question que nous nous sommes posée est de savoir si la dynamique des contacts pouvait modifier les résultats des modèles épidémiologiques classiques qui font généralement l'hypothèse d'une structure de contact fixe dans le temps. Plus généralement, on s'intéresse à la coévolution d'un processus dynamique et d'un réseau dynamique, sachant que le processus a pour support ce réseau. On peut comprendre que si les temps carac-

téristiques d'évolution sont bien différents, l'approximation d'un support constant dans le temps peut être pertinente. En revanche, plus les échelles de temps sont proches, moins l'approximation est satisfaisante.

Nous avons étudié les effets de la prise en compte de différents niveaux d'informations sur la structure des contacts pour l'évolution d'un modèle de contagion, dit SEIR. Dans un tel modèle, très courant en épidémiologie, chaque individu est soit sain (S), soit dans un état latent (E), c'est-à-dire qu'il a été contaminé mais qu'il n'est pas encore infectieux, soit infectieux (I) et dans ce cas, il peut infecter un individu sain avec lequel il est en contact, soit guéri (R). Les paramètres de transition entre des différents états permettent de caractériser le modèle. Nous avons considéré dans notre cas que le passage d'un état latent à un état infectieux, la contamination d'un individu sain par un infectieux avec lequel il est en contact et la guérison sont décrits respectivement par des processus de Poisson de taux σ, β et ν. Nous considérons deux scénarios de maladie, l'un très rapide ($\sigma^{-1} = 1$ jour, $\beta = 30.10^{-5}s^{-1}$ et $\nu^{-1} = 2$ jours) et un deuxième moins rapide ($\sigma^{-1} = 2$ jours, $\beta = 15.10^{-5}s^{-1}$ et $\nu^{-1} = 4$ jours), compatible avec une grippe très virulente.

À partir de données de contacts collectées dans une conférence scientifique, nous construisons trois modèles de structure des contacts, contenant différents niveaux d'information sur la dynamique.

- Le premier modèle, contenant le moins d'information, est appelé HOM. Il consiste en l'alternance de réseaux quotidiens non pondérés, dans lequel les nœuds, qui représentent des individus, sont connectés si au moins un contact a été enregistré entre les personnes correspondantes pendant la journée en question.

- Le second modèle, appelé HET contient l'information de HOM et l'information sur l'hétérogénéité des durées cumulées d'interaction. En effet, HET consiste en un réseau pondéré, similaire au réseau précédant, à l'exception faite qu'à chaque lien correspond un poids représentant la durée cumulée des interactions de la journée en question. Dans ce modèle, le taux de contamination β est proportionnel à cette durée cumulée.

- Le troisième modèle et le plus riche, est appelé DYN. Il s'agit de la succession temporelle des contacts enregistrés avec une résolution de 20 secondes. L'information sur l'hétérogénéité des contacts est bien entendue contenue dans un tel modèle, mais ce dernier contient en surcroît la temporalité des interactions, de sorte que si A et B ont un contact puis B et C, alors une maladie ne peut pas être transmise de C à A alors que c'était le cas dans le modèle HET.

Malheureusement, les données sur les contacts de la conférence scientifique sont étendues sur une durée de deux jours seulement. Étant donnée la gamme de paramètres épidémiologiques considérée, deux jours de simulation est une durée bien trop courte pour voir l'évolution de la contagion dans son ensemble. Pour pallier à cette limite, nous avons étudié trois méthodes d'extension temporelle des données de contact.

La première et la plus simple consiste à simplement répéter le film des interactions tous les deux jours sur plus de 100 jours. Une seconde méthode, à l'inverse, consiste à échanger de façon complètement aléatoire l'identité des nœuds tous les deux jours et à rejouer la même séquence d'interactions. Une troisième méthode se situe à l'intermédiaire entre la réplication pure et simple et le mélange complètement aléatoire. Il est important de noter que ces méthodes de réplications changent certes les niveaux et la vitesse de contagion, mais ne modifient pas les différences assez générales que l'on observe entre les trois modèles de contact.

Un raisonnement sur le nombre moyen d'individus infectés au bout d'une journée permet d'établir des équivalences entre les paramètres σ, β et ν pour les différents modèles de contacts. Cette équivalence est nécessaire pour pouvoir faire des comparaisons entre les modèles de contacts à modèle épidémiologique donné. Le tout est simulé numériquement, en sélectionnant de manière complètement aléatoire un seul individu infectieux que l'on appelle la graine et qui est à l'origine de la contagion pour chaque tour de simulation. On s'arrête uniquement lorsque plus aucun individu n'est susceptible d'être contaminé, c'est-à-dire lorsque le nombre d'infectieux et de latents est nul. Les comparaisons sont faites pour 5000 simulations.

La première quantité que l'on considère est le taux de reproduction de base. Cette quantité est définie comme le nombre moyen d'individus infectés par la graine. Il s'agit d'une quantité très souvent regardée en épidémiologie puisque dans les modèles épidémiologiques les plus simples, elle définit un seuil entre une maladie qui tendrait à s'éteindre très rapidement et une maladie susceptible de contaminer une fraction importante de la population. Dans notre cas, la distribution du nombre moyen d'individus infectés est exponentielle, avec une valeur moyenne pour le modèle HOM sensiblement supérieure à celles obtenues pour HET et DYN, très semblables entre elles. La taille finale de l'épidémie, définie comme le nombre total d'individus qui ont été infectés, est quant à elle bimodale. En effet, une proportion non négligeable de simulations donne peu de cas, généralement parce que la contagion s'est éteinte après peu de transmissions. En revanche, lorsque la contagion atteint une proportion considérable de la population, le modèle HOM donne des valeurs à nouveau sensiblement supérieures à celles obtenues pour HET et DYN qui restent très semblables. Ces deux résultats suggèrent que l'hétérogénéité des durées cumulées de contacts donne une moindre transmission, suggérant ainsi que le partage non équitable du temps passé avec les autres diminuerait les canaux de transmission. En revanche, la temporalité des contacts ne semble pas avoir d'effet à cette échelle.

Malgré ces différences de niveau, l'évolution temporelle est bien similaire entre les trois modèles. Le pic épidémique, c'est-à-dire l'instant où le nombre d'infectieux atteint son maximum, est tout à fait comparable dans les trois cas. Ceci suggère que l'estimation de ce pic se fait assez bien avec un modèle très pauvre en information sur l'hétérogénéité des durées des contacts. Il est néanmoins nécessaire pour sa construction de connaître de manière assez fiable la durée moyennes des interactions cumulées ainsi que leur structure dans la population.

Cette analyse a permis de mettre en évidence que si la temporalité des contacts

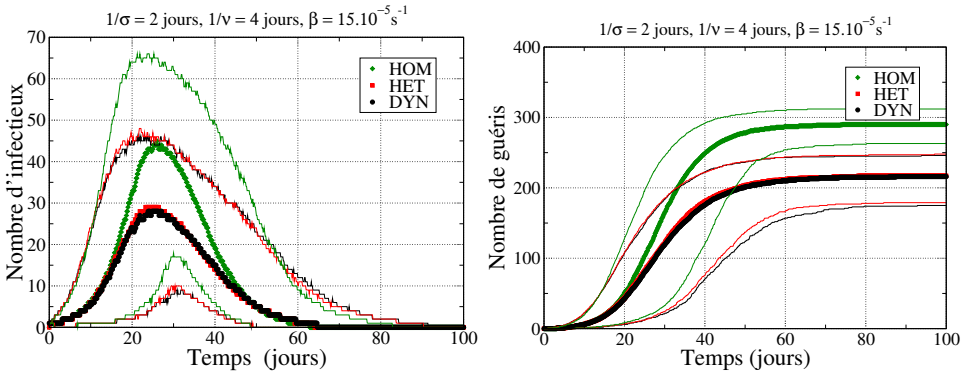


Figure 4: Evolution temporelle du processus de propagation pour le scenario compatible avec une grippe. Le graphique de gauche donne la prévalence, c'est-à-dire le nombre d'individus infectieux, et le graphique de droite donne le nombre d'individus guéris. Seules les simulations dans lesquels plus de 10% de la population a été contaminée sont considérés. Les symboles donnent les valeurs médianes et les lignes donnent les 5e et 95e percentiles du nombre d'infectieux et d'individus guéris.

n'est pas nécessaires à cette échelle, l'hétérogénéité des durées cumulées d'interaction est nécessaire pour correctement estimer l'ampleur de la propagation épidémique. En revanche, il est suffisant de savoir simplement de qui a été en contact avec qui et de connaître la durée moyenne des interactions pour correctement estimer l'instant du pic épidémique. Dans cette mesure, nous apportons une contribution nouvelle à la littérature puisque l'hétérogénéité des durées entre deux contacts avait certes été étudiée mais ce n'était pas le cas des hétérogénéités des durées de contacts. Les résultats que nous obtenons restent à nuancer. L'échelle de temps du processus de contagion que nous avons considérée reste réduite. Il est possible par exemple que des différences entre la dynamique de HET et DYN apparaissent lorsque l'on considère une dynamique bien plus rapide (et alors peu réaliste pour une propagation de maladie). Dans le futur, il serait intéressant de considérer des modèles théoriques de contact, comme celui que j'ai présenté dans la section précédente. Ainsi, l'existence de paramètres permettant de contrôler finement la structure des contacts permettrait d'analyser l'effet des hétérogénéités des durées de contacts et entre contacts sur la propagation de maladie, et de voir si l'un effet domine l'autre ou non. Plus généralement, ce type d'étude devient actuellement important puisque l'accès à des données de plus en plus détaillées sur les contacts dans une population devient possible. Il est dorénavant nécessaire de pouvoir estimer l'apport d'un effort supplémentaire, puisque un niveau de détail extrême n'est pas forcément nécessaire pour les estimations que l'on souhaite obtenir ou pour établir des stratégies de santé publiques sophistiquées, telles que la fermeture de classe ou les vaccinations ciblées.

Conclusion

Cette thèse s'articule autour de l'étude de la dynamique de proximité humaine, mesurée par des badges radio. Cette méthode de collecte non-supervisée permet d'obtenir de manière très détaillée des informations sur les interactions humaines. Mes travaux présentent la caractérisation statistique de la dynamique de proximité physique, mise en relation avec le contexte et les autres métadonnées disponibles, telles que l'âge, le sexe des individus, ou bien la structure de leurs réseaux sociaux virtuels. On retiendra que si la structure des contacts diffère considérablement selon le contexte, les distributions empiriques des durées des interactions et entre interactions sont très similaires. Afin d'interpréter cette similarité, j'ai travaillé sur un modèle individu-centré qui propose des règles d'interactions microscopiques simples susceptibles de donner lieu à cette structure macroscopique complexe des temps d'interaction. Enfin, la caractérisation de la dynamique des contacts entre individus constitue une étape cruciale pour comprendre les mécanismes de propagation de maladies telles que la grippe dans une population. Les données de proximité humaine ont permis d'étudier la quantité d'informations nécessaires sur la dynamique des contacts pour la construction de modèles épidémiologiques de contagion.

Deux types de prolongement peuvent être envisagés suite à ces travaux. D'une part, les travaux sur la propagation de maladies demandent à être poursuivis, avec la quantification pure et simple des contacts entre les différentes strates d'une population, et l'étude des différentes approximations de modélisation et l'impact de degrés de détails supplémentaires sur la pertinence des estimations. D'autre part, en sciences sociales, les travaux de comparaison entre la proximité physique et les relations sociales pourraient être étendus à d'autres mécanismes, d'autres sources d'homophilie, mais surtout, cette méthodologie pourrait faciliter l'étude de la dynamique de changement des relations, mise en liens avec les comportements des individus. Plus précisément, l'étude quantitative de la coévolution des réseaux sociaux et des comportements des individus est un domaine encore peu étudié et qui est confronté à la difficulté majeure de collecte de données. Quitte à perdre éventuellement en qualité de l'indicateur de proximité sociale, ce qui reste à démontrer, la mesure non supervisée de relations de proximité physique, couplées à d'autres informations, pourrait être une piste intéressante pour pallier la difficulté.

Contents

1 Introduction **1**

1.1 A biased overview of the network science landscape 1

1.1.1 The complex network paradigm 2

1.1.2 Dynamical processes on complex networks 5

1.1.3 Network dynamics 6

1.2 Mining human interactions 8

1.2.1 Before the technological era 8

1.2.2 Using Bluetooth, WiFi and RFID devices 9

1.2.3 The SocioPatterns collaboration 11

1.2.4 Self-reported data vs behavioral data 14

1.3 Structure of the following chapters 16

2 Statistical analysis of face-to-face proximity data **17**

2.1 Some useful network concepts 17

2.2 Interactions in different environments: museum and conferences . . . 21

2.2.1 Dynamical burstiness 21

2.2.2 Network static characteristics 22

2.2.3 Distances network 28

2.2.4 Discussion 31

2.3 Primary school 32

2.3.1 School composition and category structure 33

2.3.2 Results 33

2.3.3 Discussion 40

2.4 Partial conclusion and perspectives 44

3 Testing socio-psychological theories **46**

3.1 Motivations 46

3.1.1 Drivers of social networks 46

3.1.2 Socio-psychological theories on virtual networks 47

3.2 Gender homophily among children 48

3.2.1 Background about gender homophily among children 49

3.2.2 Results 50

3.2.3 Discussion 60

3.3 Physical interactions versus online social networks 61

3.3.1 The Live-Social Semantics platform 62

3.3.2 Results 63

3.3.3 Discussion 71

3.4 Partial conclusion and perspectives 71

4 Modeling the dynamics of encounters **73**
4.1 Motivations . . . 73
4.2 Model of interactions in a homogeneous population . . . 75
4.2.1 Description of the model . . . 75
4.2.2 Analytical solution with constant transition probabilities . . . 77
4.2.3 Analytical solution with a rich-get-richer effect . . . 78
4.2.4 Numerical simulations . . . 81
4.2.5 Aggregated networks . . . 86
4.3 Variation on the model . . . 87
4.3.1 Heterogeneous population . . . 87
4.3.2 Fluxes in the population . . . 91
4.4 Partial conclusions and perspectives . . . 95

5 Disease spreading **97**
5.1 Motivation . . . 97
5.1.1 Dynamic processes on networks . . . 97
5.1.2 Modeling disease spreading . . . 98
5.1.3 Toward more realism in contact patterns . . . 103
5.1.4 The need of data on contact patterns . . . 108
5.1.5 Why does contact dynamics matter? . . . 109
5.2 Simulation of an SEIR model on empirical data . . . 112
5.2.1 Data collection . . . 112
5.2.2 Description of the model . . . 113
5.2.3 Results . . . 119
5.2.4 Limitations . . . 124
5.3 Partial conclusion and perspectives . . . 127

6 Conclusion **129**

A List of publications **133**

Publications **133**

Glossary **134**

Bibliography **136**

Chapter 1

Introduction

Contents

1.1 A biased overview of the network science landscape 1
1.1.1 The complex network paradigm 2
1.1.2 Dynamical processes on complex networks 5
1.1.3 Network dynamics 6
1.2 Mining human interactions 8
1.2.1 Before the technological era 8
1.2.2 Using Bluetooth, WiFi and RFID devices 9
1.2.3 The SocioPatterns collaboration 11
1.2.4 Self-reported data vs behavioral data 14
1.3 Structure of the following chapters 16

1.1 A biased overview of the network science landscape

Network is a vast research object studied in many fields, ranging from the so called exact sciences, such as mathematics, statistics and physics, to social sciences, with organizational theory, economics and sociology. The common feature among scientists of these very different fields is that network science provides them with a framework to describe relationships between objects. The latter can be as varied as cells, human beings, computers or cities, that can be linked to each other with very different relationships. Behind the variety of these objects which would naturally require the use of different methodologies, the description of systems in network terms relies on the simple assumption – or simplification– that all these objects are similar enough to be described as one or few types of nodes, equivalent to each other, and that the relationships between all of them can be summarized to one or few types of relationships. This may sound rather vague, but in a nutshell, a network description is a way to look at a system, that gives the priority to the relationships between objects rather than to their individual variety or to their intrinsic specificity.

As this network science is too vast to be described adequately in all the specificities of each field, I can not do better than giving my personal overview of this research area, which will only cover a partial set of works, and discard many very interesting results I have not found a suitable place to recount these. The bias of this section comes from my background in physics, which made me enter network

science via the door of complex systems. Along the course of my doctorate, I have had the opportunity to learn more about economics, statistics and sociology and the works done in these domains on networks, but even so, my thesis lies within the framework of complex systems, which will be discussed in this chapter.

1.1.1 The complex network paradigm

Network science has been initiated in two disciplines: graph theory in mathematics and social networks in sociology. Graph theory dates back to the seminal problem of Leonhard Euler presented in 1735 and published in 1741, called the Seven Bridges of Königsberg problem [Euler 1741]. This problem consisted in finding a walk through the city of Königsberg, whose simplified map is given in figure 1.1 (left), that would cross each bridge once and only once. Given the configuration of the bridges, Euler showed that there was no solution. The second father of graph theory is Paul Erdös, who was the first with Alfréd Rényi to introduce probabilistic models and whose work started the branch of random graph theory [Erdős 1959]. In sociology, network concepts were first introduced by Jacob Moreno who introduced sociograms to describe relationships among children [Moreno 1953]. These sociograms are drawings of relationships with people represented as circles and relationships as an edges between two circles (see right panel of figure 1.1).

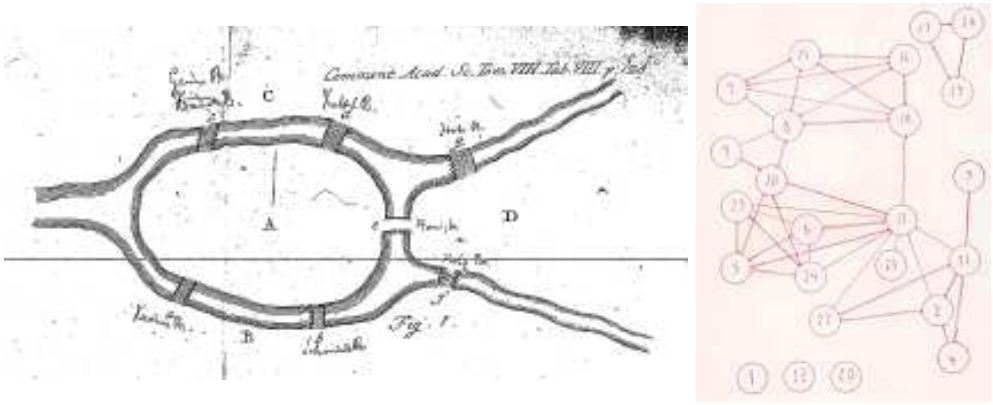


Figure 1.1: Left: the Seven Bridges of Königsberg problem from Euler (figure from [Euler 1741]). Right: an acquaintance sociogram of Moreno (figure from [Moreno 1953]).

At the end of the 90's (see a timeline on complex system modeling in figure 1.2), two fundamental articles started a new literature [Watts 1998, Barabási 1999], that grew very rapidly to the point that in a couple of years, articles from this branch became more numerous that those of the traditional branches. This new literature finds its roots in complex systems modeling and for its physics components, in statistical physics. The paradigm of complexity is to study how relationships between parts of a system can give rise to the collective emerging behaviors. These emergent

phenomenon are those that cannot be explained with the separate analysis of each components, which is generally summed up as *a system which is not the sum of its parts*. In statistical physics, this finds an echo in the study of critical phenomena, which describes the power-law divergence of some quantities of a system under very specific conditions. These critical phenomena can only be understood when taking interactions between components into account. It has been found that many systems share a similar behavior near the critical point, i.e., the set of parameter values that determine the very specific conditions in which some quantities of the system diverge. In particular, the set of critical exponents of different chemical compounds at various phase transitions (for example the ferromagnetic phase transition or the liquid gas critical point) may be the same. This observation defines so called *universality classes*. Theoretical models and methods that ignore all of the chemical nature of compounds but that take the physical interactions between the parts into account, are used to describe these mechanisms and can even predict the values of the critical exponents in some cases. This paradigm largely explains why physicists such as Duncan Watts and Albert-László Barabási were intrigued by some apparently generic properties of networks. For example, the model introduced by Barabási and Albert [Barabási 1999] suggests a possible explanation for the frequent presence of scale-free distribution of degrees in various kinds of networks, originally in the World Wide Web network and in scientific citation network. Later on, many other networks were found to share this scale-free behavior, for example email networks [Ebel 2002] and even sexual contacts [Liljeros 2001]. Similarly, the Watts and Strogatz model aimed at giving an explanation of how large networks can be small-worlds.

The second explanation of this recent branch of literature on complex networks is due to the development of computing tools and of the increasing use of huge databases [Lazer 2009, Watts 2007]. Some are specialized databases, such as gene banks, DNA banks, financial trades, and are not life intrusive, but at present, most of our daily acts are systematically recorded in huge databases too. Our journeys are registered via our RFID transport badges, our email exchanges via our email services, our phone calls via our telecommunication subsidiaries, our credit card purchases via our banks, our commercial transactions via the client cards we are kindly offered – or not –, our contacts via Facebook. These multiplication of datasets allows present-day researchers to do quantitative analysis on millions of data, on several temporal and spatial scales and resolutions, which was not even conceivable 50 years ago. Likewise, it changes the scientific approach because it often happens that researchers obtain a dataset before even precisely constructing scientific questions. The problematics become gradually more clear when *playing with a dataset*, i.e., when computing various quantities without having a predefined methodology.

This literature on complex networks that focuses on common non-trivial properties of diverse networks is reviewed thoroughly in [Newman 2003] and [Boccaletti 2006].

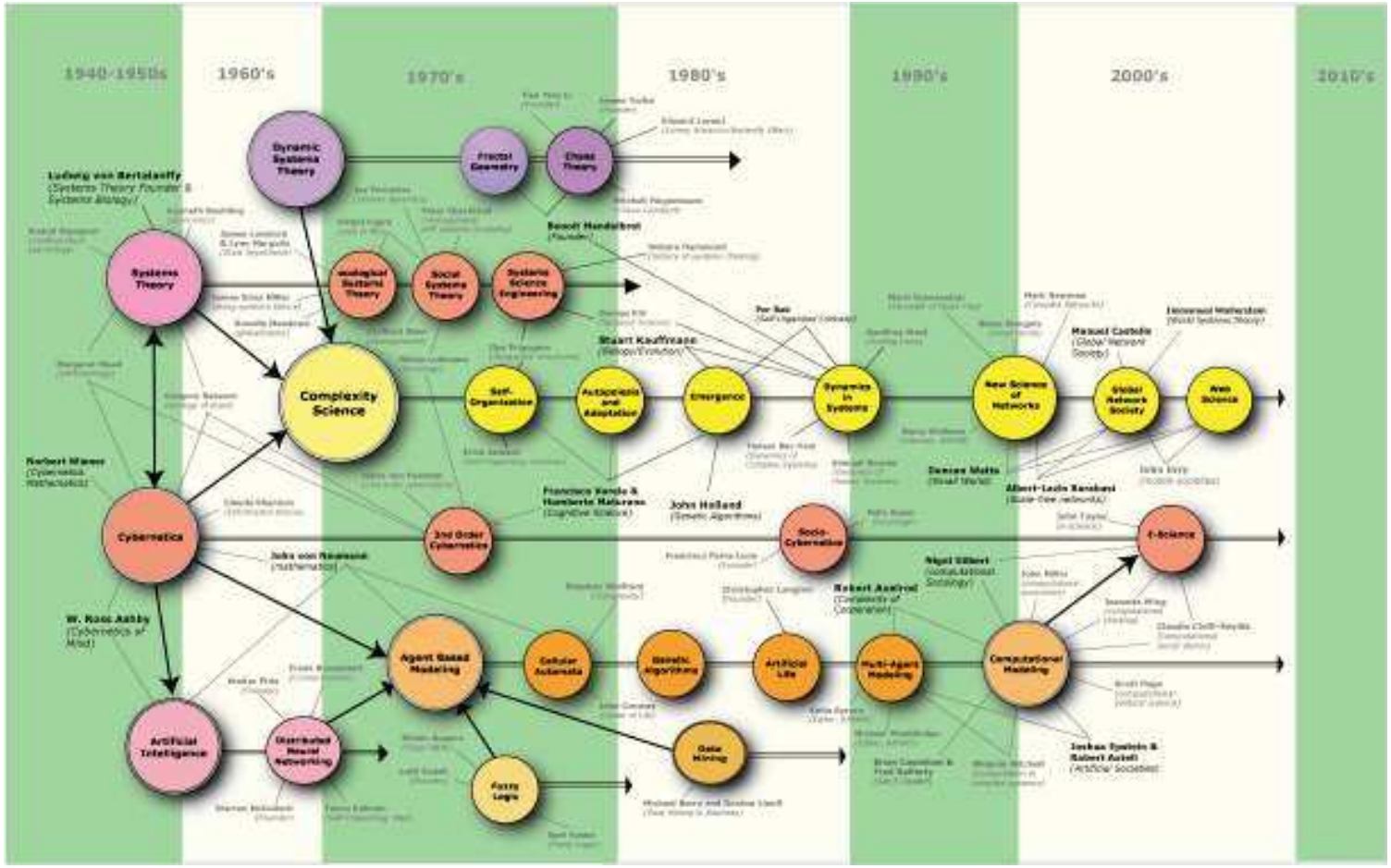


Figure 1.2: Chronology of the main domains of complex systems (figure from Wikipedia).

1.1.2 Dynamical processes on complex networks

Not long after the field of complex networks has been initiated by the Barabási-Albert's and Watts and Strogatz's articles, the research line went on dynamical processes evolving on networks, thus bridging the gap between complex networks and another domain of complex system modeling dealing with dynamical processes. I will briefly mention some of the main dynamical processes that take place on networks, but the interested reader will find much more in [Barrat 2008, Dorogovtsev 2008].

One of the favorite models of statistical physicists is the Ising model, in which particles have a spin that can take only two values, $+1$ or -1. Interaction exists between particles, and an external magnetic field can be introduced and influence spins to be in a specific direction. This model has been considered in the case in which the spins are located on the nodes of a network and analytically solved in equilibrium in some cases. In the case of a scale-free network, it has been shown that the phase transition, that is characteristic of the Ising model (except in the one-dimensional case) and that divides the phase space into an ordered and a disordered behavior, depends on the exponent of the degree distribution [Bianconi 2002]. A slightly modified version of this model is also used to represent social influence, and then goes by the name of the voter model. Other ingredients may be added to improve the phenomenology of social influence, for example when one considers that only individuals having opinions that are not too opposite can influence each other [Holme 2006, Vazquez 2008, Nardini 2008, Kozma 2008]. Similar models that define ordered and disordered phases have been studied. The synchronization of linearly coupled oscillators is, for instance, shown to be more effective in small world-networks than in standard deterministic graphs and purely random graphs [Barahona 2002].

This dependence of phase transition on topology is not specific of the Ising model. For example, the resilience and robustness of networks has been shown to depend on the topology of the network as well [Albert 2000]. This has considerable consequences for technological networks, such as the Internet, which have heavy-tailed degree distributions and thus are very sensitive to targeted attacks. Analogous results are obtained for simple epidemic models. The epidemic threshold of these models, that separates the phase space in a phase in which a disease spreading vanishes after few transmissions, and another phase in which the disease reaches a sizable amount of persons, vanishes in case of highly heterogeneous scale-free networks [Pastor-Satorras 2001]. Another example of model is the random walk on networks. In such a discrete system, a random walker hops from node to node if those are connected by an edge. In the case of an uncorrelated network (where there is no correlation between a node degree and its neighbors' degrees) and a random walker that selects any neighbor with the same probability, the probability that this walker lies, at a given time (when there is no dependence on the starting node), on a node of degree k is proportional to the degree [Noh 2004]. This feature has huge consequences for web search algorithms, such as PageRank, which much rely on such a random walk model.

The main interest of these dynamic process models is to investigate the effect of stylized features of networks on the general evolution of such processes. It generally seems that the topology on the network affects critical phenomenon, by shifting the phase transition, or making it disappear. Most results have been obtained on static networks, and it is likely that the network dynamics could affect these critical phenomenon, as we will see in chapter 5.

1.1.3 Network dynamics

Until fairly recently, networks were mostly considered as static objects with a fixed topology, but most of the time, the system they represent evolves, with the entrance and exit of entities and transformation of relationships between these entities. For example in the case of the airport transportation network, new airports are built, others are closed, traffic connections change with time, with seasons. It also happens that a set of airports becomes inactive in the sense that their traffic is interrupted, as it was the case because of the Eyjafjallajökull eruption in Spring 2010. Because of these changes in the system, one would like to allow a network to evolve, with the creation and disappearance of links and nodes. Such networks are called dynamic networks, evolving networks or temporal networks or graphs.

Since the last few years, a growing body of literature focuses on the statistical analysis and phenomenology of dynamic networks. A first category would be related to digital and communication technologies, such as Facebook [Golder 2007], the Messenger instant-messaging service [Leskovec 2008] or e-mail exchanges [Kossinets 2006, Malmgren 2009]. As said above, the systematic record of digital traces gives researchers massive datasets to analyze. Similarly, phone records [Onnela 2007], mobility traces such as airport fluxes [Gautreau 2009] and individual human travels recorded by mobile phone companies [González 2008] are now analyzed as dynamic networks. More specifically, these works focus on temporal activity patterns, on the relation of one node's temporal rhythm and the one of its neighbors, on temporal characteristics of interactions and on the evolution of network-level properties, may they be scalars or distributions. Some formalization has been proposed in a couple of papers, defining quantities specific to dynamics networks, such as the temporal proximity or the reachability, in the case of directed and undirected networks [Kossinets 2008, Kostakos 2009, Tang 2010, Holme 2012, Pan 2011]. This dynamic perspective makes us understand that many networks that were previously treated as static are in fact the temporal aggregation of temporal sequences of interactions over a certain period of time. In the case of the phone call networks, events are aggregated over typically one year, whilst it is also possible to consider a daily or weekly aggregated network that would evolve over one year.

In parallel to this empirical literature, different theoretical models have been developed. The nature of these models differs in the purposes they serve. A first family of models, sometimes called minimalist models, are aimed at studying how basic interaction rules may induce emerging dynamic patterns. This literature is a

follow-up of complex system models. One example of these models would be the one proposed in [Hill 2010] that suggests simple interaction rules to explain how a system can have a stable power law degree distribution but no continuity in the degree centrality of nodes over time. Another example is described in [Davidsen 2002] that gives local mechanisms of acquaintance creation and deletion producing a network with a small-world structure and triadic closure, characteristic of social networks. The realism of these models is only contained in the plausibility of the pertinent interaction mechanisms in order to reproduce empirical observations.

A second family of models is more related to the statistical core of literature. Their objective is to test on empirical data whether some dynamic mechanisms are statistically significant and to quantify their relative magnitude, as panel regressions do in the econometric literature. For example, is *triadic closure* likely to exist in an empirical dynamic network? Is it more important among girls than among boys? These models are theoretically designed to estimate empirical data. They include the most plausible mechanisms, and the statistical estimation must display which of those mechanisms are likely to exist. This family of models is older than the previous one [Holland 1977] but experience a new revival (see [Goldenberg 2009] for a review).

In these two categories of models, interaction rules are more or less based on rational choices. For example, in [Gräser 2009], a model that belongs to the family of minimalistic models relies on individual actions determined by choice-payoff and expectations. In the second family of statistical models, the model developed in [Snijders 2007] relies on individual preferences that affect the creation or deletion of ties.

A small third category of model of dynamic networks aims at reproducing as many realistic quantities as possible, such as contact / inter contact durations, degree distribution, subgraph structures, etc. To the best of my knowledge, I only know of one model in this category, described in [Scherrer 2008]. The main use of this model would be to simulate realistic dynamic networks, as the statistical models are also able to do. It would then be possible to examine, on a wide range of dynamic networks, the evolution of dynamic processes, and to alleviate the lack of empirical data.

A fourth category of models consists in randomization models. They consist of reference null models to compare empirical data with, in order to identify whether a topological or temporal feature is over- or underrepresented. It is the dynamic extension of configuration models that are used in the case of static networks. The realism of their psychological foundations is not a stake.

The interested reader will find in [Holme 2012] a recent review of the literature on dynamic networks, going through real-world examples, definitions of measures, theoretical models and some examples of dynamical processes on dynamical networks. A less recent book develops rational models within a game theoretic approach, which are omitted in the review [Gross 2009].

1.2 Mining human interactions

During my PhD, I worked on a specific human interaction, namely face-to-face proximity. Various methodologies have been used to record human interactions. It is possible to classify them into two main categories. First, I review the most used pen-and-paper approaches: surveys and self-reported diaries. Second, I will shortly describe the automatic ways of recording interaction proxies, based on a dedicated technological design[1]. Third, the methodology that was used to create the datasets I worked on during my thesis, will be presented. Last but not least, the pros and the cons of this protocol with respect to more traditional approaches are discussed.

1.2.1 Before the technological era

The predominant and, likely, the oldest (dating back to Moreno seminal book [Moreno 1953]) method to investigate interactions among persons relies on surveys and questionnaires. Generally, it consists on one or more name generators followed by name interpreters. Name generators are questions asking for a list of persons, such as *"Who would you ask for advice in case of a personal problem"*. It may be limited in numbers (*name up to X persons*) or not. The name interpreters are questions to add specific information on the named persons (age, gender) and/or on the nature of the tie (duration of acquaintance, frequency of contacts), on the relationships between named persons (do they know each other...). Without considering direct costs (to hire interviewers, recruit participants) this method is time-consuming. The time to pass such a questionnaire ranges from a dozen of minutes to an hour, in case of multiple name generators and name interpreters, to which we need to add the time to fix appointments with the participants. This constraint is particularly important in the case of longitudinal data and limits considerably the number of persons that can be followed in time (see [Lubbers 2010] for recent questionnaire based longitudinal analysis on the integration of 25 Argentinians in Spain). While this was mostly a pen-and-paper approach, computers have been progressively introduced to facilitate the interviews (generally computer assisted personal interviewing, which still requires the presence of interviewers). Progressively, softwares providing a visual representation of networks are used, and increase both the participation ratio and the quality of answers [Hogan 2007].

A second method consists in studying archives. The time-cost of this method greatly depends on the fact that archives are digitalized or not. For example, we can cite the remarkable work on the Medici social network based on thorough work of historians on the Florence family [Padgett 1993]. Some fields mostly rely on archive analysis, such as the study of interlocking directorate (see [Mizruchi 1996] for a review).

Third, direct observation of a small set of people provides an alternative way to study interactions, in a complementary perspective, or when the set of persons can

[1] I do not include studies done on phone calls, mail and e-mail exchanges, that infer social proximity by non-physical interactions.

not be studied via surveys or archive, as it is the case for infants [La Freniere 1984, Martin 2001].

A fourth technique that is similar to survey/questionnaires relies on self-reported diaries. This technique, that is more common in epidemiology than in sociology, is described in more details in a separated paragraph (see paragraph 5.1.4 in chapter 5).

While the survey and questionnaire based studies focus mostly on perceived social ties, the three other methods analyze actual exchanges. Other differences exist, especially concerning the issue of the limited recall of participants that is characteristic of surveys and diaries. The interested reader will find more details on these methods in [Wasserman 1994, Marsden 1990, Marsden 2011].

1.2.2 Using Bluetooth, WiFi and RFID devices

It has been a couple of years since, technological devices have been used in order to measure on a unsupervised manner human proximity. These devices, worn by individuals, are actively emitting and receiving data packets in a limited spatial range. This data exchange is then considered as a proxy of human proximity. I briefly list the datasets I am aware of, with the technology they used. This list is limited to direct device-to-device contact datasets. It excludes access-point based datasets, such as WiFi logs [Henderson 2004, McNett 2005] and records of visible GSM cell towers [Eagle 2006]. The fact that two devices send information to a common antenna effectively means that these devices are in the same area, but given the size of the area, it only corresponds to a physical copresence in the same room/place. Table 1.1 summarizes the main characteristics of the datasets relying on direct device-to-device interaction.

The first technology relies on the Bluetooth proprietary open signal exchange. Some studies have used existing support that enable Bluetooth communication, such as PDA or mobile phones, on which a dedicated software allows this specific use of Bluetooth [Eagle 2006]. Other studies use small wireless portable radio devices, called iMotes, that are lighter than PDA and mobile phones [Chaintreau 2007, Hui 2005, Yoneki 2008, Kostakos 2010, O'Neill 2006]. The temporal resolution on interactions recorded by Bluetooth signals is generally over 2 minutes, except for the Cityware project where it said to reach 5s. The spatial resolution, although not informed systematically in articles, is around 10 meters which should lead to an overestimation of interactions (discussion, physical contact).

A second technology used to record physical proximity relies on Radio-Frequency IDentification (RFID). Only dedicated sensors are used in studies [Salathé 2010, Kazandjieva 2010, Friggeri 2011, Fraboulet 2007]. The time granularity is finer than with Bluetooth motes and varies between 5 seconds and 1 minute. The physical range is also narrower, ranging between 2 and 3 meters. A main advantage of these motes is that it allows one to work with frequencies and power such that the radio signal is shielded by human bodies[2]. Researchers can then capture physical proximity on a non-isotropic fashion, and if motes are carried on the chest for

[2]This advantage has obviously not been exploited in [Friggeri 2011].

Data set	Device	Duration (days)	Granularity (seconds)	Distance (meters)	Number of participants
MIT [Eagle 2006]	Phone (Bluetooth)	246	300	5-10	97
Toronto [a]	PDA (Bluetooth)	M.I.	120	M.I.	23
CAM [Leguay 2006]	iMote (Bluetooth)	11	120	M.I.	36
Cambridge [Chaintreau 2007]	iMote (Bluetooth)	3	120	M.I.	12
Hong-Kong [Chaintreau 2007]	iMote (Bluetooth)	5	120	M.I.	37
INFC05 [Chaintreau 2007]	iMote (Bluetooth)	4	120	M.I.	41
INFC06 [Yoneki 2008]	iMote (Bluetooth)	3	120	M.I.	78
BATH [Kostakos 2010]	PC (Bluetooth)	5.5	Continuous	10	743
D_1 [Takaguchi 2011]	Infrared	73	60	2	163
D_2 [Takaguchi 2011]	Infrared	120	60	2	211
Secondary school [Salathé 2010]	Radio motes	1	60	3	788
HCWs [Friggeri 2011]	Radio motes	98	5	M.I.	56

Table 1.1: Some characteristics on the main types of infrastructures to collect human proximity data at the individual scale (M.I. stands for missing information).

[a] Dataset described in [Chaintreau 2007] without any reference.

example, to detect only face-to-face proximity. Other motes use higher frequencies, i.e., infrared radiation and the characteristics of the datasets is very close to those obtain with standard radio emission [Wakisaka 2009, Takaguchi 2011].

These infrastructures relying on direct device-to-device communication have been deployed in conference [Chaintreau 2007, Yoneki 2008], in school and university [Salathé 2010, Eagle 2006], in an hospital [Friggeri 2011], in a city [Leguay 2006, Kostakos 2010], in offices [Takaguchi 2011]. The interest of studies varies from organizational theory, peer-to-peer device communication, epidemiology and sociology. It is though rare that datasets are analyzed in different perspectives, involving different scientific communities, which would further reduce the time and organization these deployments require.

1.2.3 The SocioPatterns collaboration

In 2008, a collaboration called SocioPatterns started, between researchers and developers from the following institutions:

- the Institute for Scientific Interchange (ISI) in Turin, Italy,
- the Center of Theoretical Physics (CPT) in Marseilles, France,
- the Physics Laboratory of the École Normale Supérieure (ENS) in Lyon, France,
- and Bitmanufaktur in Berlin, Germany.

They developed a protocol to detect physical face-to-face proximity with wearable sensors, based on RFID technology.

The setup, described in details in [Cattuto 2010] is the following. Small RFID badges are worn by individuals on their chest (see figure 1.3 on the left). These badges exchange low power radio packets at close range (about 1 to 1.5m), containing information about the badge identity. They do so by alternating listening and emitting phases, during which they either scan their environment for signal or emit a radio signal. The power of the signal can be tuned in order to control the radius of interaction. As the human body acts as a radio shield because of the water contained in the tissues, no signal is exchanged between badges if individuals are not facing each other. Higher power data packets are then emitted by the badges to RFID antennas located in the premises (see figure 1.3 on the right). Because the radio range of this signal is much higher, only few RFID antennas are needed to cover an entire building. For example, only 15 antennas were used for a deployment in a primary school of 232 children and 10 teachers. These antennas are connected to a Local Area Network and a central computer collects and stores the information sent by antennas. This information consists in the identity of the badge who has sent the information about its neighborhood, the identity of the badges in this neighborhood, the identity of the antenna and the time at which the information has been sent from the badge to the antenna.





Figure 1.3: RFID badge and antenna used in the infrastructure developed in the SocioPatterns collaboration.

By this protocol, an information stream is collected, which can be considered as a good proxy for the evolving face-to-face proximities between the individuals wearing the badges. The low-power radio range of 1 to 1.5 meters is small enough to correspond to a close social interaction, and the strong anisotropy of the signal that is shielded by human bodies allows to consider almost only face-to-face proximity. For example in [Cattuto 2010], a measurement during a conference session is described. While many attendees were present, very few interactions were collected because people were facing the speaker. Finally, the rate at which badges alternate listening and emitting phases is tuned in order to assess a face-to-face proximity between 2 individuals with a probability in excess of 99% over a time interval of 20 seconds.

From the signal recorded by the central computer, one can define a **dynamic network** by aggregating contact signal over a sliding time window of 20 seconds. This duration is long enough to be almost certain that a contact occurs but still short enough to observe a very rich dynamic over an hour. The construction of the dynamic network, defined by its adjacency matrix (A_{ij}) is the following. If badges worn by individuals i and j have exchanged at least one data packet during the time interval $[t, t + \Delta t]$ where $\Delta t = 20$ seconds, then $A_{ij}(t) = 1$ and i and j are said to be in contact. Otherwise, if no radio packet has been exchanged, $A_{ij}(t) = 0$. A **contact** is defined between i and j as an uninterrupted sequence of 20-second intervals during which $A_{ij}(t) = 1$. More precisely, a contact between i and j exists during the time interval $[t_1, t_2]$ if

$$\begin{cases} A_{ij}(t_1 - \Delta t) = 0 & \\ \qquad A_{ij}(t) = 1 & \forall t \in [t_1, t_1 + \Delta t, t_1 + 2\Delta t, \ldots, t_2] \\ A_{ij}(t_2 + \Delta t) = 0 & \end{cases} \tag{1.1}$$

with possibly $t_1 = t_2$. This precise contact definition allows us to statistically characterize the dynamics of encounters (see section 2).

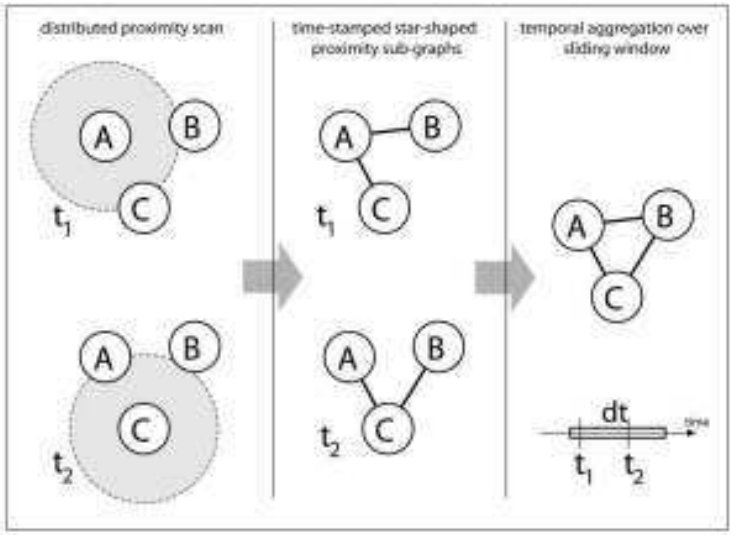


Figure 1.4: Temporal aggregation process by which the radio signal exchanged by badges is transformed into a dynamic network. (Figure from [Cattuto 2010])

The setup has been deployed in different types of social contexts such as an hospital, a primary school, scientific conferences, a museum or in offices. Some characteristics of these deployments are summarized in table 1.2. The participation ratio (i.e. the proportion of volunteers in the population) is in general very good, except for the SFHH conference in Nice, for which only a third of the audience participated to the deployment because of the limited number of devices. This overall good acceptance of wearing technological devices when the purpose of the deployments is explained is encouraging for future work. A relatively lower participation ratio could have been expected from experienced audience such as those of web related conferences (HT and ESWC) because they are highly aware of the issue of digital fingerprint collects, or from less or non-experienced audience because of technological fear (whether the radio signal would be harmful to health or not).

During some deployments, other informations may have been collected on participants. For example, the simultaneous use of the Live-Social Semantic platform, described in more details in section 3.3.1, allowed to cross face-to-face physical proximity with online social networks such as Facebook, Flickr and LastFM. During the school deployment in October 2009, gender, age and class information were provided for a majority of children (more details can be found in section 3.2.2). Last but not least, during the H-Farm deployment which took place in a startup incubator, many information on the structure of the companies (organizational charts, activities) and on each participant (e.g. age, gender, birth place, occupation) has been matched with the badge IDs.

The infrastructure described above is relatively light-weight and once installed, does not need much human supervision (there are still non-avoidable problems, such

as lost badges, badge exchange, discharged batteries). RFID badges are relatively cheaper than bluetooth devices, and require smaller batteries. The infrastructure imposes deployments in relatively small premises, in which antennas must be plugged in a way to cover the entire geographical space. A technological improvement that consists in having embarked memories inside badges is examined. It would allow to avoid the use of antennas and to deploy the infrastructure in open-space areas, but technical synchronization difficulties must be overcome.

1.2.4 Self-reported data vs behavioral data

As with any data collection method, the use of such technologies has advantages and limitations that can be examined with respect to the issues raised by [Marsden 1990] in the context of data collected through questionnaires and surveys. First, while traditional methods are often limited in terms of population size or time of study (at most few waves of few weeks), wearable sensors have already been used to monitor populations of several hundreds of individuals over several weeks [Isella 2010]. This method presents the important advantages of being unsupervised and of relying on unobtrusive wearable badges [Cattuto 2010]. While the direct monitoring of behaviors requires the continuous presence of observers, and questionnaire-based studies require the participation of subjects, once the wearable devices are distributed, no additional human intervention is needed during the data collection. Thirdly, the deployment of the sensing infrastructure is light-weight and does not require training of the investigators. The non-supervised and automatic recording of interactions avoids the difficulty of semantic ambiguity that systematically exists with surveys: there is no unique definition of what a friend is, it depends on many factors such as the culture or the age. For these reasons, re-test studies and comparisons need to be carried out in order to validate trends or to determine which features are context-specific. Finally, the problem of recall bias and cognitive limits is avoided. This advantage is especially important for the study of weak ties [Granovetter 1973]. In [Smieszek 2011], it is shown that the probability that someone forgets a contact lasting less than 5 min exceeds 50% and that this recall probability depends on the total number of persons met during the considered time period.

It is important to note that data collection relying on wearable sensors gives access, from a methodological point of view, to *behavioral* networks defined in terms of spatial proximity, and not to *social* networks. The behavioral proxy we use enables the precise definition of a tie in terms of location and body posture: a tie at time t between a pair of individuals exists if they face each other in close proximity. While in the case of friendship networks it is quite common that a declared friendship is not reciprocated, the behavioral ties we use are by construction reciprocal[3] and accurately time-stamped. These networks, however, are not completely disconnected:

[3] A non reciprocate tie obtained with a multiple name generator may be considered as a different perception of the relationship between the individuals, which can be interesting *per se*, but it can also result from an informant bias (see [Knoke 2008] for a chapter on informant bias) leading to measurement errors.

Date	Venue	Context	Participants		Duration
			Number	Ratio	
Jun 2008	ISI, Torino, IT	Offices	25	$\sim$ 100%	3 weeks
Oct 2008	ISI, Torino, IT	Workshop	51	$\sim$ 100%	3 days
Dec 2008	25C3 , Berlin, DE	Conference	575	30-40%	4 days
Apr-Jun 2009	Science Gallery, Dublin, IE	Museum	11537	$\sim$ 100%	3 months
Jun 2009	ESWC09, Crete, GR	Conference	187	$\sim$ 60%	4 days
Jun 2009	SFHH, Nice, FR	Conference	405	34%	2 days
Jul 2009	HT2009, Torino, IT	Conference	120	75%	3 days
Oct 2009	Primary school, Lyon, FR	School	251	96.4%	2 days
Nov 2009	Bambino Gesù, Rome, IT	Hospital	188	96.4%	10 days
Jun 2010	ESWC10, Crete, GR	Conference	175	$\sim$ 55%	4 days
Apr 2010	Practice Mapping, Gijon, ES	Workshop	100	100%	10 days
Jun-Jul 2010	H-Farm, Treviso, IT	Offices	141	86%	6 weeks
Dec 2010	Hédouard Herriot Hospital, Lyon, FR	Hospital	79	97.5%	5 days
Nov 2011	APS conference, Salt Lake City, US	Conference	320	15%	5 days
Dec 2011	Post secondary classes, Marseilles, FR	School	120	100%[a]	5 days

Table 1.2: Deployments of the SocioPatterns setup with the number of participants, the participation ratio, the type of context and the duration.

[a] Full participation of three classes, which represent only a small part of the entire school.

in particular, it has been shown that children interact four times more with members of their friendship group (identified with the social-cognitive map procedure) than with non-members [Gest 2003]. In this respect, studies that allow to make accurate comparisons between the various ways of capturing social interactions would be of great interest.

1.3 Structure of the following chapters

This thesis hinges on the datasets obtained with the protocol developed in the SocioPatterns collaboration and described above. After a short summary of the main network concepts I use throughout the thesis, the next chapter gives the quantitative description of datasets obtained in a museum, in conferences and in a school. The similarities and dissimilarities between these datasets are considered. A third chapter presents two sociologically oriented studies, one concerns gender homophily among children, the other looks at the relation between physical proximity and the existence of virtual ties on online networking websites. The fourth chapter consists in the presentation of a model that could explain the origin of temporal similarities in contact/intercontact durations. In a fifth chapter, the consequences of contact dynamics on simple epidemiological diffusion models is investigated.

Chapter 2

Statistical analysis of face-to-face proximity data

Contents

2.1 Some useful network concepts . . . 17
2.2 Interactions in different environments: museum and conferences . . . 21
2.2.1 Dynamical burstiness . . . 21
2.2.2 Network static characteristics . . . 22
2.2.3 Distances network . . . 28
2.2.4 Discussion . . . 31
2.3 Primary school . . . 32
2.3.1 School composition and category structure . . . 33
2.3.2 Results . . . 33
2.3.3 Discussion . . . 40
2.4 Partial conclusion and perspectives . . . 44

2.1 Some useful network concepts

In this section, I briefly expose the definitions of the concepts I use in the rest of the thesis. Those are particularly common in the field. For more definitions and details on their specific uses, the reader can refer to the book written by Barrat, Barthélemy and Vespignani [Barrat 2008].

An **unweighted network** (or graph) is usually noted $G(V, E)$, where V is the set of nodes (also called vertices) and E is the set of links (or edges) connecting pairs of nodes. The two simplest quantities to describe a network are the number of nodes $|V|$ and the number of edges $|E|$. They provide information on the network sparsity because the number of possible links is limited by the number of nodes. This maximum bound is reached when all nodes are connected and $|E| = |V|(|V|-1)/2$. However, these quantities generally offer a poor description of the network: two networks with the same number of edges and vertices can be very different.

Two important concepts borne by the degree and the clustering coefficient provide local informations on the network topology. First, the **degree** k_i of a node i

corresponds to its number of adjacent edges, i.e., the number of directly connected nodes (often called *neighbors*):

$$k_i = \sum_{j \in V} a_{ij} = |\mathcal{V}(i)| \tag{2.1}$$

where $A = (a_{ij})$ is the adjacency matrix ($a_{ij} = 1$ if an edge exists between i and j, and $a_{ij} = 0$ otherwise) and $\mathcal{V}(i)$ is the set of neighbors of node i. The degree distribution P_k is often used to characterize the network:

$$P_k(k) = \frac{\sum_i \mathbb{1}\{k_i = k\}}{|V|} \qquad \text{for } k \in \mathbf{N}. \tag{2.2}$$

For example, the degree distribution of a regular lattice is a Dirac mass function because the number of neighbors is always the same[1]. In an Erdős-Rényi network [Erdős 1959], i.e., a random network $G(E, V)$ in which each pair of nodes has the same probability p to be connected, the degree distribution follows a binomial distribution[2]. The degree distribution is so looked at that it is used to define classes of networks. For example, when node degrees are broadly distributed, the network is often said to be scale-free.

Second, the **clustering coefficient** c_i of a node i characterizes the interconnectedness of its neighbors. It is defined as the ratio between the effective number of ties connecting i's neighbors and the possible number of such ties:

$$c_i = \frac{\sum_{\{j_1 \in \mathcal{V}(i),\, j_2 \in \mathcal{V}(i)\}} a_{j_1 j_2}}{k_i(k_i - 1)/2}. \tag{2.3}$$

It ranges from 0 if no tie exists between i's neighbors, to 1 if those are all connected to each other.

Two point correlation functions can be defined to compare local structures. For example, the **cosine similarity** quantifies the similarity between the direct neighborhoods of two nodes, i and j. It is defined as the normalized scalar product of their adjacency vectors $a_{i\bullet} = (a_{i1}, a_{i2}, \ldots, a_{i|V|})$:

$$\mathrm{sim}_{i,j} = \sum_{k \in V} \frac{a_{ik} a_{jk}}{\sqrt{k_i k_j}}. \tag{2.4}$$

If these two nodes share exactly the same set of neighbors, this quantity is equal to 1. If they have no common neighbor, then it is equal to 0.

[1] This regular lattice has to be finite with periodic boundary conditions. Otherwise, the Dirac mass function is only a large size limit of the degree distribution.

[2] A network model is often confounded with its realizations in notations and in the terminology I use. More precisely, a random network model is a collection $\{P_\theta(G), G \in \mathcal{G} : \theta \in \Theta\}$ where $\mathcal{G}$ is an ensemble of possible graphs, P_θ is a probability distribution on $\mathcal{G}$, and θ is a vector of parameters, ranging over possible values in Θ. More details on random graph models can be found in the sixth chapter of [Kolaczyk 2009].

The former quantities can be used for any kind of network, especially when the presence or absence of relation between nodes is the sole information and yet it often occurs that edges are characterized by an attribute, called **weight** (we note w_{ij} the weight between nodes i and j). For example in a social network, the weight can indicate the level of intimacy in the relationship. In that case, the weight is directly imputed from questionnaires containing a rating scale (called Likert-type scale in psychometry): *from 0 to 4, how much do you know this person? (0 for I do not know this person, and 4 for I know him/her very well.* However it can describe very different things: in the cases that I have studied in my thesis, the weight was generally defined as the cumulated duration of contacts between two persons over a limited period of time. In the airport transportation network, it often corresponds to the yearly number of passenger traveling from one airport to the other.

The weighted equivalent of the degree in a weighted network is the **strength**. For node i, it is defined as the sum of adjacent edge weights

$$s_i = \sum_{j \in \mathcal{V}(i)} w_{ij} \tag{2.5}$$

In an aggregated contact network defined as below in section 2.2.2, the strength gives the cumulated duration of interactions of the individual corresponding to node i.

Analogously, the cosine similarity in a weighted network is the scalar product of the normalized weight vectors $w_{i\bullet} = (w_{i1}, w_{i2}, \ldots, w_{i|V|})$ with the convention that $w_{ij} = 0$ if no edge exists between i and j:

$$\mathrm{sim}_{i,j} = \sum_{k \in V} \frac{w_{ik} w_{jk}}{\sqrt{\sum_l w_{il}^2 \sum_l w_{jl}^2}}. \tag{2.6}$$

It is equal to 1 if nodes i and j have the same set of neighbors and with the same weights.

The **participation ratio** $Y_2(i)$ of a node i, also called the Herfindahl-Hirschman index [Herfindahl 1959, Hirschman 1964], characterizes the repartition of weights among its neighbors. It is an index of the local heterogeneity of weights around a node.

$$Y_2(i) = \sum_{j \in \mathcal{V}(i)} \left(\frac{w_{ij}}{s_i} \right)^2 \tag{2.7}$$

The two limit situations are the following: if all adjacent edges have the same weight, $Y_2(i)$ is equal to k_i^{-1} and if one edge weight is much larger than the others, $Y_2(i)$ tends to 1.

All these quantities are often compared between groups of nodes. A rather natural way is to define groups according to the degree, and to compute a node quantity, such as the strength, over all nodes with the same degree. For instance, we can define the average strength of nodes of degree k as

$$\langle s(k) \rangle = \frac{\sum_i s(i) \mathbb{1}\{k_i = k\}}{\sum_i \mathbb{1}\{k_i = k\}} \tag{2.8}$$

If weights are uniformly distributed across nodes, the average strength grows approximately linearly with the degree, i.e., $\langle s(k) \rangle \sim k \langle w \rangle$. If stronger ties are more frequently linked to highly connected nodes, a superlinear behavior is observed [Barrat 2004].

The topology of a network is commonly described in terms of distances. The geodesic **distance** between two nodes i and j, generally noted $\text{dist}(i, j)$, is defined as the minimum number of intermediate edges that need to be traversed in order to go from i to j on the network. It is equal to 1 if an edge exists between i and j, to 2 if no edge exists between i and j but if at least one node k exists at distance 1 from both i and j, and so on. The list of edges between two connected nodes is called a **path**. If no path exists between two nodes, the distance is infinite. In that case, these nodes are in two different connected components (CC). A **connected component** is the set of nodes that are at a finite distance of each other. The size (in terms of distance) of a connected component is generally characterized by its **diameter**, which is the highest geodesic distance among all pairs of nodes.

The very popular six degrees of separation concept is related to these topological notions. When the average distance in a network is very small compared with the number of nodes, the network is said to be a small-world. In 1963, Stanley Milgram experimented this idea in the United States and concluded that every pair of persons in the US is, on average, a few intermediate apart [Milgram 1967, Travers 1969]. His experiment consisted in asking a randomly selected set of individuals to transmit a document to a target person in Massachusetts, whom they did not necessary knew. To do so, they were given some information about the target person in addition to its name, and they had to mail the document to a first-name acquaintance in the purpose that the document would arrive to the target person. Given the experiment design, the conclusions are to be taken carefully. First, the experiment measures the size of a social search chain rather than the smallest social distance between two persons, because people have only an ego-centered view of the social network and they are not aware of the shortest social paths, especially when its real size in larger than 3. Second, the attrition, which is known to be rather high (only 64 out of 296 chains arrived at destination in the original Milgram's experiment), may increase with the path length and lead to an underestimation of the real social search distance. Nevertheless, a recent global-scale experiment with emails instead of mails tends to confirm the numbers given by Milgram when the attrition is taken into account (still with some hypothesis) [Dodds 2003], and the measure of the average distance in large social networks gives a similar order of magnitude [Leskovec 2008].

Little has been done concerning dynamic network measures. Generally the approach is to define a time window over which an aggregated network is constructed. One is generally interested in the time evolution of the quantities defined above. Other quantities such as the temporal distance defined in [Pan 2011] are specific of dynamic networks (see [Holme 2012] for a review).

2.2 Interactions in different environments: museum and conferences

The protocol developed in the SocioPatterns collaboration and described in paragraph 1.2.3 has been deployed in various environments (see table 1.2). In this section, we focus on the deployments at the Science Gallery in Dublin (http://www.sciencegallery.com/infectious) (referred hereafter as SG), at the HT09 conference in Turin (http://www.ht2009.org/) and at the ESWC09 conference in Heraklion (http://www.eswc2009.org/). The two different types of venues (museum and conferences) are compared in order to outline differences and similarities in the interaction patterns, and the possible universality of these results will be discussed.

These venues were chosen for their very difference of nature and characteristics. First they involved vastly different numbers of individuals and stretched along different time scales. The SG venue lasted for about three months and recorded the interactions of more than 14,000 visitors (more than 230,000 face-to-face contacts recorded), whereas the HT09 took place over the course of three days and involved about 100 conference participants (about 10,000 contacts) and ESWC09 involved 175 voluntary participants over the same time period (about 15,000 contacts). Behaviors are also very different: in a museum, visitors typically spend a limited amount of time on site, well below the maximum duration permitted by the museum opening hours, they are not likely to return, and they follow a rather pre-defined path, touching different locations that host the exhibits. In a conference setting, on the other hand, most attendees stay on-site for the entire duration of the conference (a few days), and move at will between different areas such as conference room, areas for coffee breaks and so on.

The present section describes work done in collaboration with Lorenzo Isella, Ciro Cattuto and Wouter Van den Broeck of the ISI Foundation, my advisor Alain Barrat and Jean-François Pinton from the ENS Lyon. Some parts of this work have been published in [Isella 2010] and in [Zhao 2011].

2.2.1 Dynamical burstiness

A very robust observation on contact dynamics measured by wearable sensors concerns their temporal heterogeneities. In this section, results are presented for ESWC09 only, but the same behavior was observed in all other datasets.

Contact durations The first heterogeneity is given by contact durations. Those are defined as the time interval elapsed between the beginning and the end of a continuous sequence of radio packet exchange, at the 20 second temporal resolution. The duration is then a multiple of 20 seconds, and the minimal value is 20 seconds. These contact durations are broadly distributed. Figure 2.1 (left) shows the distribution of contact durations measured at the ESWC09 conference. We use a partially log-binned representation that is very common way to represent distri-

butions over several orders of magnitudes (the interested reader can find technical details in the appendix of [Pastor-Satorras 2004]). This distribution is heavy-tailed, spanning over 2 orders of magnitude on the x-axis but almost 6 orders of magnitude on the y-axis. It means that even if a contact lasts on average 46s, this average quantity does not summarize adequately the distribution because many other contacts may last several times longer. Neither the average nor any other time scale (like the standard deviation) emerge from the distribution because of its shape: it is said to have no characteristic scale.

Relay time intervals In [Miritello 2011], relay time intervals are defined for instantaneous interactions (they study mobile phone calls and do not consider their durations). A relay time interval is defined as the time interval between the start of two consecutive contacts of a given individual A with two distinct persons B and C. If A starts a contact with B at time t_{AB} and starts a successive contact with C at time t_{AC}, then the relay time is $\tau = t_{AC} - t_{AB}$, independently of the fact that the first contact with B has ended or not. This quantity is relevant for diffusion processes such as information diffusion or infectious disease spreading because it constrains the temporal evolution. At the ESWC09 conference (see right panel of figure 2.1) as in other environments (not shown), relay time intervals are broadly distributed. A peak at 12 hours gives a characteristic time scale that corresponds to the time interval between the end of one conference day and the beginning of the sessions on the consecutive day. As the setup does not allow to register interactions occurring outside of the conference premise, each participant contributes to this peak once if s/he stays two days at the conference, and twice if s/he stays three days.

Group durations Unlike phone calls which generally happen between two persons, more persons may interact at the same time. We define the duration of a group of of size $p+1$ ($p = 0$ for isolated individuals, $p = 1$ for a pair ...) as the time interval elapsed between the time a group of this size is formed and the time a person leaves or joins it. Figure 2.2 shows the distribution of group durations. Again, broad distributions are observed. The slopes of the distributions plotted with a logarithmic scale increase with p, meaning that larger groups have a shorter average lifetime.

2.2.2 Network static characteristics

We face a dynamic system of interacting individuals, with generally an entrance and exit of persons. Such a system can be considered as a dynamic network, in which individuals are represented by nodes, and an edge connects two nodes at a given time if the corresponding persons are in contact. The network is said to be dynamic, because edges appear and disappear with time, and, because all individuals are not necessarily present during the entire deployment, nodes enter and leave the network over time. At the temporal resolution scale of 20 seconds, the network

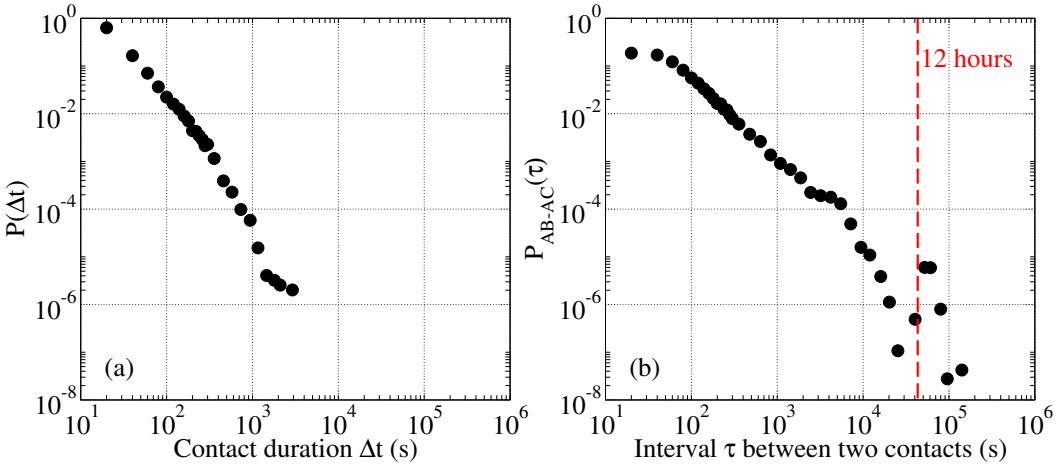


Figure 2.1: Distribution of contact durations (a) and of relay times (b) at ESWC09.

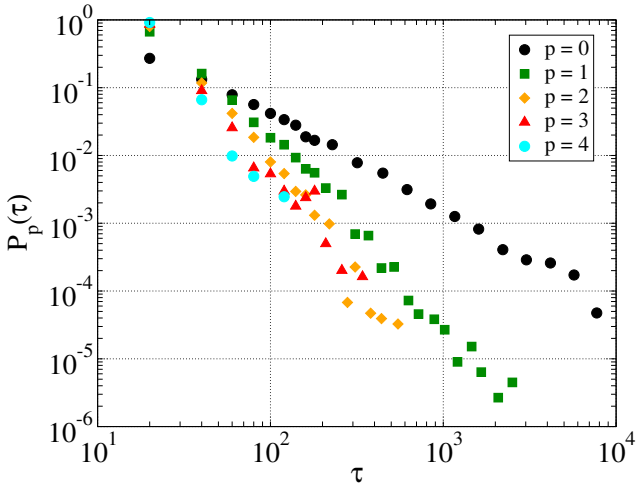


Figure 2.2: Distribution $P_p(\tau)$ of the duration in seconds of groups of size $p+1$ at ESWC09.

consists mostly in isolated nodes and dyads (pairs of connected nodes), occasionally of larger groups of people. Unlike for static networks, few typical observables are commonly defined for this kind of system. A solution is to aggregate interactions over a time window to construct a weighted network and use usual metrics. Weights are generally defined as the cumulated duration of contacts between two persons over the time window, but other quantities could be used, such as the number of distinct contacts.

In this section, we focus on the analysis of daily aggregated networks. The choice of daily time window is rather natural in our settings. This time scale is common for constructing networks of interacting people based on declarative data. For example, each participant would declare who s/he has encountered during the course of the day. Longer aggregation periods such as weeks or months would be interesting to investigate the stationarity of the collected data. Shorter aggregation times of the order of a few minutes are also useful, for instance, to distinguish the daily rhythm in various environments, such as the succession of breaks and lunch during a school day.

Connected components and diameter Figure 2.3 displays the aggregated contact networks for June 30th at the HT09 conference (top left), and for three representative days for the museum deployment. Despite the large variation in the number of daily museum visitors, ranging from about 60 to 400, the chosen days illustrate the topology of the museum aggregated networks, in particular the presence of either a single or two large connected components in the network. Days with smaller numbers of visitors can also give rise to aggregated networks made of a larger number of small isolated clusters. As shown in figure 2.4, depending on the number of visitors, the number of connected components can in fact vary substantially. For a large number of visitors, typically only one connected component is observed. For a low number of visitors, on the other hand, many clusters are formed. In the case of conferences, only one connected component is observed. This has to be related to the fact that the main purpose of conferences is the meeting of scientists. The presence of several connected components would be rather worrying otherwise. Because of the behavioral differences between a museum visitor and a conference attendant, the topological difference between the aggregated networks is outlined by the network diameters (highlighted in all the plots of figure 2.3), which is considerably larger for SG than for HT09 aggregated networks. This is mainly due to the fact that visitors stay in the museum only for a rather small duration and two visitors interact with each other at least if they are in the museum at the same time. The correlation between the diameter and the time of the visit is analyzed below in section 2.2.3.

Degree distribution An extensively reported result in the literature of complex networks concerns the degree distribution. For several social networks, for example for sexual networks in [Liljeros 2001], for scientific collaboration in [Newman 2001]

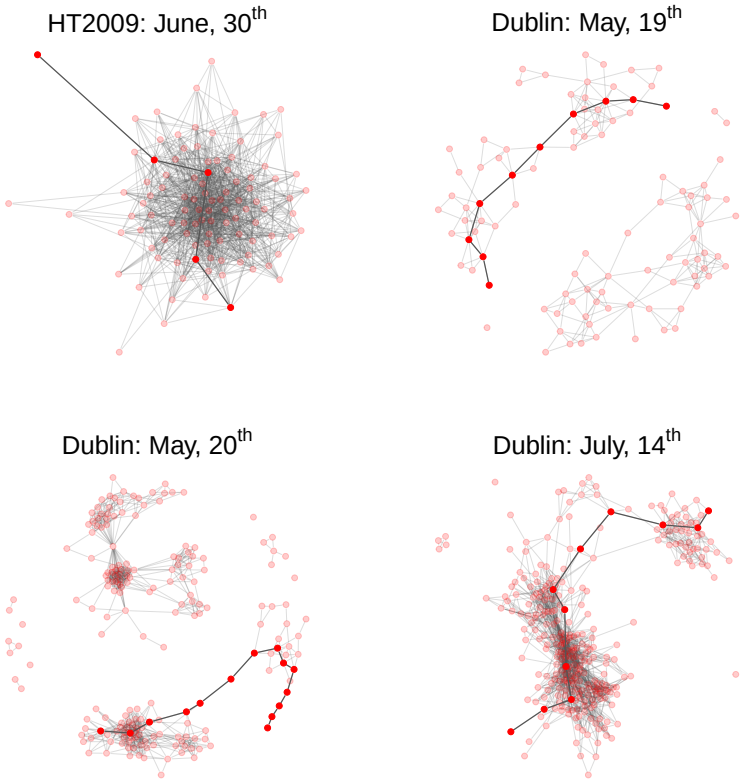


Figure 2.3: Daily aggregated networks in the HT09 and museum deployments. Nodes represent individuals and edges are drawn between nodes if at least one contact event was detected during the aggregation interval. Clockwise from top: aggregated network for one day of the HT09 conference, and for three representative days at the museum deployment. In each case, the network diameter is highlighted.

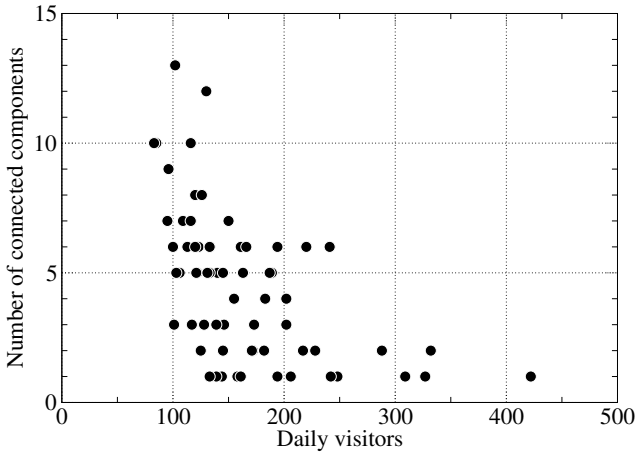


Figure 2.4: Number of connected components in the daily aggregated networks of the museum deployment with respect to the number of visitors.

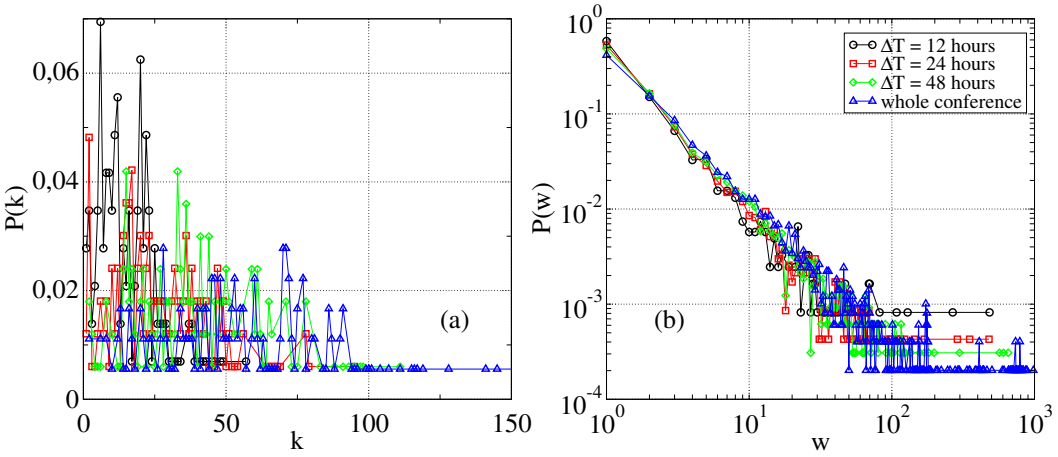


Figure 2.5: Distribution of (a) degree and (b) weight on aggregated networks on various time windows at ESWC09.

and for phone calls in [Onnela 2007] to name a few, the degree distribution is found to be heavy-tailed, meaning that some people have a far larger number of neighbors than the average. This feature was found in networks that are very different from each other and Barabási and Albert proposed a simple model that gives a generic mechanism that would explain such degree distributions [Barabási 1999]. In the daily aggregated contact networks, this feature is not observed. The degree distribution, given by figure 2.5 for the ESWC09 conference (left), is narrowly distributed. Compared to the social networks given as examples above which are defined over a large period of time (more than a year), the contact networks we study are defined on a much smaller timescale and on a small population size. If we change the aggregation window from half a day to the entire conference duration (three days), the average degree increases and higher degree values are reached (see figure 2.5, left panel) but the time scale is still too small to exhibit a heavy tailed behavior. It is possible that if contacts were recorded over a long time period such as a year and on a large population (at least 1000 persons) the degree distribution of the corresponding yearly aggregated network would be heavy tailed.

Weight distribution It is not possible to identify broad distributions for degrees because there are too few participants in the deployments. In the case of ESWC09, there are 175 nodes and degrees are integers between 1 and 174 by definition and a rule of thumb is to consider at least 3 orders of magnitude over the x- or the y-axis to identify broad distributions. In the case of distributions defined over edges, the number of participants does not limit that much the identification of broad distributions (there are at most $N(N-1)/2$ edges in a network with N nodes, which exceed 30.000 in the case of ESWC09 although the majority of these edges do not exist). In all settings, the weight distribution of daily aggregated networks is indeed broadly distributed (it can be objected that the span of the distribution is still rather small, but the logbinning of data helps to deal with the trumpet like outlook of the tail and logbinned data generally span over one to two additional orders of magnitude). Furthermore, as illustrated by the right panel of figure 2.5 for ESWC09, the shape of the weight distribution is independent of the time window over which the aggregated network is constructed. This result is reminiscent of the distribution of contact durations, which are broadly distributed as well, but is not a direct consequence of it. If contacts happened only between the same pairs of persons, the weight distribution could have had almost no contribution on small weights.

Strength and participation ratio When we average measures on nodes that have the same degree (see section 2.1 for more details), the results offer some insight into correlations between nodes and edges. Here, the linear or slightly faster behavior of the average strength $\langle s(k) \rangle$ of nodes of degree k versus k highlights a weak correlation between weights and degree (see left panel of figure 2.6 for ESWC09). The more neighbors these nodes have, the longer the corresponding persons interact.

The correlation does not obviously change with the span of the aggregation window. Moreover, the positive correlation between the degree and the average participation ratio (average over nodes of same degree) multiplied by the degree $\langle kY_2(k)\rangle$ shows that the heterogeneity of weights connecting a node to its neighbors (see right panel of figure 2.6 for ESWC09) increases with the degree. Again, this correlation is not deeply modified by a change of the aggregation window size.

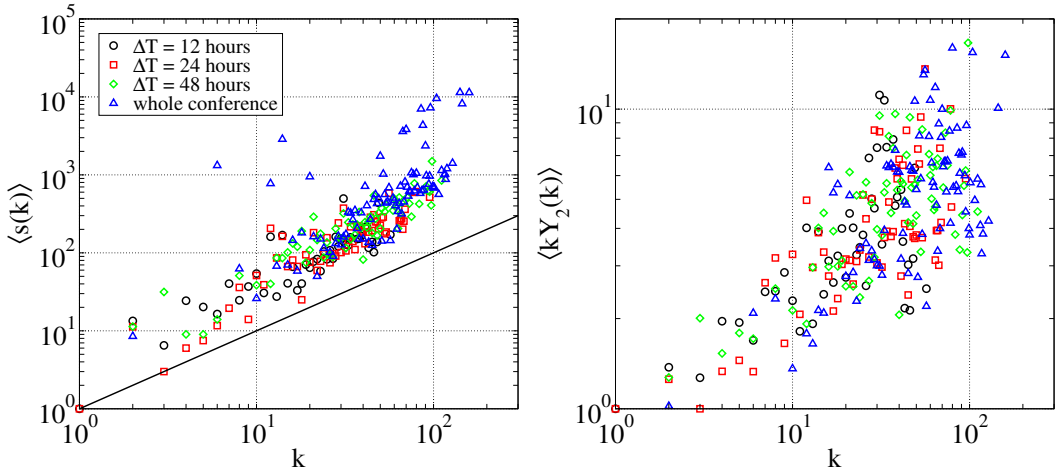


Figure 2.6: Average strength of nodes of degree k vs k (left), and average participation ratio of nodes of degree k multiplied by k vs k (right) for aggregated networks on various time windows at ESWC09.

2.2.3 Distances network

Small-world nature of the aggregated networks The small-world nature – or lack thereof – of these aggregated networks can be investigated statistically by introducing a proper null model. To this end, we construct a randomized network using the rewiring procedure described by [Maslov 2004]. The procedure consists in taking random pairs of links $(i-j)$ and $(l-m)$ involving four distinct nodes, and rewiring them as $(i-m)$ and $(j-l)$ (or as $(i-l)$ and $(j-m)$) if none of these links already exist. This procedure preserves the degree of each node and the degree distribution $P(k)$, while destroying the degree correlations between neighboring nodes, as well as any other correlations linked to node properties. The procedure is carried out so that initially distinct connected components do not get merged. Since the rewiring procedure cannot be implemented for the rare connected components with less than four nodes, these small connected components are removed from the aggregated networks before rewiring. The rewired version of the aggregated HT09 network is very similar in terms of distances to the original version, whereas the null model for the aggregated network of the SG data on July 14th is more *compact* than the original network and exhibits a much shorter diameter. Similar considerations hold for the other aggregated networks of the SG deployment.

More quantitatively, we measure the mean number of nodes one can reach from a randomly chosen node by making l steps on the network, a quantity hereafter called $M(l)$. For a network consisting of a single connected component, the definition of $M(l)$ implies that

$$\begin{aligned} M(1) &= \langle k \rangle + 1 \\ M(\infty) &= N \end{aligned} \tag{2.9}$$

where $\langle k \rangle$ is the average node degree, N is the total number of nodes in the network and $M(\infty)$ the saturation value of M on the network. The saturation value $M(\infty)$ is reached when l is equal to the length of the network diameter, and may vary for different realizations of the random networks. For a network consisting of several connected components, one has to take into account the probability N_i/N that the chosen node belongs to a given connected component, whose node number is noted N_i. As a consequence, the previous equations generalize to

$$\begin{aligned} M(1) &= \frac{1}{N} \sum_i N_i \left(\langle k \rangle_i + 1 \right) \\ M(\infty) &= \frac{1}{N} \sum_i N_i^2 \end{aligned} \tag{2.10}$$

where $\langle k \rangle_i$ is the average node degree on the i-th connected component. This ensures that the quantity $M(l)/M(\infty)$, regardless of the number of connected components, assumes the same value when $l = 1$, and saturates to unity for both the aggregated and rewired network.

Figure 2.7 displays $M(l)/M(\infty)$ for the aggregated networks of HT09 and of the museum deployment on July 14, as well as its value averaged on 100 randomized networks (the average value of $M(l)$ converges rapidly already when calculated on a few tens of randomized networks). We notice the striking similarity between the results for the HT09 original and randomized networks, where about 90% of the individuals lie, in both cases, within two degrees of separation. In the museum case, conversely, the same 90% is reached with six degrees of separation for the original network, but with only three degrees of separation on the corresponding randomized networks. The same calculation, performed on other days of the museum deployment, yields qualitatively similar results, always exposing a dramatic difference from the null model.

Diameter As mentioned above, the elongated aspect of the aggregated networks of visitor interactions (see figure 2.3) is related to the existence of a limited visit duration. Indeed museum visitors are unlikely to interact directly with other visitors entering the venue more than one hour after them, thus preventing the aggregated network from exhibiting small-world properties. Figure 2.8 reports the museum aggregated networks for two different days, where the network diameter is highlighted and each node is colored according to the arrival time of the corresponding visitor (this information is given by the first time an antenna detects a new RFID). This

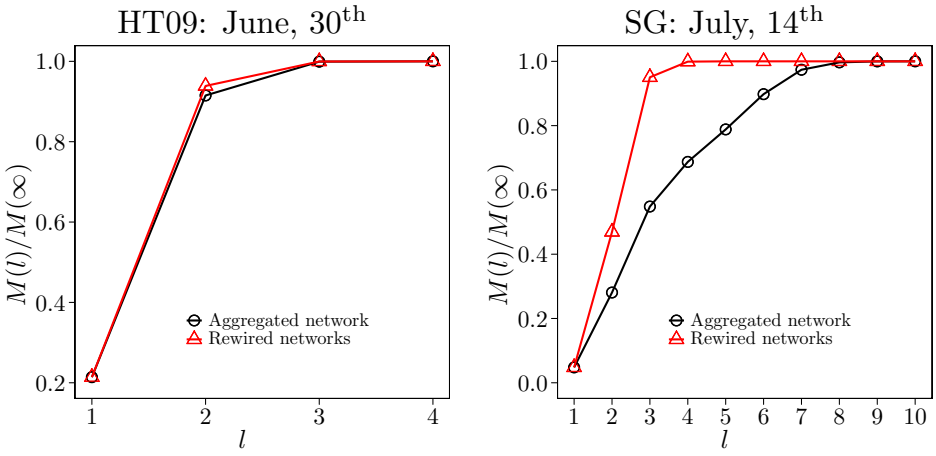


Figure 2.7: Average number of nodes reachable from a randomly chosen node by making l steps on the network, $M(l)$, divided by its saturation limit $M(\infty)$, for daily aggregated networks (circles) and their randomized versions (triangles). For the randomized case, data are averaged on 100 realizations. 90%-confidence intervals give intervals smaller than symbol size. Left: network aggregated on June 30th for the HT09 case. Right: museum deployment, July 14th. The solid lines are only guides for the eye.

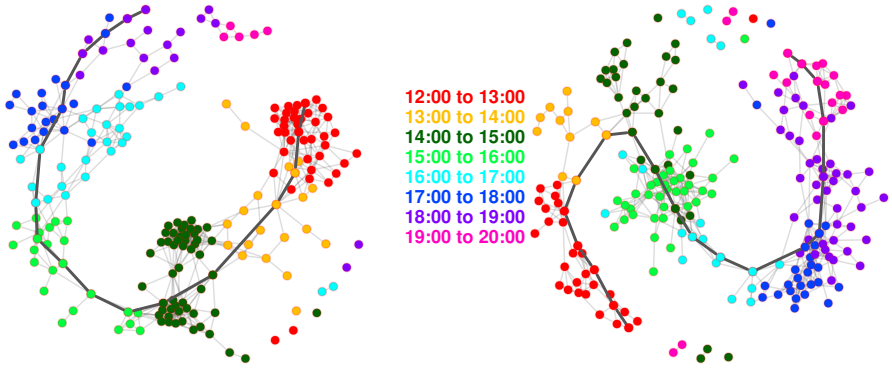


Figure 2.8: Aggregated networks for two different days of the SG museum deployment. Nodes are colored according to the corresponding visitor's entry time slot. The network diameter is highlighted in each case.

figure clearly shows that interactions among visitors entering the museum at different times are very limited. Besides, the network diameter clearly defines a path connecting visitors that enter the venue at subsequent times, mirroring the longitudinal dimension of the network. These findings show that aggregated network topology and longitudinal/temporal properties are deeply interwoven.

2.2.4 Discussion

The analysis of time-resolved data on face-to-face interactions can be carried out within a network perspective. The two types of considered data, i.e., a museum and conferences, are chosen here for their important difference of nature: a conference is a *closed* system in which scientists gather and interact in a repeated fashion, while a museum is an *open* environment in which visitors enter and leave, as a flux of individuals through the premises.

As expected, the difference between these two types of venues has observable consequences in the results of our analysis. The daily aggregated conference networks are rather dense small worlds, while the museum networks have a larger diameter and are possibly made of several connected components, mainly because individuals enter the museum at different times and remain visiting only for a limited duration. The deep intricacy between the topology and the temporal properties may be important for diffusion processes evolving at a temporal scale close to the temporal scale of contacts. This hypothesis motivates a chapter of this thesis on the study of how spreading processes occur on dynamical networks (see chapter 5).

In spite of these unsurprising results, our analysis shows that the behavior of individuals in conference and in a museum setting exhibits unexpected similarities. The distribution of the contact event durations, of the time intervals between two contacts, of the duration of groups, of the total time spent in face-to-face interaction by two individuals are very similar. The broad nature of these distributions is not intuitive as well. It has been observed in other contexts such as for emails, instant messaging exchanges, phone calls [Eckmann 2004, Rybski 2009, Onnela 2007, Leskovec 2008]) that inter-event times are broadly distributed. This is the case too for face-to-face interactions measured by our setup, and furthermore we show that interaction durations exhibit the same behavior. This finding motivates a model presented in chapter 4.

A possible critic to this analysis would be that the types of datasets given as examples here (a museum and scientific conferences) are of limited scientific interest. The environments in which contacts are recorded are very structured and do not allow to learn much on the social drivers of interactions. The most important result concerns the shape of duration distributions, because most of the other measures only quantify patterns that are rather expected in these specific contexts. In my opinion, this is only partially true, because the main interest of this work is to present methods to analyze time-resolved contact networks. Today, the setup is limited to register contacts in predefined premises (because of antennas) but a next infrastructure is under current development, in which the wearable sensors will

record directly the information on contacts on a memory card. With this new generation of sensors, it will be possible to measure contacts among a defined set of people in any environment. In that case, the methods we propose will give some results that can not be known in advance. For example, can anyone characterize the topology of a daily aggregated contact network in a residential building? on a university campus? in a firm? among family members? I dare believe that these questions will interest a larger community of researchers, ranging from organizational scientists to epidemiologists.

2.3 Interactions among individuals of defined categories: the case of a primary school

The protocol developed in the SocioPatterns collaboration has been deployed in Fall 2009 in a primary school. From a scientific perspective, schools are more interesting than conferences and museums and the analysis of this dataset highlights the rich potential of the methodology. First, schools are of deep interest for epidemiologists. Because of the numerous close contacts occurring among children and because children are the first to be infected and transmit the disease to the households, schools are crucial for the spreading of diseases (see [Longini 1982, Viboud 2004]). The collaboration with epidemiologists of the Hospital Edouard Herriot has indeed motivated the study and they have largely contributed to the deployment and to the analysis of results. Second, sociologists and developmental psychologists are interested too in the behavior of children at school. These scientists are mainly interested in understanding how different variables such as the gender, the socio-economic status of their family, the ethno-racial markers shape the behavior and the structure of social groups of children. One of these aspects related to gender homophily is developed in chapter 3.

The present analysis focuses on the analysis of a time-resolved contact network in which categories of nodes are explicitly defined: grade, class, children vs teachers, boys vs girls. This has been done in different contexts with different categories. For example in [Isella 2011], contacts are analyzed in a pediatric hospital. In their situation, these categories correspond to different roles (nurses, medical doctors, family, children).

Most of the present section corresponds to a published work in PlosOne [Stehlé 2011b]. This work was done in collaboration with Nicolas Voirin, Corinne Régis, Bruno Lina and Philippe Vanhems from the University Claude Bernard of Lyon (and from the Hospital Edouard Herriot), with Ciro Cattuto, Lorenzo Isella, Marco Quaggiotto and Wouter Van den Broeck from the ISI Foundation in Turin, with my advisor Alain Barrat and with Jean-François Pinton from the Ecole Normale Supérieure in Lyon.

2.3.1 School composition and category structure

The primary school is composed of 10 classes, divided in 5 grades, labeled 1A, 1B,..., 5A, 5B (two classes for each grade). The age of children ranges between 6 and 12 years. The data were collected over two school days, from 8:30am to 5:15pm. Only interactions taking place on the school grounds were recorded. It is worth noting that slightly more than one child in three leaves the school premises for lunch, which leads to a relative drop of activity during the lunch break. All of the 10 teachers and 96% of the children (232 out of 241) took part in the data collection. The 9 remaining children were either missing on both days or received a badge that was defective and had to be removed from the dataset.

Each individual is uniquely associated with one wearable badge and, through that, to a unique numeric identifier. The identifier is only associated with anonymous metadata for each individual: school class, gender, year and month of birth. All the statistical treatment of the data is performed in an anonymous way. The metadata was collected for 227 out of 232 participating students. The difference is accounted for by participants whose badge was accidentally replaced during the deployment, breaking the connection between the badge identifier and the participant metadata. The sample is restricted to this subpopulation of 227 children (94% of the children in the school). We can reasonably assume the absence of any selection bias, as the exclusion from the studied population is not related to gender or behavior.

Class	Number of individuals	Participants Day 1	Day 2
1A	24	22 (B: 10, G: 11, U: 1)	23 (B: 10, G: 11, U: 2)
1B	25	25 (B: 12, G: 13)	25 (B: 12, G: 13)
2A	25	22 (B: 8, G: 14)	23 (B: 9, G: 14)
2B	26	25 (B: 11, G: 14)	26 (B: 11, G: 15)
3A	24	23 (B: 14, G: 9)	23 (B: 14, G: 9)
3B	22	21 (B: 10, G: 11)	21 (B: 10, G: 11)
4A	23	21 (B: 11, G: 8, U: 2)	21 (B: 11, G: 8, U: 2)
4B	24	22 (B: 13, G: 9)	22 (B: 12, G: 10)
5A	24	22 (B: 11, G: 10, U: 1)	21 (B: 10, G: 10, U: 1)
5B	24	23 (B: 12, G: 11)	23 (B: 12, G: 11)
Teachers	10	10	10

Table 2.1: Participant repartition among class and gender (B for boys, G for girls and U when the gender metadata is not available).

2.3.2 Results

We present here the results of our quantitative analysis. As averaged values do not reflect adequately dispersed data, the coefficient of variation squared (CV^2)

is systematically reported with averages. It is defined as the square of the ratio between the standard deviation σ and the mean μ.

We recorded a total of $77,602$ contact events involving 242 individuals ($37,414$ contacts on day 1 and $40,188$ on day 2), with an average of about 317 contacts per individual on the first day ($\mathrm{CV}^2 \sim 0.22$) and 338 contacts per individual on the second day ($\mathrm{CV}^2 \sim 0.27$).

2.3.2.1 Temporal evolution

The school day runs from 8.30am to 4.30pm, with a lunch break from 12pm to 2pm, and two breaks of 20–25 min around 10.30am and 3.30pm. Lunches are served in a common canteen, and a shared playground is located outside the main building. As the playground and the canteen do not have enough capacity to host all the students at the same time, only two or three classes have breaks at the same time, and lunches are taken in two consecutive turns.

Figure 2.9 shows the time evolution of the 20-minute aggregated networks for each day of the study. The number of individuals is stable during teaching hours, morning and afternoon breaks, and drops during the lunch time as some children were going back home to have lunch (the school is located in an urban area and many children actually live nearby and can go home for lunch). The average degree of the network displays a more interesting behavior, as it peaks at various moments, each corresponding to a break or the beginning or end of the lunch of a series of classes.

Thanks to the information on classes, we distinguish the degree and the strength between contacts with children in the same class (k_i^{in}, s_i^{in}) and with children of different classes (k_i^{out}, s_i^{out}). Figure 2.10 displays the time evolution of the quantities k_i , k_i^{in} and k_i^{out}, averaged over all children, since the beginning of the deployment. The average number of distinct persons contacted grows initially rapidly, mostly because of contacts occurring within each class. The average k_i^{in} however saturates at the average class size after a few hours, meaning that each child has been in contact with all members of his/her own class, while new contacts across classes occur only during the breaks, leading to plateaus in the evolution of the cumulated average degree. In the second day, contacts within each class are the same as in the first day, and the average of k_i continues to evolve only during the breaks due to contacts involving children of different classes that had not occurred on the first day. At the end of each day, each individual, on average, has been in contact with 50 distinct individuals ($\mathrm{CV}^2 \sim 0.14$) in day 1 and with 46.5 individuals ($\mathrm{CV}^2 \sim 0.18$) in day 2.

Figure 2.11 gives more insight into the evolution of the contacts of the children by taking into account the cumulated time spent in contact. It shows the time evolution of s_i , s_i^{in} and s_i^{out}, averaged over all children. The average contact time spent by a child with other children grows regularly with time, in a similar way in both days. While the time spent with other children of the same class also has a regular increase (only slightly faster during morning and afternoon breaks), the time

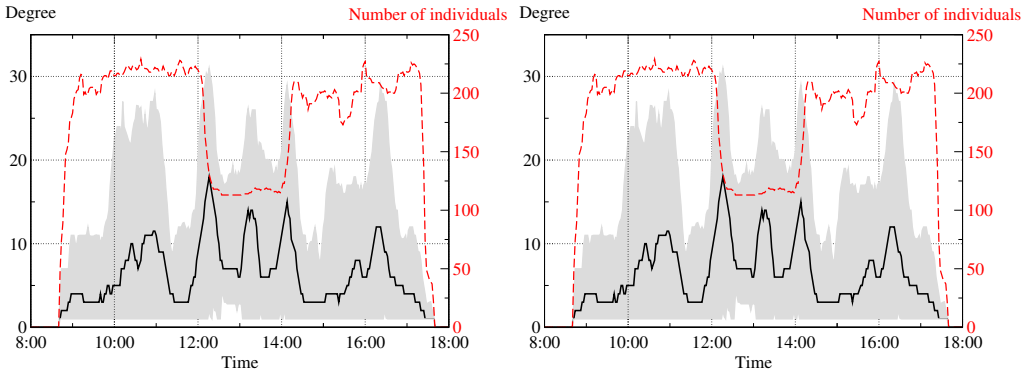


Figure 2.9: Degree of individuals in the contact networks aggregated over sliding time windows of 20 minutes during the first day (left) and the second day (right) of data collection. The median value is represented with a black line, the 95% confidence interval is shown in gray and the number of individuals over which the statistics are calculated is shown in red dashes. Breaks and beginning and end of lunch are characterized by a sudden increase of the degree, showing the occurrence of large numbers of contact events.

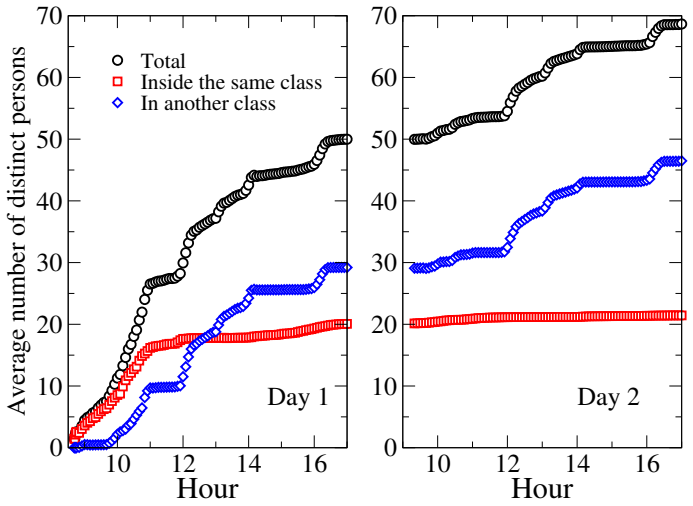


Figure 2.10: Time evolution of the average number of distinct children with whom a child has been in contact. The average total number is displayed in black, the average number of children of the same class in red, and the average number of children of other classes in blue.

spent with children of a different class evolves significantly only during the lunch break (the evolution occurring during the morning and afternoon breaks are much smaller). Overall, at the end of each day, a child has spent on average three times more time in face-to-face proximity with children of his/her class than with children of other classes.

2.3.2.2 Aggregated network

Figure 2.12 displays the daily aggregated contact network for the first day of the study. For ease of interpretation, edges between individuals who spent together a cumulated time smaller than 2 minutes during the day have been removed. This corresponds to keeping only the strongest 33.2% of all edges. The figure highlights the mixing patterns between children of different classes and how children preferentially mix within the same class or age group. Classes within the same grade tend indeed to be more connected than classes belonging to different grades.

2.3.2.3 Comparison between day 1 and day 2

A comparison between the characteristics of the overall face-to-face contact patterns in the two days of the deployment is reported in table 2.2. Statistical quantities such as the average total number and durations of contacts, the number of different persons contacted, or the contact durations are extremely close across the two days.

Network characteristics	Day 1		Day 2	
Number of individuals	236		238	
Average number of contacts of an individual (CV^2)	317	(0.22)	338	(0.27)
Average total time in contact of an individual, in minutes (CV^2)	172	(0.25)	183	(0.33)
Average number of distinct persons contacted (CV^2)	50.0	(0.14)	46.5	(0.18)
Average cumulated time spent in contact by two persons, in seconds (CV^2)	207	(5.4)	236	(4.7)
Average duration of a contact, in seconds (CV^2)	32.6	(1.2)	32.6	(1.1)
Average clustering coefficient	0.50		0.56	

Table 2.2: Comparison of some characteristics of the networks of day 1 and 2.

At a more detailed level, the Pearson correlation coefficients between the number of contacts of an individual in the first and second day is 0.53; for the time spent in contact, it is 0.54; for the number of distinct persons contacted it is 0.53. These values show an overall strong correlation between the behavior of individuals from one day to the next.

Moreover, each child, on average, has 26 repeated contacts on the second day with children met during the first day (19 in the same class and 7 in a different class), and new contacts with 20 other children (1.4 in the same class, 18.4 in a different class). The average cosine similarity between his/her neighborhoods across the two days is 0.67 (0.74 for the neighborhood restricted to his/her own class, 0.2 for the

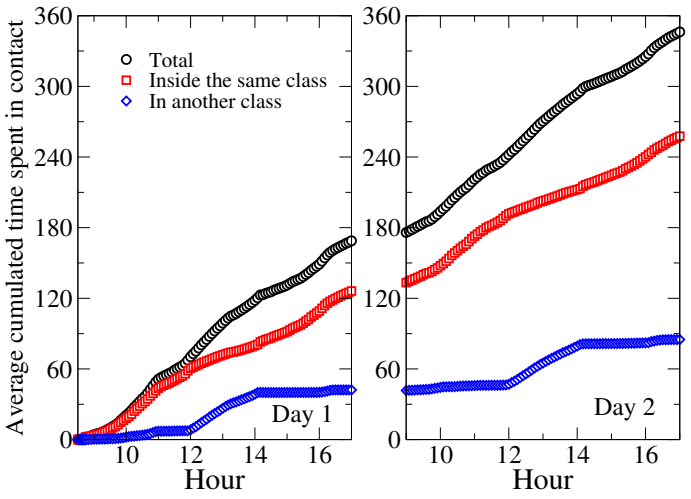


Figure 2.11: Time evolution of the average cumulated time spent by a child in contact with other children. Colors and lines are the same as in the previous figure.

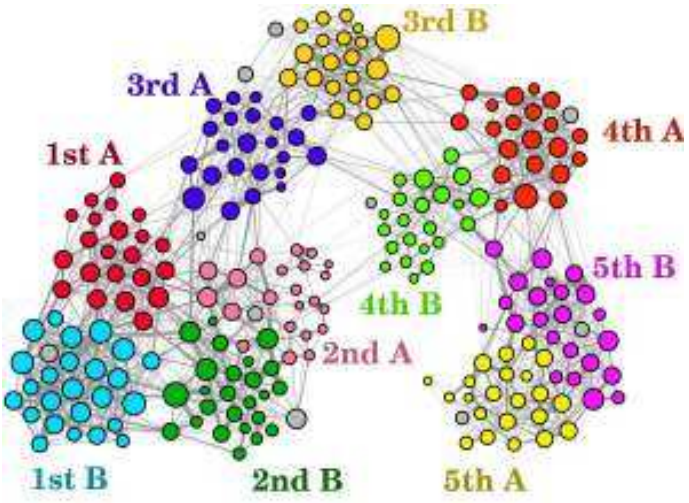


Figure 2.12: Network of contacts aggregated over the first day. Edges between individuals having interacted less than 2 minutes have been removed, thus keeping only the strongest links. The width of links corresponds to the cumulative duration of contacts, and nodes with higher number of edges have larger size. Colors correspond to classes, teachers are shown in grey. (Figure created using the Gephi software, http://www.gephi.org/)

neighborhood restricted to children in a different class). This indicates a repetitive pattern inside each class but a non negligible renewal of the contacts between classes across consecutive days.

2.3.2.4 Contact matrices

The school environment gives a natural way to categorize individuals, with the structure of classes and grades. Aggregated measures are defined to characterize contacts inside and among classes.

The **total number** of contacts between children of classes A and B is defined as the sum of the number of times n_{ij} that a contact event between i and j is recorded, for all i in class A and j in class B:

$$\begin{aligned} n_{AB} &= \sum_{i\in A, j\in B} n_{ij} \quad \text{for } A \neq B \\ n_{AA} &= \sum_{i,j\in A} n_{ij} = 2 \sum_{(i-j)|i,j\in A} n_{ij}. \end{aligned} \tag{2.11}$$

In the case of $A = B$, as each child of the class is counted twice, one as i and the a second time as j, n_{AA} is twice the total number of contacts between children of class A. Another convention could have been taken for n_{AA} but the advantage here, is that the average number of contacts of a child of class A with children of class B is given by the ratio n_{AB}/N_A , where N_A is the number of children in class A, even when $A = B$.

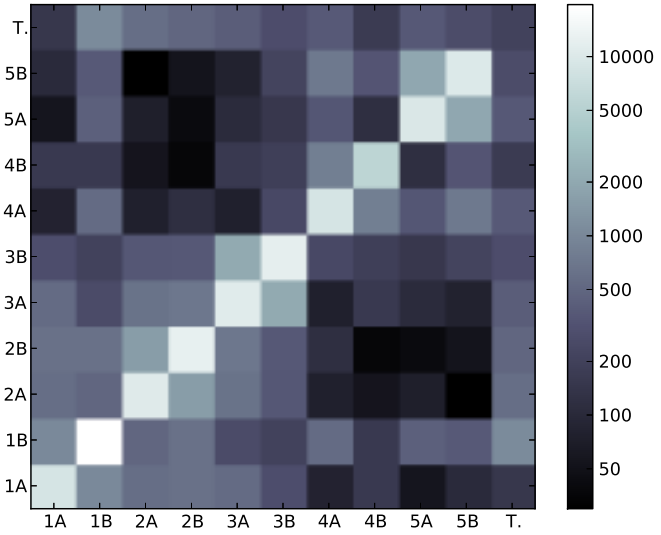
Similarly, the **total time spent in contact** between children of classes A and B is defined as the sum of the cumulated duration of contacts w_{ij} between all children i in A and j in B:

$$\begin{aligned} w_{AB} &= \sum_{i\in A, j\in B} w_{ij} \quad \text{for } A \neq B \\ w_{AA} &= \sum_{i,j\in A} w_{ij} = 2 \sum_{(i-j)|i,j\in A} w_{ij}, \end{aligned} \tag{2.12}$$

and the average contact time of a child of class A with children of class B is given by the ratio w_{AB}/N_A.

Figure 2.13 displays grayscale-coded matrices giving, at the intersection of row A and column B, respectively the total number of contacts (n_{AB}) and the total duration of contacts (w_{AB}) occurring between individuals of classes A and B during the two-day study. A clear hierarchical structure can be observed. Most contacts involve children of the same class, as shown by the whitish diagonal. Two-by-two light blocks around the diagonal also show that larger numbers and durations of contacts are observed between children of the same grade rather than with other grades, consistently with reports in [Conlan 2011]. A separation between smaller grades (1st to 3rd) and upper grades (4th and 5th) grades is also apparent, mainly because of the lunch break schedule. Finally, teachers have sparse contacts with one another because they spend most of the time in class with children.

(a) Number of contacts n_{AB}



(b) Cumulated duration of contacts w_{AB}, in minutes

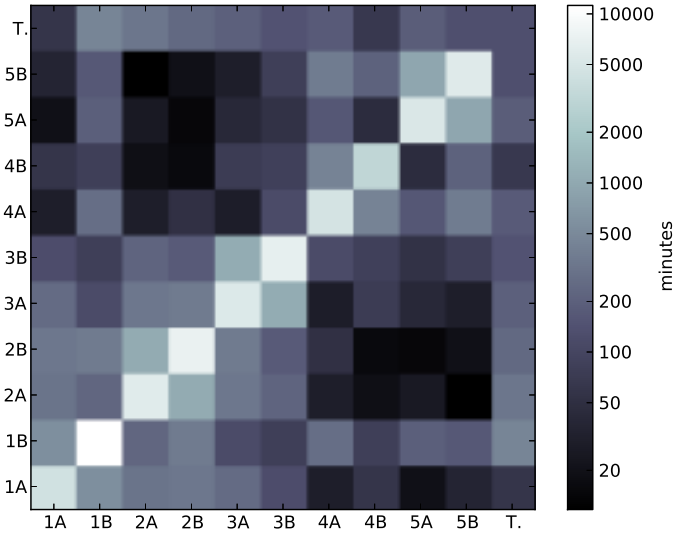


Figure 2.13: Grayscale-coded contact matrix between classes. The matrix entry for row A and column B gives the number of contacts n_{AB} (top) and the cumulated duration of contacts w_{AB} (down) measured between individuals of classes A and B over the two days of data collection. A logarithmic grayscale is used to compress the dynamic range of the matrix entries and enhance the off-diagonal hierarchical structure.

2.3.2.5 Number and duration of contacts

Figure 2.14 reports the total number $n_{A\cdot} = \sum_B n_{AB}$ and cumulated duration $w_{A\cdot} = \sum_B w_{AB}$ of contacts involving children of a specific class A or teachers. Figure 2.15 displays boxplots of the distributions of the individual contact numbers and cumulated durations, for each class and for each day. Figure 2.14 shows that the total number and duration of contacts involving teachers are smaller compared to those involving children, but figure 2.15 indicates that this is mostly a consequence of the number of teachers being smaller than the number of children in a class. The latter figure indeed shows the boxplots of the sum for each individual i in class A of its daily contacts $n_{i\cdot} = \sum_j n_{ij}$ and of the cumulated durations $w_{i\cdot} = \sum_j w_{ij}$ of these contacts, a quantity already referred to as the strength $s(i)$. Both the number and the duration of contacts show a limited degree of heterogeneity across classes as well as across days. This is partly due to different class schedules (e.g., a class being absent during half a day because of sport activities) or to different school activities (e.g., group vs individual activities).

2.3.3 Discussion

To our knowledge, this was the first study presenting detailed measures of close (face-to-face) proximity interactions between children in a primary school (see however [Salathé 2010] for the case of a high school). These descriptive results on contact patterns are of interest for modeling the spread of various infectious diseases, and possibly for investigating the role of specific control measures, such as closure of classes or immunization strategies, as discussed in chapter 5.

A number of other studies that rely on different methodologies (mostly surveys and direct observations) describe or estimate social contact numbers and durations. Comparison with previous results is clearly important but is made difficult by differences in the definitions of interaction/contact as well as by differences in the measurement techniques. As the present study considers the unsupervised detection of face-to-face proximity, it does not rely on surveys nor on the memories of participants. It is thus expected that larger total number and durations of contacts will be obtained, in comparison with survey-based methods.

Table 2.3 reports the comparison of the number and duration of contacts between previous studies and the present one. As expected, when all contacts are taken into account, we obtain larger values than the studies cited above, with the exception of [Salathé 2010]: as the infrastructure they described considers a broader detection range (3 meters proximity) than in the present case, it is not surprising that our study detects less numerous and shorter contacts. We report that each child has on average 323 contacts lasting 33 seconds per day with other children, corresponding to contacts with an average of 47 distinct other children, for an average daily total interaction time of 176 minutes.

To allow a more informed comparison between studies based on different methodologies, we compute for each child or for each pair of individuals the number

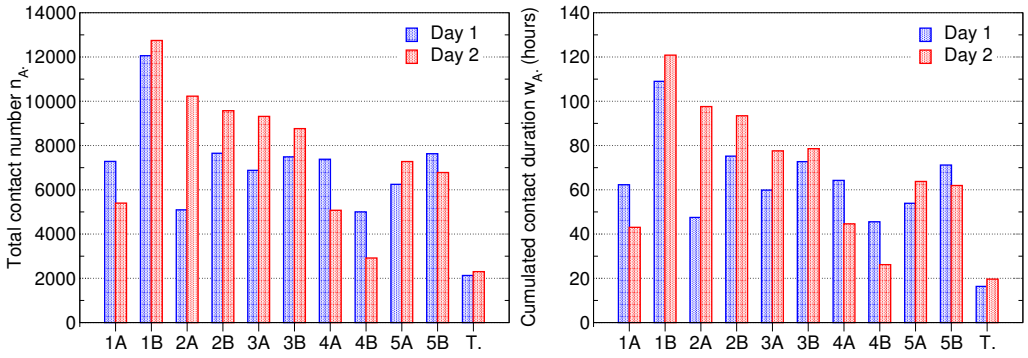


Figure 2.14: Total number of contacts n_A. (left) and total cumulated duration w_A. (right, in hours) involving individuals of each class and teachers (T) for each day.

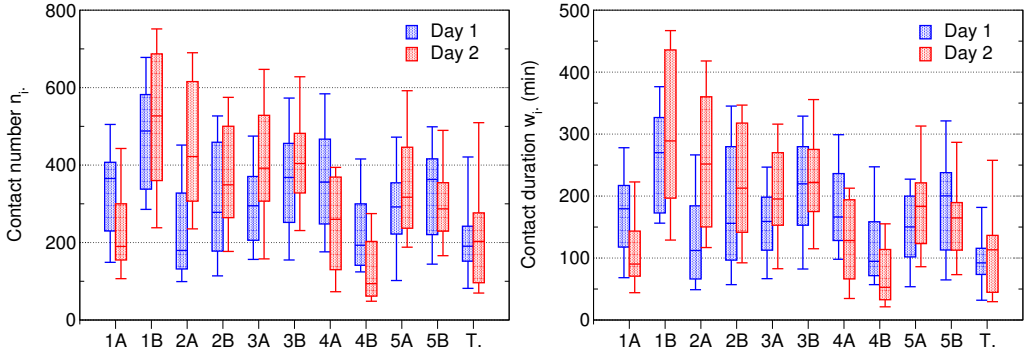


Figure 2.15: Boxplots of the distributions of the number of contacts n_i. (left) and of their cumulated duration w_i. (right, in minutes) involving an individual i, for each class and teachers (T) for each day. In each boxplot, the horizontal line gives the median, the box extremities are the 25th and 75th percentiles, and the whiskers correspond to the 5th and 95th percentiles.

	Setting	Contact definition	Results
[Mikolajczyk 2008]	Survey in a primary school; 6–10 year-old children.	A person with whom the child spoke or played with in a day	25.1 contacts per day per child
[Wallinga 2006]	General population survey, divided into 6 age classes.	Number of different conversation partners the participant encountered during a typical week by age classes	23.77 conversations per week (3.40 per day) held with different persons for 6–12 year-old children with other 6–12 year-old children
[Glass 2008]	High, middle and elementary schools survey, divided in 6 age classes (10 to 18 years old).	An interaction during which influenza could be passed. It must be within 3 feet and for a recognizable length of time	4.43 contacts per day for a 10–12 year-old child with other 10–12 year-old children. About 1 hour per day between 10–12 year-old children.
[Zagheni 2008]	School survey, divided in 3 age classes (5 to 19 year-old persons).	Estimation through time-use data, under the assumption of proportionate time mixing, of the co-presence of people in the same location.	98 min per day between 5–9 year-old children and 113 min per day between 10–14 year-old children.
[Del Valle 2007]	General population, divided into 9 age classes. Data are obtained from the EpiSimS agent-based simulation of a city, based on US census statistics.	Co-presence in the same sub-location. The duration is defined as the total length that two people spent together in the same sub-location. The durations of multiple encounters between two persons are added up and the total aggregated length gives the final contact duration.	227 min between children at school (not detailed for age groups).
[Mossong 2008]	General population in 8 European countries, divided into 10 age-classes.	Either skin-to-skin contact such as a kiss or handshake (a physical contact), or a two-way conversation with three or more words in the physical presence of another person but no skin-to-skin contact (a nonphysical contact)	From 2.25 to 11.88 contacts per day between 5–9 year-old children ; from 3.58 to 14.56 contacts per day between 10–14 year-old children. These numbers depend on the country
[Salathé 2010]	US high school. Students, teachers and staff.	Electronic devices (motes). A contact is defined as a continuous sequence of close proximity records (≤ 3 meters) between two motes.	1900 contacts per student per day, lasting on average about 1 minute. Each individual has contact with an average of 300 distinct other individuals.
[Stehlé 2011b] (present study)	Primary school with 6–12 year-old children and teachers.	RFID devices that exchange radio packets only when the individuals wearing them face each other at close range (about 1 to 1.5 m).	323 contacts per child per day, lasting on average 33 seconds, with on average 47 other distinct individuals. Averaged cumulated contact time of each individual of 176 min per day.

Table 2.3: Comparison of the measured average numbers and durations of contacts across several studies.

and total duration of contacts lasting longer than a given threshold. The results are summarized in Table 2.4. For instance, when restricting to cumulated contact durations of at least one minute, the number of different children with whom a child has interacted drops to 21, and the corresponding total interaction time drop to 163 minutes. Moreover, numbers close to those reported by [Glass 2008, Zagheni 2008, Del Valle 2007] are obtained when one takes into account only pairs of children having interacted for at least 10–12 minutes per day. Overall, our results are therefore quantitatively different from other studies, as can be expected from the strong methodological differences, but become compatible with previous studies when applying filtering procedures which retain only the longest contacts.

Filtering procedure: only contacts of duration at least T	Average daily number of distinct other children in contact	Average daily cumulated duration of contacts with other children, in minutes
T = 0	47.4	176
T = 40 s	20.8	100
T = 1 mn	11.8	65
T = 2 mn	4.1	28
T = 3 mn	2.2	19
Filtering procedure: only cumulated contacts at least W	**Average daily number of distinct other children in contact**	**Average daily cumulated duration of contacts with other children, in minutes**
W = 0	47.4	176
W = 1 mn	21.4	163
W = 2 mn	15.2	153
W = 5 mn	8.1	129
W = 7 mn	6.1	117
W = 10 mn	4.3	102
W = 12 mn	3.5	93
W = 15 mn	2.7	81

Table 2.4: Average number of children with whom a child was in contact, computed over one day, and average total time spent daily in contact with other children. Two filtering procedures are considered. Top: only contacts with duration at least equal to T are considered ($T = 0$ or 20 s corresponds to taking all contacts into account, given the available 20 s time resolution). Bottom: only links between children who have spent an amount of time at least equal to W in face-to-face proximity ($W = 0$ or 20 s corresponds to taking all links into account).

The study shows that children mix preferentially with children within their age group, mainly because they have more opportunity to interact with mates of the

same class than with children of other classes and grades, whom they see only during breaks and lunch. Nonetheless, these contacts between different classes and age groups exist. This information may help to advise public decision-makers on interventions aimed at containing or mitigating the propagation of communicable diseases at the level of schools, in particular in case of an epidemic or a pandemic. School closure has been proposed as an effective physical intervention to reduce transmission of respiratory pathogens, especially influenza [Cauchemez 2009]. However, it is not well understood how the benefit of closing entire schools, in terms of reducing cases, morbidity and mortality, compares to the economic costs of such interventions. In addition, the effectiveness of school closure depends on the effectiveness of other measures such as vaccination or antiviral drugs. Our results could be of interest in this context, especially if combined with other sources of information on the contact patterns of children [Salathé 2010, Mikolajczyk 2008, Conlan 2011]. The fact that a child spends three times more time in contact with classmates than with children of other classes suggests for instance that closing selected classes instead of the whole school could be a viable alternative. Additional intermediate steps between class and school closures could be devised through the analysis of aggregated contact networks, such as the one depicted in Figure 2.12, and exposure matrices such as the ones of Figure 2.13: classes most strongly linked to the class of the first detected case could for instance be closed in order to reduce the risk of propagation to the remaining classes. It would be interesting to assess by means of numerical simulation whether the closure of a single class or of a group of classes could efficiently mitigate the propagation of a disease at the school level. Finally, preventive measures such as shifts of the class schedules could substantially reduce contacts between classes, which could be particularly relevant for preventing transmission events from asymptomatic cases.

2.4 Partial conclusion and perspectives

In this chapter, I have presented various statistics to describe face-to-face interactions in three very different environments: in scientific conferences, in a museum and in a primary school. The variety of situations is mirrored in the variety of observed structures.

On the one side, the existing network toolbox is used to characterize time-aggregated networks, mainly on the scale of one day. Typical network measures, such as the degree and strength distributions, the analysis in terms of groups, of distances and the number of connected components, quantify the interaction patterns of each event, and are generally well explained by the knowledge of the contexts in which these interactions take place.

On the other hand, measures on the dynamic of interactions, such as the distribution of the contact durations, of the cumulated contact durations, of the time elapsed between two contacts, are revealed to be robust across contexts. These properties appear to be rather general in human interactions because similar results have

been obtained in other settings [Salathé 2010, Hui 2005, Cattuto 2010, Isella 2011] and with other types of interactions [Eckmann 2004, Rybski 2009, Onnela 2007, Leskovec 2008], although the generalization of such results should be carried out with caution.

Last but not least, the raw statistics on contact patterns are crucial for informing mathematical models that aim at describing the spread of infectious diseases and its prevention. Epidemiological models of disease transmissions in structured populations depend heavily on the knowledge of the amount and duration of contacts between individuals of different age groups. Here, the methodology allows to accurately estimate contact patterns, as we did for example in the case of the school dataset with the construction of contact matrices, which are widely used in epidemiology [Anderson 1991].

This methodology presents some limitations. First of all, the deployed infrastructure only measured contacts in a close environment, in the school building or in the playground for the school, in the museum premises or in the conference building. Contacts outside with the rest of the community are not recorded, and are of great importance for spreading dynamics (virus, rumors, fads). Moreover, badges were not worn during sport activities in the school deployment, which often involve close proximity situations and physical contacts. It would be interesting to use the data collection infrastructure to combine these sensor-recorded contact information with household data and data on contact patterns during school closure [Jackson 2011, Eames 2011]. Coupling the dynamical contact patterns at school and at home would allow to improve our understanding of the role of children as a reservoir during the spread of infections in a larger community.

Another potential issue concerns the possibility that individuals changed their behavior because they were wearing badges and knew they were participating in a scientific measure. This is especially important for children who are known to behave differently when adults are nearby (see [Maccoby 1990] reports on [Greeno 1989]). According to observers familiar with the environment (teachers and staff), however, no significant change could be detected in the children's behavior, and the children seemed to rapidly forget about the badges. In addition, while detailed explanations were given to the parents about the study and the badges, details on the role of the RFID badges (e.g., their detection range) were not given to the children.

From a public health perspective, it has to be emphasized that the collected data provide information on the mutual proximity of badges (and therefore of the persons wearing the badges), but not on the occurrence of physical contacts. Our measurements may thus be used in the context of, e.g., respiratory-spread pathogens but not for infectious agents transmitted by skin contact. Note however that physical contact can only occur between persons who are already in spatial proximity. Therefore, it would be very interesting to study the fraction of close encounters that result in a physical contact.

Chapter 3

Testing socio-psychological theories

Contents

3.1 Motivations 46
3.1.1 Drivers of social networks 46
3.1.2 Socio-psychological theories on virtual networks 47
3.2 Gender homophily among children 48
3.2.1 Background about gender homophily among children 49
3.2.2 Results 50
3.2.3 Discussion 60
3.3 Physical interactions versus online social networks 61
3.3.1 The Live-Social Semantics platform 62
3.3.2 Results 63
3.3.3 Discussion 71
3.4 Partial conclusion and perspectives 71

3.1 Motivations

3.1.1 Drivers of social networks

In sociology and in psychology, the network structure is often defined *as the set of principles driving its unfolding and its reproduction* [Ferrand 1997]. Various mechanisms have been identified thus far. I shall briefly list the most important of them.

The most known of these principles is given by the preference of individuals to interact and build friendships with peers they consider to be alike, is a well known feature of human behavior and is referred to as **homophily** (see [McPherson 2001] for a review). The field of traits that may influence human relationships is very broad, ranging from physical attributes to tastes or political opinions. The question of which kinds of similarities matter the most is rather open, and indeed the answer seems to depend on age and on the nature of the considered social ties. The outcomes related to gender or socio-ethnic homophily may vary over the lifespan of an individual and, for example, sharing similar working methods may have a stronger impact among coworkers than among friends.

A second important principle is what we call **triadic closure**, i.e., the fact that a friend of a friend is likely to be a friend too. In [Schaefer 2010], this effect is shown to be very important among children, as well as **reciprocity**, the fact of recognizing each other personal value in a friendship.

The **structural balance** is a principle that exists when both positive (e.g. a friendship) and negative (e.g. an enmity) interactions are considered. Let us consider three persons, A, B and C. The situation in which A is friend with B and C, but B and C hate each other, and the other situation in which A, B and C hate each other are not very stable. For example, the tension in the first situation can be solved with B befriending C or with one of A's friendships turning to enmity. In the second situation, an alliance against a shared enemy is much more stable. An interesting test of the structural balance theory is given in [Brandes 2009], in the case of a geopolitical dataset (which is not a social network) where political actors such as countries, international organizations or ethnic groups and events are assigned a weight ranging between -10 for the most hostile interaction and $+8.3$ for the most cooperative. (accusations, threat, military aid,...).

All these principles shape social relationships and can be a factor influencing the formation and the stability of groups. Nevertheless, it is often difficult to assess quantitatively to which amount any of these principles contribute to social structures, for the reason that they often occur simultaneously and that many of the traits that are considered to be important are changeable [Manski 1993]. In the case of homophily, peer influence and peer selection are very difficult to assess separately [Steglich 2010]. For example, similar smoking habits may be important among teenagers for the formation of new friendships but an individual having only smokers as friends might be influenced and become a smoker as well. Unless successive snapshots of social relationships and individual traits are available (panel network procedures), it is impossible to disentangle both effects [Kossinets 2009, Shalizi 2010], with the exception of almost immutable traits such as gender.

3.1.2 Socio-psychological theories on virtual networks

Technological networks and especially online networks have generated a vast literature in social network studies. This can be partly attributed to the lower economic and time cost compared to traditional survey for obtaining data. The theories and facts described in the previous section are often considered on online networks. The main hypothesis behind these studies is that relations between individuals on web sites reflect social interactions individuals have in the real world. Unfortunately to the best of my knowledge, its truthfulness has not been much studied.

First analysis concern homophily and community structure. Different virtual networks have been considered so far. In [Adamic 2003], a social network of university members is constructed by considering hyperlinks among personal webpages. In [Lewis 2008], Facebook activity shows that homophily based on gender, race/ethnicity, and socioeconomic status is present in cultural preferences. In [Takhteyev 2012], the geographic based, nationality and language based homophily

is assessed in the networking activity of Twitter. Very recently, in [Traud 2012], homophily and community structure of the Facebook network of 100 American colleges and universities differ noticeably among these institutions. For example, the effect of dormitory residences in Caltech universities which was expected to be more important than in other universities is quantified.

The structural balance theory has been tested too in the case of an online massive multiplayer game dataset in [Szell 2010]. The reciprocity among individuals has been studied with mobile phone calls [Kovanen 2011], which do not constitute a virtual network nor a social network. Triangle closure (together with homophily and reciprocity) has been examined in aNobii, a social bookmarking system [Aiello 2010].

Other famous theories often have a counterpart in the virtual world. For example, in [Leskovec 2008], the authors examine distances in an instant messaging network (Microsoft Messenger) and they show its small-world nature at a planetary-scale. It provides a mirror example of Milgram's experiment in a virtual network [Milgram 1967]. In [Kumar 2010], a dynamic version of the Barabàsi-Albert model [Barabási 1999] is tested on two virtual networks (Flickr and Yahoo! 360) to analyze the effect of preferential attachment among users.

The following studies are positioned within the above framework of measuring these socio-psychological theories shaping human interaction patterns in space by means of technological proxies or novel sensing techniques.

3.2 Gender homophily among children

We investigate gender homophily in the spatial proximity of children (6 to 12 years old) in a French primary school, using time-resolved data on face-to-face proximity recorded by means of wearable sensors, as described in 1.2.3. The deployment was initially designed for epidemiological purposes (results with this direction are presented in section 2.3 and published in [Stehlé 2012]). Unfortunately, no additional information on the socio-economic category of parents, socio-ethnic origins, school performances and extrascolar activities/interests were provided. Nevertheless, it is not the first time that a study on contact patterns designed for epidemiological purposes gives results on gender homophily. Indeed, Conlan et al. led a study in 11 primary schools and conclude on the existence of gender segregation, considered as the absence of reciprocate nomination between both genders inside the same class [Conlan 2011].

Measuring children's interaction patterns by means of wearable sensors is an interesting way to analyze children's behavior, which are known to present some particularities not presented in section 1.2.4. For example, Greeno analyzed how much children change their behavior when adults are nearby ([Greeno 1989] as reported in [Maccoby 1990]) . This is an important problem with direct observations, which are the main survey technique for very young children. With this respect, RFID tags provide a reliable way to access child behavior, but the reaction to the study and these sensors, which is not obvious given field experience, has not been

studied.

This study has been done in collaboration with my advisor Alain Barrat, Ciro Cattuto of the ISI Foundation and two other ENSAE students, François Charbonnier and Tristan Picard. A manuscript was submitted to the journal *Social Networks*.

3.2.1 Background about gender homophily among children

Quantitative analyses on gender preferences date back to Moreno's seminal work on sociometry [Moreno 1953], in which he introduces network terminology to describe relations between children from kindergarten through eighth grade[1]. His study relies on direct observations and interviews with children, and makes inferences about the variables that affect friendships. He shows that although young children up to the second grade prefer same gender mates, some of them also name friends of the opposite gender. This gender mixing then almost disappears, very few children making any mixed friendship up to the sixth grade.

Several studies have since confirmed and extended these results. A recent review [Mehta 2009] shows in particular a consensus about the fact that gender homophily exists along the entire life span: it is already present in infants' behaviors, increases up to a peak between 8 and 11 years [Maccoby 2003], in agreement with Moreno's study, and decreases afterwards, mainly because of the development of so-called *romantic relationships*. It reaches a rather stable level among adults, although studies on this life period remain scarce.

More recently, some differences of interaction styles between boys and girls have been brought to light. The first and widely reported difference is that boys tend to have a broader social network than girls (more network neighbors), who instead tend to make deeper and stronger relationships [Vigil 2007, Lee 2007]. In particular, when children are asked to list their friends with no limitation on the number, boys name more friends than girls but most of the reciprocate nominations occur between girls. The evolution of interaction styles is moreover different for both genders. La Frenière et al. conclude from the direct observation of 193 children aged between 1 and 6 years that gender preferences increase earlier for girls than for boys, but later on they become stronger for boys than for girls [La Freniere 1984]. On the other hand, a study of Martin et al. present a more moderate result [Martin 2001]. In a direct observation of 61 children between 39 and 74 months of age, they do not observe a significant correlation between age and the proportion of same sex playmates for the entire sample, but the correlation reaches a significant level when they consider boys only (this does not happen for girls). Indeed boys and girls behave differently, and differences in the amplitude of same gender preference have been also observed. Hayden-Thomson et al. asked 186 children about their positive, neutral or negative attitude toward all their classmates. While in the fourth grade girls are more positive towards boys than boys towards girls, the situation is reversed

[1] The paternity of Social Network Analysis is generally attributed to Moreno who published a extensive book in 1934, but previous works in educational and developmental psychology show that the fundamental concepts already existed [Freeman 1996].

in the sixth grade [Hayden-Thomson 1987]. Shrum et al. reach similar conclusions from questionnaires identifying friendships [Shrum 1988].

During adolescence and the beginning of romantic relationships, girls show an earlier evolution in their attitude toward other sex mates than boys. Richards et al. report that girls declare to have frequent thoughts about the opposite sex one grade earlier than boys (but more moderately) [Richards 1998]. They also declare to spend nearly twice as much time with boys as boys do with girls. This asymmetry may be explained by the fact that girls often have an older boy as second best friend outside school, while boys rarely report having girls as second best friends [Poulin 2007].

Finally, stability of relationships, i.e. the maintenance of ties over time, is a facet of friendship that has recently drawn some interest (see [Poulin 2010] for a review). It is known to increase during the primary school, which may be interpreted by the fact that concepts such as reciprocity, loyalty and ability of solving difficulties become more and more important in friendships at that age. Relationships between same sex peers are also more stable than mixed relationships but the empirical literature is still too shallow to conclude about gender differences in terms of stability.

3.2.2 Results

We use for this study the dataset on the face-to-face spatial proximity of more than 200 school children described in section 2.3. The metadata on age, class and gender is available for 227 out of 232 participating students. No information is provided on whether children are subjected to pre-defined seating arrangements during class time, nor whether teachers control or influence the seating patterns within classes. In order to remove possible biases due to such factors, the spatial proximity of children is studied only when they have maximum freedom of associating with one another: the analysis is limited to contacts recorded in the playground and canteen of the school, which overall account for 32 027 contact events (41% of the recorded contacts).

A first clue of the presence of homophily is given by the contact share with respect to the gender of both pair members. Figure 3.1 shows that there is a smaller proportion of mixed contacts in the 5th grade than in the 1st grade. Further conclusions from this kind of figure are limited because the boy/girl share is not the same between these two grades and because the total number of contacts is not the same.

3.2.2.1 Structure of the aggregated contact network

From the 2 day long contact sequence, an aggregated network is built following the method described in section 2.2.2. This aggregated network is made up of 227 nodes and 7070 weighted edges.

The cumulative edge weight histogram, which behaves similarly to those described in section 2.2.2 can be broken down according to the gender of the connected

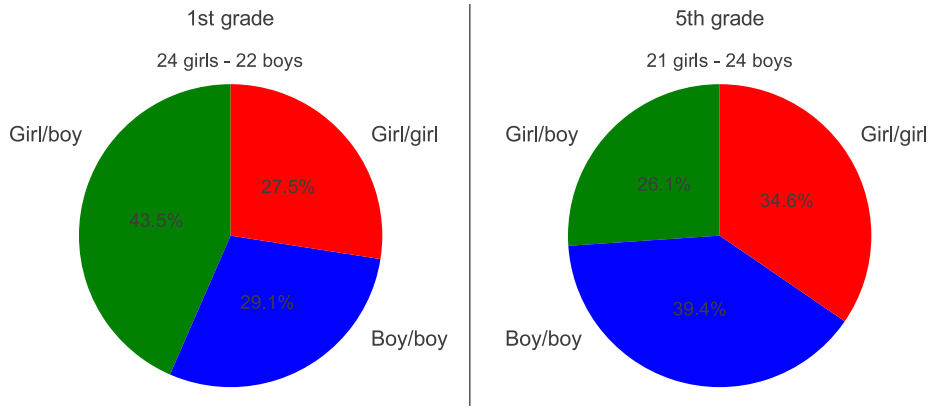


Figure 3.1: Pie chart of the contact number with respect to the gender of both pair members in the first grade (left) and in the 5th grade (right).

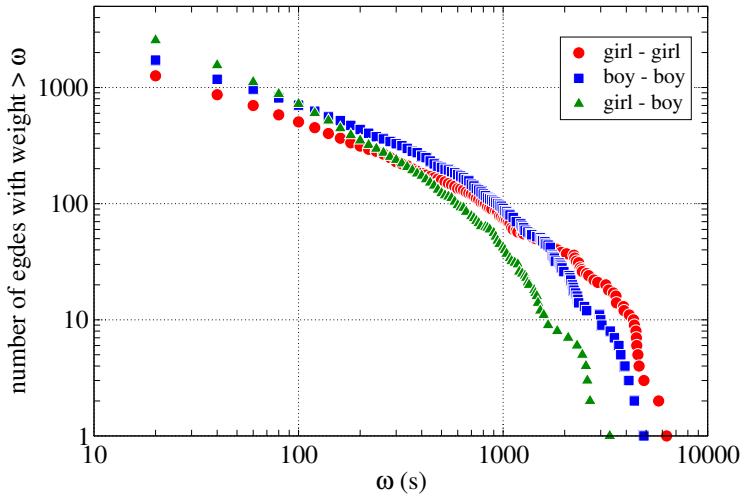


Figure 3.2: Cumulative histograms of edge weights of the contact network aggregated over two days. Edges are divided into three categories according to the gender of the connected nodes (boy–boy, boy–girl or girl–girl).

nodes and provides a first indication of gender-related differences, as shown in figure 3.2. Edges between boys are more frequent than contacts between girls (2223 vs 1573) whilst there are almost as many girls as boys ($N_g = 112$ girls vs $N_b = 115$ boys). There are more than twice as many mixed edges than edges between girls (3274 vs 1573), but far less than twice the number of edges between boys. The three cumulative histograms shown in figure 3.2 indicate that mixed-gender edges tend to correspond to shorter cumulated interactions relative to edges linking same-gender individuals: the mean weights of mixed-gender and same-gender edges are 118 s and 242 s, respectively. Moreover, edges between girls have on average higher weights than edges between boys, with mean weights of 265 s and 225 s, respectively.

In this behavioral aggregated network the average degree is equal to 62.3, and has a higher value for boys than for girls (67.1 and 57.3, respectively). This difference is significant at the 10% threshold (tested with a one-sided Wilcoxon test, $p = 0.06$). When considering the subgraph defined by edges with weight of at least 5 min, the average degree is 8.0, with a significantly higher value for boys than for girls (9.0 vs 7.0, $p = 0.07$).

These two observations are in agreement with the literature about differences on group size and level of intimacy, as reviewed by [Vigil 2007]. They support the hypothesis that men and women (here, boys and girls) arbitrate differently between maintaining a large social group and having more intimate and secure relationships.

3.2.2.2 Statistical evidence of gender homophily

In this section, the aim is to test statistically the evidence of gender homophily. A first approach consists in comparing the aggregated contact network with random graphs in which the probability that an edge connects two nodes is independent of node genders, to assess the probability that the observed contact network arises from a random arrangement of contacts between individuals.

We first restrict the study to contacts occurring within each class: the school schedule constrains contacts between classes, hence we cannot assume that children have the same opportunity to make strong ties within and across classes. Moreover, we only consider edges with weight of at least 5 minutes, i.e. whose contacts have a cumulated duration of at least 5 minutes over the two days of data. The aggregated contact network has 531 such edges, and in the following we will indicate them as *strong ties*, in reference to Granovetter's terminology [Granovetter 1973]. The threshold of 5 minutes is arbitrary: it has been chosen to be large enough to eliminate *weak ties* that will be shown later (Section 3.2.2.5) to have different properties with respect to gender homophily, and small enough to retain a sufficient number of edges for statistical analysis. We have checked that all of our results are robust with respect to changes in the 5 minutes threshold.

The number of strong ties involving at least one girl (denoted by $E_{g\star}$), and the number of strong ties involving at least one boy (denoted by $E_{b\star}$) are different. Moreover, boys have on average a larger degree than girls in the aggregated contact network. Since this difference could be explained by a unilateral preference for same-

sex peers, we test the evidence of same-gender preference separately for boys and for girls, fixing the number of strong ties between children of the other gender.

Let us first consider the possibility that strong ties including at least one girl are compatible with a model in which ties with boys and ties with another girl are equiprobable. This can be interpreted as an equal propensity of a girl to have a strong tie with all other boys and girls in her class. To this aim, we consider the null hypothesis $H0_g$ of a random graph in which nodes are labeled by gender (boy or girl), the total number of edges between boys is fixed, denoted by E_{bb}, the total number of edges involving at least one girl is also fixed, denoted by $E_{g\star}$, and edges between a girl and a boy or between two girls have the same probability to exist and are all independent. Therefore we want to compute the confidence interval of the number of same-gender edges E_{gg} under the assumption of validity of the null hypothesis $H0_g$.

In order to compute the probability mass function of E_{gg} under the null hypothesis, we note that the problem is analogous to a Bernoulli urn with A balls, where $A = N_g(N_g - 1)/2 + N_g N_b$ is the total number of *possible* edges involving at least one of the N_g girls, partitioned into $N_g(N_g - 1)/2$ white balls representing the possible ties between girls and $N_g N_b$ black balls representing the mixed-gender ties. We want to extract $n = E_{g\star}$ balls from this urn, with no replacement. The statistics of the number of white balls extracted (i.e., of the number of girl-girl ties E_{gg}) is thus given by a hypergeometric distribution of parameters n, $p = N_g(N_g - 1)/(2A)$ and A. We compute the region W in which the null hypothesis is accepted at the 5% threshold, $\Pr_{H0_g}(E_{gg} \in W) = 95\%$. We follow the same reasoning for testing the possible indifference of boys to establish strong ties with boys or girls (null hypothesis $H0_b$). Figure 3.3 shows for each class the region of acceptance of the null hypotheses at the 5% threshold, $H0_g$ for girls (in red), and $H0_b$ for boys (in blue). For girls, the empirical values are compatible with the null hypothesis $H0_g$ for 4 classes (2 classes of the 1st grade, 1 in 2nd grade, and 1 in fourth grade). For boys, the data are compatible with the null hypothesis $H0_b$ for 2 classes (1 in 1st grade and 1 in 4th grade). In a majority of cases (6 classes for girls and 8 for boys) we find evidence of gender homophily, as the empirical values of E_{gg} and E_{bb} are above the corresponding 95% confidence interval and the null hypothesis can be rejected[2].

It is important to remark some limitations of the approach described above. The null hypothesis $H0_g$ disregards all knowledge about the specific structure of the network (distribution of the number of neighbors, size of friendship clusters, etc.), except for the number of links involving at least one girl, $E_{g\star}$. In particular, edges are considered as independent variables, which means that a strong tie may exist between two children independently from the fact that they may share a lot of contacts. It is on the contrary known that many triangles exist in the aggregated contact network, just like in many social networks.

To overcome this limitation, we design a different null hypothesis: we consider

[2]The number of classes for which the null hypothesis is accepted depends slightly on the threshold used for the definition of strong ties, and is of at most 5 classes for boys and 5 classes for girls, for thresholds values ranging from 40 seconds to 10 minutes.

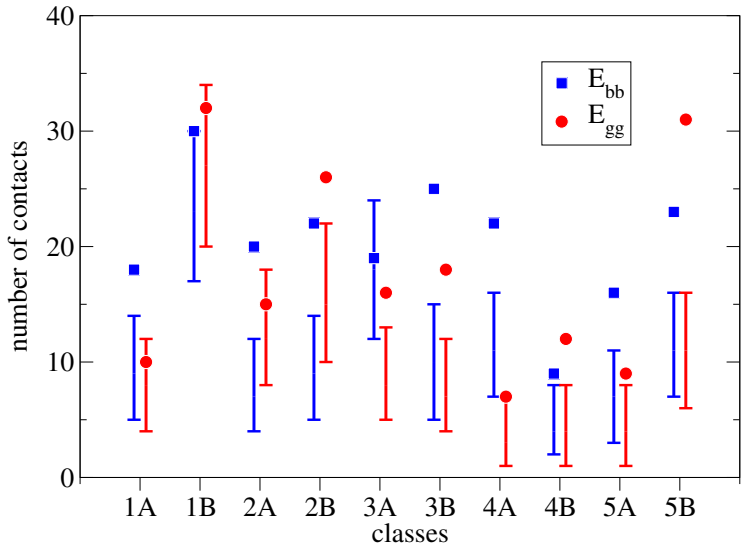


Figure 3.3: Statistical test of gender homophily for contacts within each class, restricted to contacts of cumulated duration of at least 5 minutes over two days. Error bars indicate the 95% confidence interval of acceptance of the null hypotheses of gender indifference. Symbols indicate the empirical numbers of girl-girl and boy-boy contacts.

the network of strong ties as fixed, and we randomly assign the gender of each node, preserving the total numbers of girls and boys in each class. Under such a null hypothesis genders are interchangeable and the network formed by the strong ties is independent from the gender attributes. As previously discussed, we need to control the number of ties between boys when testing the possibility that the part of the network to which girls take part is independent from gender allocation. To this aim, we fix the network structure and we consider all possible allocations of genders to nodes in which not only the number of girls and boys are equal to their empirical values, but also the number of links joining two boys is exactly equal to the empirical value E_{bb}. With such constraints, it is impossible to obtain an analytical formula for the distribution of the numbers of girl–girl and girl–boy ties, nor is it feasible in general to exhaustively list all the possible gender assignments that respect the constraints. Therefore, to estimate the 95% confidence interval for the number of girl–girl ties we resort to Monte-Carlo simulations[3]. This allows us to test at the 5% threshold the hypothesis that our data are compatible with the gender-shuffled null hypothesis. Overall, the results obtained with this method are exactly the same as with the previous null hypothesis: we have statistical evidence for same-gender peer

[3]For some classes, there are very few gender assignments that respect both the condition on the total numbers of boys and girls and that on the number of strong ties between boys. For 5 classes out of 10, less than 1000 possible allocations are found, and for 3 of them, less than 100 are found. The constraint on E_{bb} is hard to fulfill because its large value with respect to the total number of edges allows few gender assignments that are different from the real one.

preference in the same 6 classes for girls, and in the same 8 classes for boys.

3.2.2.3 Stronger same-gender homophily for boys than for girls

While the analysis of the previous paragraph was based on a global measure (the number of same-gender ties inside a group), heterogeneity between individuals can also be investigated through an individual index.

We consider here the network composed of strong ties (cumulated durations of contacts of at least 5 minutes, as above) linking children belonging to either the same class or different classes, for a total of 795 ties[4]. For each node i with at least one neighbor in this strong-tie network (215 children out of the previous 227), we compute the proportion of same gender peers among its k_i neighbors. We call this index the individual same-gender preference index, and we denote it by $P_k^{sg}(i)$. This index is equal to 1 if the considered child has only same-sex neighbors and equal to 0 if all the neighbors are of the opposite sex. A value of 0.5 indicates a perfectly gender-balanced neighborhood.

This index is close to the one introduced by [Criswell 1939], which is equal to the above ratio divided by its expected value in a random graph in which a tie exists between two nodes with a constant probability, independently from the existence of any other tie and from gender. [Criswell 1939] argues that a value close to 1 would indicate the absence of preference for same-gender peers, and values much higher than 1 would indicate the opposite. We have chosen not to consider this index because the null model Criswell is implicitly referring to might not be adequate here, as we know that the network structure is far from random: the opportunities for strong ties with classmates are not the same as with children from other classes, and the actual contact network has properties similar to those of social networks, such as triadic closure.

Figure 3.4 gives for each class the boxplot of the individual same-gender preference index, computed separately for boys and for girls. The dispersion of the index distribution is large, as indicated by the size of the box and the whiskers. While it happens that a child has no ties with other children of the same gender, most values of the index are rather high: same-gender homophily is present in all grades, for both genders. Moreover, the figure seems to indicate that boys tend to have a higher same-gender preference index than girls, and that this index increases with grades (we will examine this point in section 3.2.2.5).

The statistical difference between boys and girls can be estimated through a one-sided Wilcoxon test. This non-parametric test is preferred to a parametric one (such as ANOVA) because it does not require any assumption on the form of the distribution of the underlying random variable that generates the heterogeneity of same-gender preference index. We test the null hypothesis that the averages of the same-gender preference index are the same for boys and girls, against the one-sided

[4] In this case, children leaving the building for lunch will be less connected than others because they have a reduced opportunity to interact. Assuming that gender is independent from the behavior of eating at home, this does not bias our analysis.

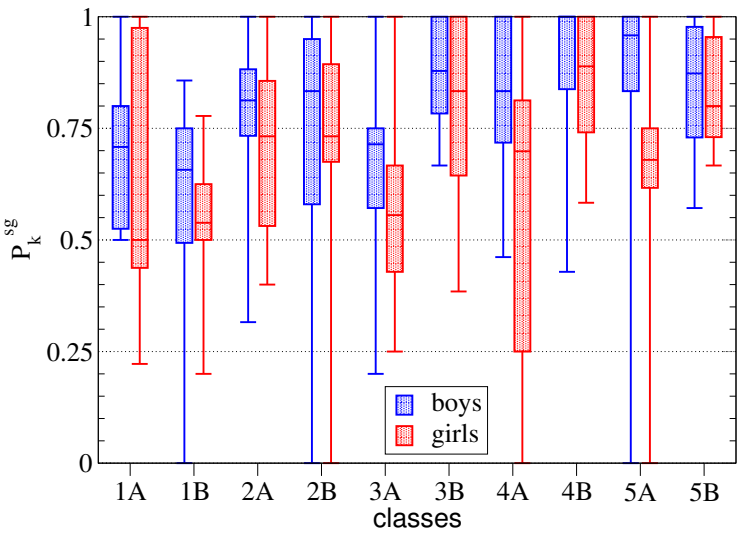


Figure 3.4: Boxplots for the different classes of the same gender preference index P_k^{sg}, computed separately for boys (blue) and girls (red). The center of the box indicates the median, its extremities the lower and upper quartiles, and whiskers indicate the smallest and largest values.

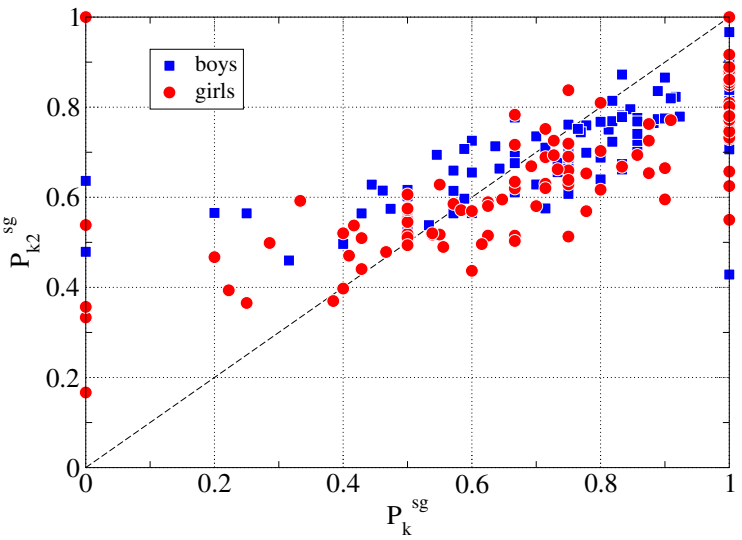


Figure 3.5: Scatter plot of the two-step homophily index P_{k2}^{sg} vs the same-gender preference index P_k^{sg}, computed separately for boys (blue) and girls (red). The first bisector is represented by a dashed line.

alternative hypothesis that the average is higher for boys than for girls. For the 4th and the 5th grades the null hypothesis is rejected at the 10% threshold, meaning that same-gender preference is statistically higher for boys than for girls only for the two highest grades.

Interestingly, and despite the distributions shown in figure 3.4 are mostly above 1/2, some children have most of their strong ties with children of the opposite sex. The interpretation proposed by Snijders[5] is that this situation could be socially allowed under the condition that the neighbors have, as well, many contacts with individuals of the other sex. To check this hypothesis, we define a two-step homophily index for each node i, slightly modifying the definition of the alters' covariate-average defined by [Ripley 2011]: $P_{k2}^{sg}(i) = \frac{1}{k_i}\sum_{j\in\mathcal{V}(i)} \frac{k_j(\text{gender}_i)}{k_j}$. This expression yields the average over the nodes j belonging to the neighborhood $\mathcal{V}(i)$ of node i, of the proportion of j's neighbors who have the same gender as node i. Figure 3.5 provides the scatter plot of this two-step homophily index with respect to the previous same-gender preference index $P_k^{sg}(i)$. Boys have on average a higher two-step homophily index than girls (p-value $< 5 \cdot 10^{-3}$ with a one-sided Wilcoxon test). Moreover, if we consider egos having a majority of neighbors of the opposite sex ($P_k^{sg} < 0.5$), the neighbors themselves have on average more ties with children of the same sex as ego than ego her/himself ($P_{k2}^{sg}(i) > P_k^{sg}$).

3.2.2.4 Stability of neighborhoods

As noted above, the gathered data correspond to a behavioral network of face-to-face proximity, and not to a self-reported social network. Since our dataset covers two successive days of school activity, the behavior of children from one day to the next can be compared, in particular to understand the interplay between gender homophily and the repetition of contact patterns.

To this aim, we quantify the similarity between the neighborhood of each individual i in day 1 and day 2 through the cosine similarity

$$\sigma(i) = \frac{\sum_j w_{ij,1} w_{ij,2}}{\sqrt{(\sum_j w_{ij,1}^2)(\sum_j w_{ij,2}^2)}}, \tag{3.1}$$

where the weight $w_{ij,1}$ is the cumulated time spent in face-to-face interaction between i and j during day 1, and $w_{ij,2}$ is the corresponding time during day 2.

We also consider two different similarity definitions that separately measure the similarities of the same-gender and of the opposite-gender neighborhoods across days ($\sigma_{sg}(i)$ and $\sigma_{og}(i)$, respectively). To this aim, we restrict the sums in Eq. 3.1 to neighbors j who have the same (or the opposite) gender as i:

$$\sigma_{sg}(i) = \frac{\sum_{j\in\mathcal{V}_{sg}(i)} w_{ij,1} w_{ij,2}}{\sqrt{(\sum_{j\in\mathcal{V}_{sg}(i)} w_{ij,1}^2)(\sum_{j\in\mathcal{V}_{sg}(i)} w_{ij,2}^2)}}, \tag{3.2}$$

[5] Tom Snijders addressed this issue during the presentation he gave at the Université Paris-Dauphine when receiving his honorary doctorate on December 16, 2011.

where the sums over j are restricted to the same-gender neighborhood $\mathcal{V}_{sg}(i)$ of i, and

$$\sigma_{og}(i) = \frac{\sum_{j\in\mathcal{V}_{og}(i)} w_{ij,1}w_{ij,2}}{\sqrt{(\sum_{j\in\mathcal{V}_{og}(i)} w_{ij,1}^2)(\sum_{j\in\mathcal{V}_{og}(i)} w_{ij,2}^2)}}, \tag{3.3}$$

where the sums over j are restricted to the opposite-sex neighborhood $\mathcal{V}_{og}(i)$ of i.

Figure 3.6 displays the boxplots of the distributions of σ_{sg} and σ_{og} for each class. We test the null hypothesis that the averages of σ_{sg} and σ_{og} are the same against the one-sided alternative hypothesis that the average is higher for σ_{sg} than for σ_{og}. Through a Wilcoxon test, the null hypothesis is rejected at the 5% threshold for 6 classes (one in the 2nd grade, one in the 5th grade, and all classes of the 3rd and 4th grades), and with a p-value of 0.125 for the other class of the 2nd grade. This shows that the same-gender part of the neighborhood of an individual is statistically more stable from day 1 to day 2 than the opposite-gender part of the neighborhood, in agreement with the literature reviewed by [Poulin 2010]. On the other hand, the stability of a child's neighborhood is not significantly dependent on her/his gender: a Wilcoxon test of the null hypothesis that the averages of σ are different for boys and girls leads to p-values larger than 0.3 except for one class of the 3d grade with $p = 0.06$.

3.2.2.5 Evolution of same-gender preferences with age

In the previous subsection, we noted in figure 3.4 a positive correlation between the same-preference index and grade. This would mean that as children become older during primary school they tend to interact more and more with same-gender mates, in agreement with previous studies [Mehta 2009, Maccoby 2003, Moreno 1953]. The information we have about the age of children allows us to investigate quantitatively the correlation between age and same-gender preference index, separately for boys and girls. Figure 3.7 shows the Pearson correlation coefficient between age and same-gender preference index, as a function of the threshold on the cumulated contact duration (5 minutes in the analysis above): when this threshold is equal to 20 seconds (the minimum duration of a contact) we retain all ties, while on increasing in less and less edges are kept (the strongest ones).

For both boys and girls, when we consider edges that correspond to a cumulated interaction time of at least 100 seconds (over 2 days), the correlation between the same-gender preference index and age is positive, and it is higher for boys than for girls. However, when weaker ties are retained (i.e., edges corresponding to shorter cumulated times), the correlation is instead negative for girls. This means that the evolution of homophilous behavior with age is different for weak and strong ties, and between boys and girls. A closer inspection reveals that for interactions of cumulated duration shorter than 3 minutes the number of same-gender mates decreases with age for both genders, and that the number of opposite-gender neighbors decreases even faster for boys, while it increases for girls. For contacts of cumulated duration larger than 5 minutes, on the other hand, the number of same-gender mates decreases

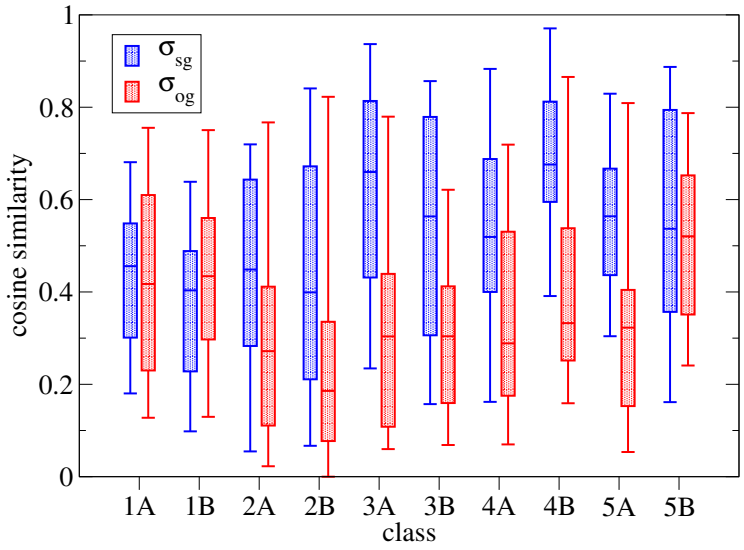


Figure 3.6: Distributions of cosine similarities between the neighborhoods of the individuals of each class in days 1 and 2, restricted to same gender neighborhood σ_{sg} and to opposite gender neighborhood σ_{og}.

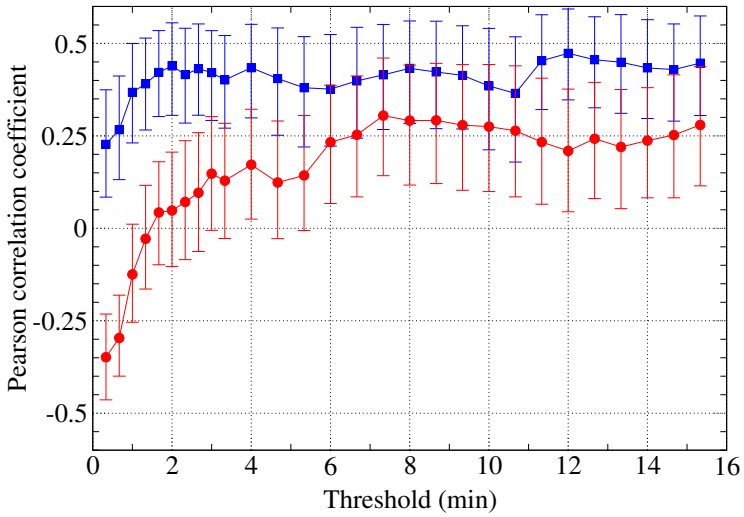


Figure 3.7: Pearson correlation coefficients between age and same-gender homophily index P_k^{sg}, as a function of the threshold on edge weights. The correlation is computed separately for boys (blue squares) and for girls (red circles). Error bars indicate the bootstrap 90% confidence interval, computed with 2000 resamples.

for boys and stays almost constant for girls, while the number of opposite-gender neighbors decreases for both genders. The increase in the number of short encounters that girls have with boys may be related to their earlier evolution in their attitude towards the other sex, as pointed out by [Richards 1998] and [Poulin 2007]. The same overall picture emerges when analyzing the correlation of the same-gender preference index with school grade rather than age, with a slightly weaker correlation of same-gender preference index and grade for boys.

It may be noted that for high enough values of the threshold some children become isolated in the network, meaning that they have no interactions with other children that account for a longer time than the threshold. This isolation does not affect equally children who are on time with the schooling schedule ($N = 204$) and children who are either in advance or late, i.e., who are younger (or older) than other children of the same class ($N = 15$ and $N = 8$, respectively). At the 5 minute threshold, 5 children out of these 23 become isolated, compared to 7 for those who have the same age as their classmates. The relative risk for these children to become socially isolated, compared to children who are on time is equal to 6.33 (a corresponding odds ratio of 7.8), indicating that children who are in advance or late might be more exposed to social exclusion.

3.2.3 Discussion

The use of wearable sensors represents a new tool in the study and description of child behavior. With respect to the previous literature, mainly based on direct observations and name-generator questionnaires, we recover here several known features with statistical significance:

1. gender homophily is present in all grades of the primary school,
2. same-gender preference reaches a higher level for boys than for girls in the 4th and 5th grades,
3. same-gender ties are more stable than mixed-gender ties across days, and
4. same-gender preference tends to increase with age for strong ties, at a higher rate for boys than for girls.

Thanks to the presented methodology, we are able to construct a complete network of face-to-face interactions of the school population over two days, which allows us to weight ties according to behavioral (who spends time with whom) rather than sociological proximity. In particular, we investigate how much boys and girls differ in their homophilous behavior when we consider their weak ties. It may be underlined that this method, even if relying on a behavioral proxy rather than on real information about social interactions (e.g., the different types of relationships are disregarded), may become an interesting tool for the study of the structure and evolution of weak ties, because it avoids informant biases such as the limited recall of individuals about their acquaintances.

3.3 Physical interactions versus online social networks

A second aspect related to socio-psychological theories that we have studied by mean of our wearable sensor design is called *multiplexity*. Social ties are not adequately described as an unidimensional relation between pairs of persons. People are linked by different types of relationships, such as kinship, friendship or work relationship. Even relationships of the same type can differ in what Granovetter called the strength [Granovetter 1973], i.e., a *(probably a linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie.* The study of the relation between these different ingredients of a social tie is not trivial. A way to tackle the problem is to consider a network with different types of ties, i.e., a *multiplex network*. A new kind of tie that was not envisaged by in Granovetter's paper because the web did not exist at the time of this seminal article, is given by ties on the Web. The existence of these ties has been widely promoted by the Web 2.0, i.e., the set of websites centered on interaction and communication between users instead of the previous generation of static webpages. The most famous example of a Web 2.0 site is Facebook. This social networking website counted on May 2012 more than 900 million active users over the world. Originally focused on relationships between university students, it has reached all age brackets and it is very common now that colleagues or relatives share a *Facebook friendship*, i.e., an individual access to personal information published by the *friend* on his/her wall.

The sensor-based infrastructure has been coupled several times to a a web platform designed for the analysis of users' activity on the Web 2.0, called the Live-Social Semantics platform. The major interest was to investigate the relationship between the amount of time spent together and the existence of virtual ties, in Facebook but also in different web 2.0 networking websites such as Flickr (mainly based on image and video hosting), Delicious (a social bookmarking service for storing, sharing, and discovering web bookmarks) and LastFM (a music website with a recommender system).

This association between the Live-Social Semantics platform and the RFID-based proximity analysis of face-to-face proximity has led to three deployments. The first was at the European Semantic Web Conference 2009 that took place in Heraklion from May, 31 to June, 6, the second time was at the ACM Hypertext 2009 in Torino (June 29-July 2) and it was deployed for a third time one year later at the Extended Semantic Web Conference 2010 in Heraklion (June 1-4).

We analyzed only the first two deployments (ESWC09 and HT09) whose characteristics are summarized in table 3.1. This work has been done in collaboration with Lorenzo Isella, Alain Barrat, Ciro Cattuto, Harith Alani, Gianluca Correndo, Marco Quaggiotto, Martin Szomszor and Wouter Van den Broeck. It is still in progress.

	ESWC09	HT09
Duration of the experiment	4 days	3 days
Number of conference attendants	305	∼150
Number of LSS participants	187 (139)	113 (97)

Table 3.1: Duration, number of participants to the conference and to the Live Social Semantic (LSS) experiment.

3.3.1 The Live-Social Semantics platform

The Live-Social Semantics platform, which will be referred as LSS hereafter, is dedicated to the study of people's activity on the web 2.0. More precisely, it is designed to register user's tagging activity and their social online social ties such as Facebook's friendships. The key assumption is that people mostly tag on topics, places, events or people they are interested in. The purpose of the LSS experiments is mainly to relate these tagging activities with those of people they are virtually linked with and to quantify the similarities and dissimilarities of activity. The general architecture and purposes of LSS are presented in details elsewhere in [Alani 2009, Van der Broeck 2010, Szomszor 2010, Barrat 2010].

During the coupled deployments between the LSS platform and the RFID-based infrastructure developed by the SocioPatterns collaboration, voluntary participants had the possibility to register on a dedicated website where they could:

- enter their RFID identification number
- indicate their account names on a collection of social networking websites such as Delicious, Flickr and LastFM,
- activate a Facebook application that collects their Facebook friendships.

Not all of the conference participants provided all of these informations. Table 3.2 gives the number of participants who registered on the website and the number of them who have provided an account information on the set of social networking websites.

Interestingly, after the conferences, some users did register on the LSS website but did not enter any social networking accounts. They were emailed a short questionnaire investigating their reasons. Table 3.3 lists the 36 received answers, which represents 43% of those participants. Out of the variety of reasons that were put forward, the reluctance to provide personal information on its web activity occurred rather seldom. The inadequacy or limited list of networking websites was a more frequent explanation for the people who subscribed to the LSS site but who did not entered any social networking website. Nevertheless, the conference participants who did not take part to the LSS experiment were not asked for the reason of their refusal. For those, the privacy issue may have been much more important.

	ESWC09	HT09	Total	
LSS participants	187	113	300	
Website registered users	139	97	236	
Facebook	78	48	126	(53%)
Delicious	59	28	87	(37%)
LastFM	57	26	83	(35%)
Flickr	52	23	75	(32%)
Haven't entered any social network	49	35	84	(36%)

Table 3.2: Number of participants to the LSS experiment and number of them who have provided an account name on various social networking websites.

Reason	Number of users	Percentage
Do not have those accounts (or rarely use them)	16	44%
Use different networking sites	10	27%
Do not like to share them	3	9%
Did not get a chance to share them (e.g., no computer, slow internet)*	6	17%
Other	1	3%
Total	36	100%

Table 3.3: Reasons why some users did not enter any social network accounts to the LSS application website. The person who picked *other* argued that he was too *busy* during the ESWC09 conference. *At ESWC09, these attendants often blamed the unreliable Internet connection at the venue for their inactive participation. This was not an issue for HT09.

3.3.2 Results

As Facebook is the most widely used networking site, we will restrict the analysis to this online network and refer to it as FB. A FB network is constructed from the data collected through the LSS application: nodes represent participants and an edge exists between two nodes if the corresponding persons are friends on Facebook. One can consider a multiplex network in which edges can be of two types: a FB friendship (non-weighted tie) or a conference network tie (weighted tie in which the weight corresponds to the cumulated duration of the contact between the corresponding persons). The aggregated conference contact network will be referred as the ACC network.

3.3.2.1 Number of shared edges

A first evidence of the correlation between the FB and the ACC networks is given by the examination of the number of shared edges, compared with their number of nodes and edges. These quantities are given in Table 3.4. In the ACC network,

	Conference network (ACC)	FB Network	ACC ∩ FB
ESWC09	N_{ACC} = 175 nodes	N_{FB} = 63 nodes	n = 63 nodes
	E_{ACC} = 4724 links	E_{FB} = 229 links	e = 152 links
HT09	N_{ACC} = 113 nodes	N_{FB} = 29 nodes	n = 29 nodes
	E_{ACC} = 2121 links	E_{FB} = 54 links	e = 49 links

Table 3.4: Number of nodes and edges in the aggregated conference contact network (ACC), in the Facebook network and number of shared nodes and edges.

the proportion of existing links is given by the ratio between the actual number of links divided by the number of possible ties, i.e., $2E_{ACC}/N_{ACC}(N_{ACC}-1)$. If the FB and the ACC networks were totally uncorrelated, the proportion of shared links would on average be given by the proportion of links existing in the ACC network, multiplied by the number of edges in the Facebook network:

$$\text{E[number of common links]} = \frac{2E_{ACC}E_{FB}}{N_{ACC}(N_{ACC}-1)} \tag{3.4}$$

This quantity is equal to 71 for ESWC09 conference and to 18 for HT09. They must be respectively compared with 152 and 49, which shows how much larger the effective number of shared links is. Would it be, however, possible by pure chance to reach such high numbers? The answer is very unlikely. Less than 1 out of 10000 purely random allocations of ACC and FB edges over n common nodes would give such numbers. This is a first clear indication that the two considered networks are strongly interrelated.

3.3.2.2 Weight distribution

A second evidence of this correlation is that friends on FB interact longer with each other than two people who are not FB friends. Figure 3.8 shows the cumulative weight distribution for the network aggregated on the whole ESWC09 conference and for links of this conference network between participants who are also friends on FB. The same distribution shape as in other contexts described in section 2.2.2 is observed. The difference between these two curves clearly highlights that the average time spent in face-to-face interaction in a conference by FB friends is several folds larger than if the two participants did not share this virtual friendship (respectively 904 seconds versus 159 seconds, significance tests: $p < 0.1$). Interestingly however, not all edges that have large weights correspond to virtual friends who have met during the conference, but the distribution of the latter is broader than the whole

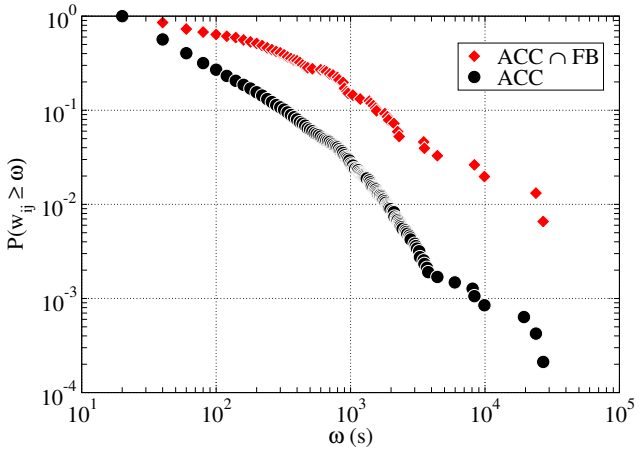


Figure 3.8: Cumulative distribution of weights for the face-to-face interaction network corresponding to the whole duration of the ESWC09 conference (black circles), and for the edges corresponding to pairs of participants who are also friends in the online social network Facebook (red diamonds).

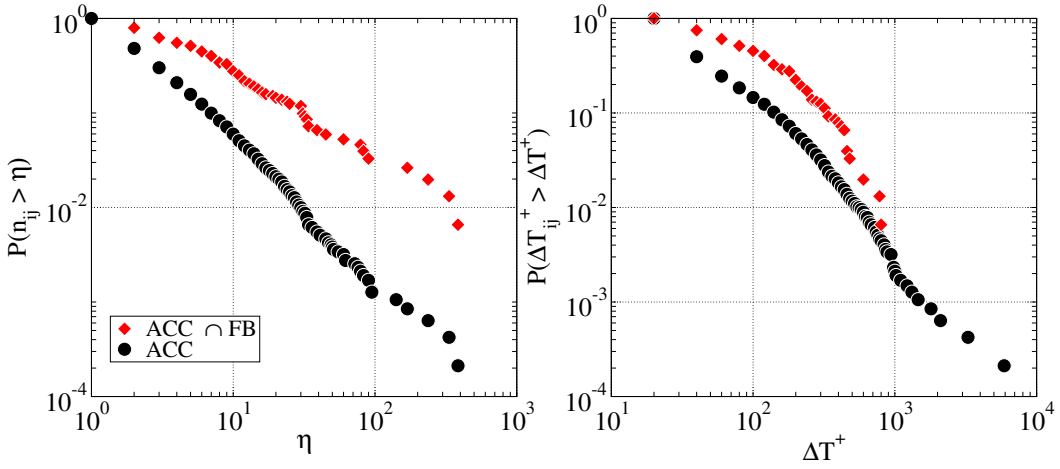


Figure 3.9: Cumulative distribution of the frequency of face-to-face interactions during the whole duration of the ESWC09 conference (left) and of their maximum duration (right), for pairs of conference participants who have met (black circles) and for pairs who are also friends in the online social network Facebook (red diamonds).

weight distribution. For any value of ω, the cumulative distribution $P(w_{ij} \geq \omega)$, giving the probability that a randomly chosen link has weight larger than ω, is larger, for any ω, when only links between FB friends are considered.

This relation between the time spent together and the fact of being friends on FB is not specific of the cumulated duration of interactions. It is also true that FB friends have met more often and that the longest interaction lasted longer than between people who do not share a FB friendship, as shown by Figure 3.9 that displays the cumulated distribution of the number of interactions n_{ij} and the maximum duration of these interaction ΔT_{ij}^{+}.

3.3.2.3 Behavioral similarity

To compare networks and uncover their interplay, it is interesting to go beyond the links and their weights, and to investigate and compare local structures. We focus here on the distribution of cosine similarities between nodes to investigate whether FB friends tend to have more similar behavioral patterns that in general. The cosine similarity of two nodes i and j ($\text{sim}_{i,j}$) previously described in section 2.1, yields indeed a simple and natural way to measure the similarity between the neighborhoods of these nodes.

In Figure 3.10, we compare the cumulative distributions of similarities $\text{sim}_{i,j}$ measured on the ACC network for the ESWC09 conference in the following cases:

1. all possible pairs (i, j);
2. pairs (i, j) of participants who are neighbors in the ACC network; and
3. pairs (i, j) of participants who share a tie in both the ACC and the FB networks.

The average similarity increases from case (1) to (3) (respectively 0.10, 0.13 and 0.20), and the proportion of pairs with zero similarity decreases (respectively 8.7%, 2.8% and 0%). Moreover, Figure 3.10 clearly shows that the value of the cumulative distribution $P(\text{sim}_{i,j} > \sigma)$ increases from case (1) to (3): for any similarity value σ, the proportion of conference participants pairs having similarity larger than σ in the ACC network increases if they are also friends on FB. The result is similar for the HT09 conference. This result is particularly striking: indeed, the weight of a link is related to the amount of interaction taking place between individuals, and it could be argued that it is somehow expected that individuals who already know each other will spend more time together. However, similarity measures go much beyond by showing that the social behaviors of individuals in a real-life social gathering are more similar if they share an online friendship.

For each tag i, we can also build at each instant t a *fingerprint* given by the vector $n_r(i, t)$ of the number of packets received by reader r in the time interval $[t, t - \Delta t]$, where we choose $\Delta t = 20$s. This fingerprint can be used to obtain a rough idea of the localization of each tag. We define the fingerprint similarity of

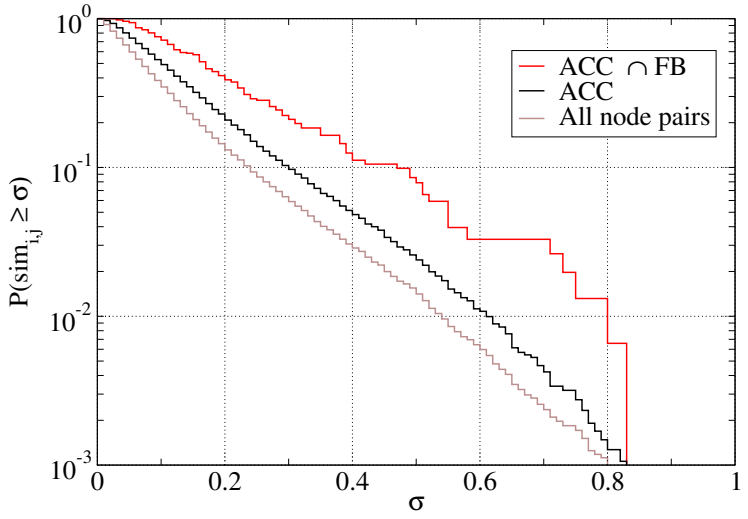


Figure 3.10: Cumulative distribution of neighborhood similarity for all node pairs (brown), for edges belonging to the ESWC09 ACC network (black), and for edges belonging to both the FB and the ACC network (red).

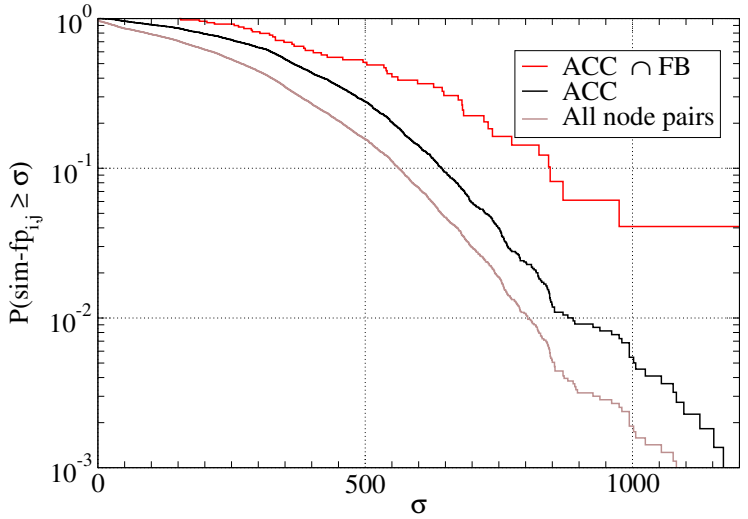


Figure 3.11: Cumulative distribution of fingerprint similarity for all node pairs (brown), for edges belonging to the ESWC09 ACC network (black), and for edges belonging to both the FB and the ACC network (red).

two tags the time-averaged cosine similarity of their fingerprints:

$$\text{sim-fp}_{i,j} = \int dt \sum_r \frac{n_r(i,t) n_r(j,t)}{\sqrt{\sum_{r'} n_{r'}(i,t)^2 \sum_{r'} n_{r'}(j,t)^2}} \tag{3.5}$$

This quantity gives therefore a proxy for the similarity of the physical trajectories of the tags i and j. Note that if $n_r(i,t)$ is exactly the same as $n_r(j,t)$ but with a time shift of more that Δt, the fingerprint similarity will in general be smaller than the presence duration at the event (if $n_r(i,t)$ is not constant). Other measures of similarities could have been envisaged, such as the Hamming distance, but the fingerprint similarity emphasizes the cost of temporal translations.

In Figure 3.11, we compare the cumulative distributions of fingerprint similarities $\text{sim-fp}_{i,j}$ measured on the ACC network for the ESWC09 conference for the three groups defined above. As for the previous similarity, the average fingerprint similarity increases from case (1) to (3) (respectively 285, 379 and 512). Again, Figure 3.10 shows that the value of the cumulative distribution $P(\text{sim}_{i,j} > \sigma)$ increases from case (1) to (3): for any fingerprint similarity value σ, the proportion of conference participants pairs having fingerprint similarity larger than σ in the ACC network increases if they are also friends on FB. This result means that people sharing a FB friendship tend be more often at the same place at the same time than people who do not share this kind of virtual friendship.

3.3.2.4 Link prediction

We now turn to the issue of link prediction, formulated as follows: is it possible to predict virtual friendships given any information on the aggregated interaction network?

A natural piece of information is given by the links weights: for any value w_{th}, one can construct a predicted online social network $\widehat{\text{FB}}(w_{\text{th}})$ by retaining all the links in the ACC network with weight larger than w_{th}, given the individuals corresponding to the connected nodes have provided their information on FB. Each pair of nodes in the ACC network can then be linked or not in the $\widehat{\text{FB}}(w_{\text{th}})$ and in the FB network, making inference right or wrong. There are two ways to be right[6]:

- links both belong to $\widehat{\text{FB}}(w_{\text{th}})$ and to FB (called True Positives TP) or
- links are both absent from $\widehat{\text{FB}}(w_{\text{th}})$ and from FB (True Negatives TN).

There are two ways to be wrong too:

- links belong to $\widehat{\text{FB}}(w_{\text{th}})$ but not to FB (False Positives FP) or
- links are absent from $\widehat{\text{FB}}(w_{\text{th}})$ but exists in FB (False Negatives FN).

This is summarized in table 3.5.

[6]Note that this terminology comes from information retrieval theory. An equivalent exists in statistics, where FN and FP are respectively called type-I and type-II errors if TP is considered as the null hypothesis.

	Link exists on FB	Link does not exist on FB
ACC link with weight $\geq w_{\text{th}}$	True positive (TP)	False positive (FP)
ACC link with weight $< w_{\text{th}}$	False negative (FN)	True negative (TN)

Table 3.5: Terminology of the link prediction.

By changing the threshold value w_{th}, we can construct precision-recall or receiver operating characteristic (ROC) curves, showing the *recall*, given by the true positive rate (ratio of true positives to the sum of true positives and false negatives)

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.6}$$

as a function of the false positive rate (FPR, ratio of false positive to the sum of false positives and true negatives)

$$\text{FPR} = \frac{FP}{FP + TN}. \tag{3.7}$$

In a perfect test, one wants to have all cases being either TN or TP. Thus, the larger the Recall, the better, meaning that one correctly identifies edges that are in FB. Similarly, the smaller the FPR, the better, because one correctly identifies edges that are not in FB. These two quantities should ideally tend to their limit values (1 for the Recall and 0 for the FRP), because a test refusing everything would give the ideal null value for the FPR, but the Recall would be null as well, and inversely, a test accepting everything would give a perfect Recall but a FPR equal to 1. The denominator of these two quantities is always constant and does not depend of the test. Random prediction would lead to an equal growth of these two rates when w_{th} decreases, and a diagonal straight line in the ROC plot. The quality of the prediction is measured by the area under the ROC curve (AUC). As shown in figure 3.12, large values are obtained: our classifier performs much better than random guess. Interestingly, figure 3.12 displays ROC curves using observed social interaction networks aggregated over different time windows. The best results are obtained when considering the whole conference duration, but even the observation of one single day gives a significant improvement over random prediction. This implies that the social interactions taking place during one conference day carry already a large amount of information on the existence or not of other social links between participants.

A third quantity, known as the *Precision*, is often used to quantify the accuracy of a test. The precision is defined as the ratio of true positives to the sum of true and false positives:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.8}$$

A precision quantifies the proportion of correct identification among the positives. The higher the precision, the better. Figure 3.13 shows the precision with respect

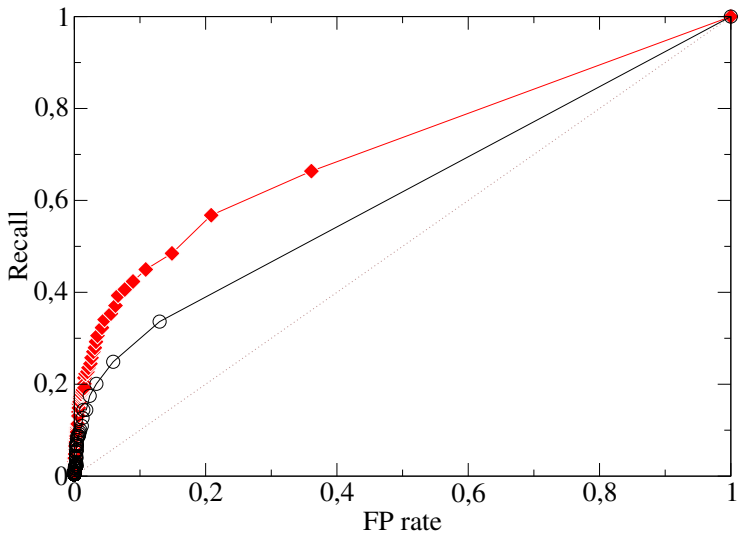


Figure 3.12: ROC curve for the prediction of FB links, given a ACC network either aggregated over one conference day (June 2, 2009) in open black circles or over the whole conference duration in red triangles. The Area Under Curve (AUC) is respectively of 0.61 and 0.71. A random prediction of pairs of nodes would follow the diagonal line, with an AUC of 0.5.

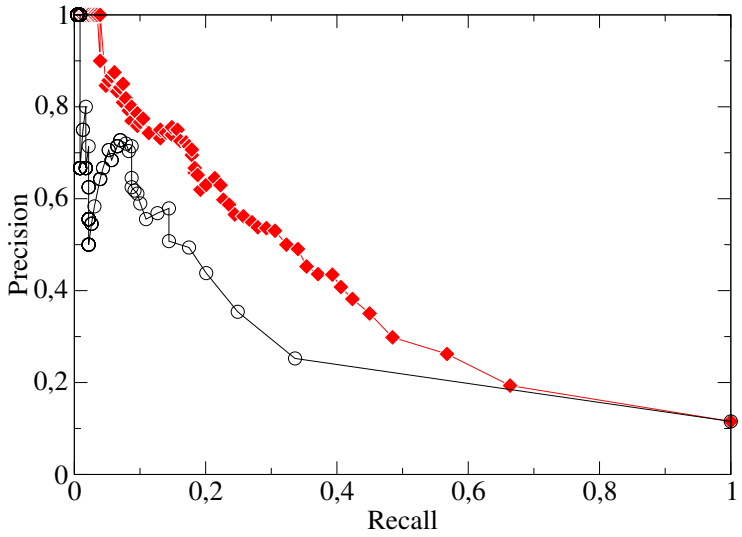


Figure 3.13: Precision-recall curve for the prediction of FB links, given as ACC network either aggregated over one conference day (June 2, 2009) in open black circles or over the whole conference duration in red triangles.

to the recall for the ESWC09 conference. This alternative visualization shows again that social interactions already give information about the presence or not of friendship ties on FB. The increased precision when taking the set of three days instead of one outlines the information carried by longer timescales.

3.3.3 Discussion

The sensor-based infrastructure developed by the SocioPatterns collaboration coupled with the Live Social Semantics platform has allowed us to get more insights into the relation between virtual ties on social networking websites provided by the web 2.0 and the amount of time spent together at conferences. This is a original contribution toward the study of social multiplexity. The non-supervised tools (i.e., the RFID sensors and the LSS website with its applications to collect virtual ties) provide a very modern way to investigate the relationships between conference attendees.

More precisely, we highlighted that

- face-to-face contacts between attendees who have provided information on their Facebook accounts, as recorded by RFID sensors, are more frequent and last longer if these attendees are friends on Facebook than if they were not,
- these Facebook friends also share a higher behavioral similarity (they interact more similarly with the rest of the population) and a higher trajectory similarity (they move from place to place in a similar manner),
- and face-to-face interactions provide relevant information on the existence of Facebook friendships, as shown by the very good results of the link prediction.

These results shed light on the interrelation between face-to-face proximity and virtual ties. Though the relation between geographic proximity at the city scale and the existence of virtual ties as already been described elsewhere (e.g. [Takhteyev 2012] for Twitter, a social networking and micro-blogging website that allows users to post and read 140 characters limited messages and [Liben-Nowell 2005] for LiveJournal, an other blogging website), to the best of our knowledge, this is the first analysis showing this interrelation at the individual scale.

3.4 Partial conclusion and perspectives

The empirical study of social interactions, and in particular the influence of peer behaviors on individual outcomes is of major interest in a broad range of social sciences, such as human behavior, sociology, economics or education economics, and organizational science [Manski 1993]. The collection of empirical social network data represents however a major barrier for the understanding of social influences, especially in the context of models such as those introduced by [Doreian 1980] or more recently by [Steglich 2010]. The development of unsupervised methodologies that

allow researchers to collect large-scale, high-resolution dynamical data on human behavior in a reproducible manner is a valuable asset. In order to develop this perspective, detailed comparisons between automated and supervised data collection methods, such as surveys or direct observations, will be highly desirable and crucially important to assess the specific limitations and potentials of the methodology.

With the analysis of homophily among children and the relation between virtual *friendships* and face-to-face proximity presented previously in this chapter, we have proposed a sociological angle to our datasets. Not only are we able to recover features already described in the literature, we also present some new results, thanks to the particularities of the protocol designed by the SocioPatterns collaboration, especially in what regards weak relationships.

Several perspectives can be envisioned at this step. First the study of the stability of encounters on longer timescales, with possibly separate waves and more individual characteristics is required to assess the robustness of our results which are obtained on a very short timescale of few days. Longer datasets are needed as well to investigate the temporal evolution of social ties and its effect on face-to-face proximities. The analysis in [Schaefer 2010] on the relative importance of reciprocity, popularity and triad closure is an a good example of what kind of analysis would be allowed by extended datasets. Second, social homophily is generally considered to be mainly the outcome of two mechanisms: peer influence and peer selection. In [Steglich 2010] is shown that their relative importance can differ considerably among social contagion processes (they study the case of drug and alcohol consumption). The type of face-to-face dynamic contact data with individual, possibly changing, characteristics is likely to provide interesting results with the methodology developed by Steglich et al. Thirdly, the question of behavioral or virtual *friendships* prediction can be further investigated with larger dataset. The question of human predictability has already been tackled in [González 2008] and [Song 2010] in the field of human trajectories, in [Takaguchi 2011] in the choice of conversation partners and in [Wang 2012] for the user behavior on two websites (a who-trust-who consumer review site and a location based social network). The kind of datasets on face-to-face proximities would be adequate to quantify how deterministic this behavior is, given the information available on individuals and past interaction patterns.

Chapter 4

Modeling the dynamics of encounters

Contents

4.1 Motivations 73
4.2 Model of interactions in a homogeneous population 75
4.2.1 Description of the model 75
4.2.2 Analytical solution with constant transition probabilities . . . 77
4.2.3 Analytical solution with a rich-get-richer effect 78
4.2.4 Numerical simulations 81
4.2.5 Aggregated networks 86
4.3 Variation on the model 87
4.3.1 Heterogeneous population 87
4.3.2 Fluxes in the population 91
4.4 Partial conclusions and perspectives 95

4.1 Motivations

In section 2.2.1 we have seen that the dynamics of face-to-face encounters exhibit long-time correlations and memory effects in agreement with what has been found elsewhere [Hui 2005]. This is not a peculiar feature of face-to-face encounters, similar properties have been observed in other contexts related to human interactions. For example, Eckmann et al. have shown using a dataset of emails exchanged in one of their universities that the distribution of interval times Δt between an email and its answer can be approximated by a power law with an exponent close to 1 on about 5 orders of magnitude [Eckmann 2004]. Rybski and al. with a statistical analysis of electronic messages in two Internet communities (the first is mainly composed of men having sex with men and the other one of teenagers) found that the number of messages users send to each other show long term correlations [Rybski 2009]. Onnela et al. analyzed a mobile phone call network of a mobile provider that contains a fifth of the population of an anonymous European country [Onnela 2007]. In this network, nodes correspond to phone numbers and an edge exists between two nodes if there has been at least one call between the corresponding phone numbers. A

weight, giving the total number of calls over the temporal span of the dataset, is associated to each edge, They noticed that the weight distribution is fat-tailed with an exponent close to 2 and showed that this weight distribution is responsible for a higher vulnerability to targeted removal of edges.

This repeated occurrence of broad distributions of activity in social systems has raised a considerable interest, especially in the physics community because it was reminiscent of the behavior of order parameters in phase transitions. Interestingly, many phase transitions of physical systems (such as alloys, superfluids and ferromagnets) can be classified into universality classes, characterized by the same set of critical exponents. This common behavior of very different physical systems was explained by renormalization group theory, which showed that simple mechanisms could drive the dynamics of the system and be more important than specific microscopic chemical or physical properties for the macroscopic behavior. By analogy, similar organizing principles have been looked for in social systems. For broad distributions, mainly two kinds of explanations have been invoked: the first being a *rich-get-richer* effect, introduced first by Simon in 1955 and the second one coming from optimization processes as proposed by Mandelbrot in 1953 [Simon 1955, Mandelbrot 1953][1].

In the context of temporal dynamics, Barabási proposed in 2005 a model based on queuing processes to explain why interevent time distribution for electronic communications such as emails or phone calls is often heavy tailed [Barabási 2005]. Previously in the literature of computer science that deals with managing traffics, queuing times were basically considered as Poisson processes which can not account for the presence of the non-markovian dynamics. In Barabási's model, individuals face a set of tasks to which they assign a perceived priority. Most tasks will be executed shortly while a few will wait very long to be done, so that the waiting time of the various tasks will be Pareto distributed. This simple mechanism can explain why in a situation in which an individual is presented with multiple tasks and chooses among them based on some perceived priority, the dynamics may not be Poissonian. From this model, it is possible to sort different types of queuing systems into *universality classes* depending on the exponent of the waiting time distribution [Vázquez 2006]. This model belongs to the category of *rich-get-richer* effect explanation, because a task with a very low priority will wait very long to be done and other tasks with an average medium priority will arrive in the list of things to do and be executed before the low priority one.

In collaboration with Ginestra Bianconi and Kun Zhao from the Northeastern University and my advisor Alain Barrat, I have worked on a similar kind of model to explain the dynamics of groups. Two articles have been published in Physi-

[1] In his article in 1953, Mandelbrot suggests an explanation about the long-tailed distribution of word occurrences in books. The basic idea is to consider a language as an information coding system: writing corresponds to the coding of information, while reading corresponds to the decoding step. The efficiency of a language, in this information encoding perspective, could rely on an implicitly optimizing strategy. Different criterion of optimality are tested and they would all lead to the same distribution family of word occurrences.

cal Review E on the subject [Stehlé 2010, Zhao 2011]. Previously, various types of models have been introduced to analyze the clustering and splitting of groups (see [Morgan 1976] for a classification of these models), but to the best of our knowledge, no work ever took the large temporal correlations into account. We introduce this aspect by means of a microscopic self-reinforcement mechanism that generates broad distributions such as those observed in the SocioPatterns deployments. Relatedly but with a perspective closer to the rational action theory [Hechter 1997], Johnson et al. proposed a model based on microscopic or individual-based mechanisms to explain the formation of groups on longer temporal scales. The meaning of group they consider is closer to the one used by sociologists, which includes a temporal persistence of the group or a repeated interaction, and they use their model to analyze empirical groups with data collected about guilds in World of Warcraft and urban street gangs in Long Beach, California [Johnson 2009]. Moreover, their model focuses on group size distributions and not on the dynamics.

4.2 Model of interactions in a homogeneous population

The model developed with G. Bianconi, K. Zhao and A. Barrat belongs to the framework of parsimonious modeling adopted within physics [Castellano 2009]. The principle is to develop simple, generic and easily implementable models that reproduce the empirical facts in order to distinguish explanatory mechanisms. In a second step, this kind of model can be used as benchmark to analyze how these bursty mechanisms could affect other dynamical properties. In the case of group dynamics, we are in the first place interested in mechanisms that generate heavy-tailed distributions for temporal quantities, but after having successfully confronted our model with empirical data, one might be interested in investigating how these properties affect the way information or infectious diseases diffuse in a population. It can be done either with empirical data (see the following chapter 5) for more realism, but for correctly assessing the effect of any network characteristic such as the weight distribution, a tunable model such as the one we propose may be more appropriate.

Most parts of this chapter are excerpts from published work [Stehlé 2010, Zhao 2011].

4.2.1 Description of the model

The model proposes a description of a dynamic network formed by disconnected groups of agents which evolve by splitting and merging. It aims at describing the dynamics of human social interactions in the context of small discussion groups, at short time scale and can reproduce features observed in empirical datasets such as the distributions of contact durations and of the time interval between two contacts. The mechanism is relatively simple, which allows for both analytical investigations and numerical simulations.

The model is composed of N agents that can interact with each other. It could represent people in a closed area, such as a conference hall, or a building, or social

animals in a relatively small environment. In this first assumption, we neglect spatial dispersion of agents and assume a well mixing dynamics. Each agent can either be isolated or interacting in a clique. During the evolution of the system, he can either stay alone, leave a group or be introduced to a group, depending on its state. We exclude the possibility of group merging or group splitting. More precisely, agents are defined by two variables: p_i the number of individuals an agent i is interacting with (also called coordination number) and t_i the last time the variable p_i evolved. At each time step t, an agent i is chosen at random. Two situations are possible, depending on the state of this agent.

- If i is isolated ($p_i = 0$), then with a probability $b_0\, f(t, t_i)$ it chooses a companion j to form a pair. This companion j is chosen among all other isolated agents, with a probability proportional to $\Pi(t, t_j)$. Then both p_i and p_j are set to 1 and t_i and t_j are updated to t.

- If i is interacting in a group of size $p+1$, a change can occur with probability $b_1\, f(t, t_i)$. With probability λ this agent leaves its group to become isolated ($p_i \to 0$ and $t_i \to t$ and for all other members of the group $p_j \to p-1$, $t_j \to t$), and with probability $1-\lambda$ it introduces to the group an isolated agent k chosen with probability proportional to $\Pi(t, t_k)$ among isolated agents (for i, k and all members of the group, $p_i \to p+1$ and $t_i \to t$).

The parameters b_0 and b_1 control respectively the probability to stay isolated or to keep the group constant, and λ controls the tendency to leave groups rather than make them increase in size. In the limit $\lambda = 0$, groups can only grow while on the contrary, if $\lambda = 1$, then only pairs are allowed as groups. The model dynamical behavior also depends on the functions f and Π.

In order to test the model against empirical data, the main quantities of interest are the times spent by agents in each state. To compute it analytically in the approximation of continuous time and number of individuals, we write the rate equation of the number of agents in each state. At time t, the evolution of the number $N_p(t, t')$ of agents that are in state p since t' is in the mean-field approximation:

$$\begin{cases} \partial_t N_0(t,t') = -\dfrac{N_0(t,t')}{N} b_0 \left[f(t,t') + \Pi(t,t')(r(t) + (1-\lambda)\alpha(t)\right] + \displaystyle\sum_{p\geq 1} \pi_{p,0}(t)\delta_{t,t'} \\ \partial_t N_1(t,t') = -2\dfrac{N_1(t,t')}{N} b_1 f(t,t') + \left[\pi_{0,1}(t) + \pi_{2,1}(t)\right]\delta_{t,t'} \\ \partial_t N_p(t,t') = -(p+1)\dfrac{N_p(t,t')}{N} b_1 f(t,t') + [\pi_{p-1,p}(t) \\ \qquad\qquad +\pi_{p+1,p}(t) + \pi_{0,p}(t)]\,\delta_{t,t'} \quad \text{for } p > 1 \end{cases} \tag{4.1}$$

where $\delta_{t,t'}$ is a Dirac delta function, equal to 0 everywhere except when $t = t'$, $\pi_{p,q}(t)$ is the average number of agents going from state p to state q at time t, and

$$r(t) = \frac{\sum_{t'} N_0(t,t') f(t,t')}{\sum_{t'} N_0(t,t')\Pi(t,t')} \tag{4.2}$$

$$\alpha(t) = \frac{\sum_{p\geq 1, t'} N_p(t,t') b_1 f(t,t')}{\sum_{t'} N_0(t,t') b_0 \Pi(t,t')} \tag{4.3}$$

where in the sums $t' < t$. In this equation, $\alpha(t)$ indicates the rate at which isolated agents are introduced by others in already existing groups of interacting agents.

If we consider functions f and Π that depend only on $t - t'$, then a stationary solution (if it exists) would be, by definition, invariant by time translation. The variables α, r and $\{\pi_{p,q}\}_{p,q}$ which depend only on t would remain constant and the time dependency of $\{N_p\}_{p\in\mathbb{N}}$ would be reduced to one time variable, for example $\tau = (t - t')/N$, which is the normalized duration since the last state change.

The average number of agents going from state p to state q at time t, namely $\pi_{p,q}$, can be written explicitly:

$$\begin{cases} \pi_{0,1} = 2b_0 \sum_\tau f(\tau) N_0(\tau) \\ \pi_{1,0} = 2b_1\lambda \sum_\tau f(\tau) N_1(\tau) \\ \pi_{p,0} = b_1\lambda \sum_\tau f(\tau) N_p(\tau) \quad \text{for } p > 1 \\ \pi_{p+1,p} = pb_1\lambda \sum_\tau f(\tau) N_{p+1}(\tau) \quad \text{for } p \geq 1 \\ \pi_{0,p} = b_1(1-\lambda) \sum_\tau f(\tau) N_{p-1}(\tau) \quad \text{for } p > 1 \\ \pi_{p,p+1} = pb_1(1-\lambda) \sum_\tau f(\tau) N_p(\tau) \quad \text{for } p \geq 1 \end{cases} \tag{4.4}$$

4.2.2 Analytical solution with constant transition probabilities

In the case where f is a constant and equal to Π ($f(\tau) = \Pi(\tau) = f$), it is easy to see that a stationary solution $\{N_p\}_{p\in\mathbb{N}}$ decays exponentially with τ. We have

$$\begin{cases} N_0(\tau) = \sum_{p\geq 1} \pi_{p,0} \exp(-(2 + (1-\lambda)\alpha) b_0 f \tau) \\ N_1(\tau) = (\pi_{0,1} + \pi_{2,1}) \exp(-2b_1 f \tau) \\ N_p(\tau) = (\pi_{p-1,p} + \pi_{p+1,p} + \pi_{0,p}) \exp(-(p+1) b_1 f \tau) \end{cases} \tag{4.5}$$

In the limit $N \to \infty$, we approximate sums by integrals to integrate $N_p(\tau)$ over τ for $p \geq 1$ and obtain relations between the number of agent in each state:

$$\begin{cases} N_0 = \dfrac{\lambda}{2 + (1-\lambda)\alpha} \dfrac{b_1}{b_0} \left(\sum_{p>1} N_p + 2N_1 \right) \\ N_1 = \dfrac{b_0}{b_1} N_0 + \dfrac{\lambda}{2} N_2 \\ N_p = \dfrac{1}{p+1} \left(p(1-\lambda) N_{p-1} + p\lambda N_{p+1} \right) \quad \text{for } p > 1 \end{cases} \tag{4.6}$$

where $N_p = \sum_\tau N_p(\tau)$ for $p \geq 0$.

The distribution $P_p(\tau)$ of the time spent in a given state p, follows an exponential decay, with parameter $(p+1)b_1 f$ if $p \geq 1$. In the case where $p = 0$, the parameter is slightly more complicated analytically because it involves the ratio α. It can be analytically solved with the previous equations. Using the equation set 4.6, the following relation is obtained:

$$-\lambda N_1 + \frac{b_0}{b_1} N_0 = 2\lambda \sum_{p=2}^{\infty} N_p. \tag{4.7}$$

This relation introduced in the definition of α given in 4.3 induces that $\alpha = \lambda^{-1}$. Consequently, the distributions $P_p(\tau)$ of the time spent in a given state p behaves exponentially as:

$$\begin{cases} P_0(\tau) \propto \exp\left(-(2 + \dfrac{1-\lambda}{\lambda}) b_0 f \tau\right) \\ P_p(\tau) \propto \exp\left(-(p+1) b_1 f \tau\right) \qquad \text{for } p \geq 1 \end{cases} \tag{4.8}$$

where $\propto$ means proportional to. One observe that the distribution P_0 decays faster when λ is close to 0, i.e., when the probability to leave a group is much larger than the probability to introduce an isolated person to a group.

4.2.3 Analytical solution with a rich-get-richer effect

The more interesting case where f and Π are decaying functions of τ corresponds to a situation with a *rich-get-richer effect* because the more an individual stays in a given state, the less likely its state will change. More precisely it implies that the longer an agent is interacting in a group, the smaller is the probability that he/she will leave the group; the longer an agent is isolated, the smaller is the probability that he/she will form a new group. For the sake of simplicity, we focus on the situation $f = \Pi$ so that $r = 1$ in equation (4.2), which permits some simplifications.

For some functions of f, calculations can be carried out completely. For example, a choice that is relevant for modeling interactions such as those collected through the *SocioPatterns* collaboration, is to consider $f(\tau) = \Pi(\tau) = (1+\tau)^{-1}$. In that situation, a stationary solution is characterized by $N_p(\tau)$ for $p \geq 1$ scaling as a

shifted power law with exponent $(p+1)b_1$[2]:

$$\begin{cases} N_0(\tau) = \sum_{p\geq 1} \pi_{p,0}(1+\tau)^{-b_0(2+(1-\lambda)\alpha)} \\ N_1(\tau) = (\pi_{0,1}+\pi_{2,1})(1+\tau)^{-2b_1} \\ N_p(\tau) = (\pi_{p-1,p}+\pi_{p+1,p}+\pi_{0,p})(1+\tau)^{-(p+1)b_1} \quad \text{for } p \geq 2 \end{cases} \tag{4.10}$$

More calculations are needed to obtain the analytical expression of other quantities, but with the expression of $\{\pi_{p,q}\}$, given in equation (4.4), as functions of $\{N_p\}_{p\in\mathbb{N}}$ and f, we can obtain the relations

$$\begin{cases} \pi_{1,0} = \lambda\pi_{0,1} + 2\lambda\pi_{2,0} \\ \pi_{2,0} = \dfrac{1-\lambda}{2}\pi_{1,0} + \lambda\pi_{3,0} \\ \pi_{p,0} = (1-\lambda)\pi_{p-1,0} + \lambda\pi_{p+1,0} \quad \text{for } p \geq 2 \end{cases} \tag{4.11}$$

These relations define a linear recurrence relation on the $\pi_{p,0}$ and solving this relation in $\pi_{p,0}$ allows one to find the analytical expression of α.

Consequently, under the conditions that $b_1 > 1/2$, $\lambda > 1/2$ and $b_0 > (2\lambda - 1)/(3\lambda - 1)$, the distribution functions $P_p(\tau)$ of the time spent in a given state p behave as

$$\begin{cases} P_0(\tau) = (1+\tau)^{-1-b_0\frac{3\lambda-1}{2\lambda-1}} \\ P_p(\tau) = (1+\tau)^{-1-(p+1)b_1} \quad \text{for } p \geq 1 \end{cases} \tag{4.12}$$

The conditions on the parameters b_0, b_1 and λ define a phase diagram of the model and outside these boundaries, the hypothesis of stationarity is violated. The figure 4.1 gives a three-dimensional view of the phase diagram.

In the stationary region ($\lambda > 1/2$, $b_0 > (2\lambda-1)/(3\lambda-1)$, $b_1 > 1/2$), the average state of an individual $\langle p \rangle$ is, by definition:

$$\langle p \rangle = \sum_{p=0}^{\infty} p \sum_{\tau} \frac{N_p(\tau)}{N} \tag{4.13}$$

Using the conservation of the number of agents, $N = \sum_{p=0}^{\infty}\sum_{\tau} N_p(\tau)$, and the expressions of the $N_p(\tau)$, one finds:

$$\langle p \rangle = \frac{\pi_{1,0}}{2\lambda} \sum_{p\geq 1} \frac{p(p+1)}{(p+1)b_1 - 1} \left(\frac{1-\lambda}{\lambda}\right)^{p-1} \tag{4.14}$$

[2] Of course, any decreasing functions f and Π can be used. For example we can show analytically that for $f(\tau) = \Pi(\tau) = (1+\tau)^{-\nu}$ with ν positive and different from 1, if the stationarity hypothesis holds, $\{P_p(\tau)\}_{p\in\mathbb{N}}$ become stretched ($\nu < 1$) or compressed exponentials ($\nu > 1$). In the case of $\lambda = 1$, the number of individuals in state 0 or 1 for τ is given by

$$\begin{cases} N_0(\tau) = \exp\left(-\dfrac{2b_0}{1-\nu}(1+\tau)^{1-\nu}\right) \\ N_1(\tau) = \exp\left(-\dfrac{2b_1}{1-\nu}(1+\tau)^{1-\nu}\right) \end{cases} \tag{4.9}$$

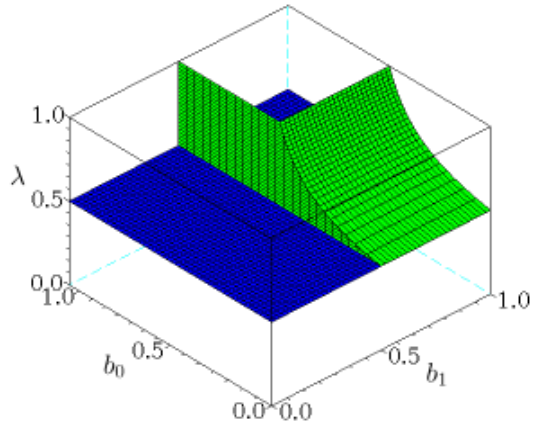


Figure 4.1: Phase diagram of the model with $f(\tau) = \Pi(\tau) = (1+\tau)^{-1}$. Boundaries are defined by $b_1 > 1/2$, $\lambda > 1/2$ and $b_0 > (2\lambda - 1)/(3\lambda - 1)$. Outside these boundaries, the stationarity hypothesis does not hold anymore.

where

$$\pi_{1,0} = \left[\frac{1}{2\left(b_0 - \frac{2\lambda-1}{3\lambda-1}\right)} + \frac{1}{2\lambda} \sum_{p\geq 2} \frac{p}{pb_1 - 1} \left(\frac{1-\lambda}{\lambda}\right)^{p-2} \right]^{-1} \tag{4.15}$$

For $\lambda \to 0.5^+$, the average state $\langle p \rangle$ diverges indicating that in this limit, the non-stationary state is governed by the presence of a large cluster of size $\mathcal{O}(N)$. This dynamical transition has not been observed among human beings, but it may be present in some animal behaviors, even though empirical measurements are limited [Morgan 1976, Gueron 1995, Bisson 2012].

Analytical results can also be obtained outside the stationary region. If $\alpha(t)$ given in equation (4.3) converges to a time-independent variable, that is $\lim_{t\to\infty} \alpha(t) = \hat{\alpha}$, a scaling assumption on the transition rates $\pi_{m,n}(t)$ allows to analyze situations where these quantities evolve with t. This is done in the following manner.

First we make the assumption that transition rates are either constant or decaying with time according to power-laws, that is

$$\pi_{m,n}(t) = \tilde{\pi}_{m,n} \left(\frac{t}{N}\right)^{-\beta_{m,n}} . \tag{4.16}$$

To check this assumption, we insert these expressions in equations (4.4) and in the

following equations that give the large time limit expression of $\{N_p\}_{p\geq 0}$

$$N_0(t,t') = \sum_{p\geq 1} \pi_{p,0}(t') \left(1 + \frac{t-t'}{N}\right)^{-b_0[2+(1-\lambda)\hat{\alpha}]} \tag{4.17}$$

$$N_1(t,t') = \left(\pi_{0,1}(t') + \pi_{2,1}(t')\right) \left(1 + \frac{t-t'}{N}\right)^{-2b_1} \tag{4.18}$$

$$N_p(t,t') = \left(\pi_{p-1,p}(t') + \pi_{p+1,p}(t') + \pi_{0,p}(t')\right) \left(1 + \frac{t-t'}{N}\right)^{-(p+1)b_1} \tag{4.19}$$

For large N, the approximation of sums by integrals allows to obtain that for $\lambda > 0.5$, the transition rate exponents are all equal to

$$\beta_{m,n} = \max\left(0, 1 - b_0\frac{3\lambda-1}{2\lambda-1}, 1-2b_1\right) \qquad \forall m,n. \tag{4.20}$$

For $\lambda \leq 0.5$, the self-consistent assumption breaks down and we will resort to numerical simulations.

The non-stationary region breaks down in two parts:

- for ($\lambda > 1/2$, $b_0 < (2\lambda-1)/(3\lambda-1)$ and/or $b_1 < 0.5$), the transition rate is decaying with time as a power-law but the distributions of lifetimes of groups $P_p(\tau)$ and of intercontact times $P_0(\tau)$ remain stationary. The coordination number in the limit $t/N \gg 1$ remains small, even as $\lambda \to 1/2$. In particular, the theoretical solution of the model predicts that for $\lambda > 1/2$ and $t \to \infty$,

$$\langle p \rangle = \begin{cases} 1 & \text{if } b_1 < b_0\dfrac{3\lambda-1}{2(2\lambda-1)} \\ 0 & \text{otherwise} \end{cases}$$

- for ($\lambda < 0.5$), the dynamics strongly depends on the number of agents N. The self-consistent assumption on the transition rates breaks down and we find numerically that the average coordination number $\langle p \rangle$ depends on the number of agents and on time.

4.2.4 Numerical simulations

We have performed numerical simulations to validate the analytical results for various f and Π and values of the parameters b_0, b_1 and λ, and with different system sizes N.

In the stationary region ($b_0 > (2\lambda-1)/(3\lambda-1)$, $b_1 > 1/2$ and $\lambda > 1/2$), up to $10N$ to $100N$ time steps are needed until transition rates remain constant, as shown in Figure 4.2 displaying the time evolution of $\pi_{1,0}$ with different parameter values.

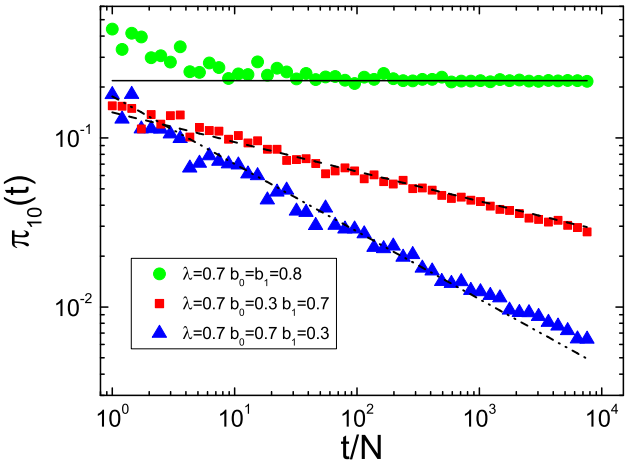


Figure 4.2: Evolution of the transition rate $\pi_{10}(t)$ in the different parameter regions for $\lambda = 0.7$. Numerical simulations are performed with $N = 1000$ agents for a number of time steps $T_{\max} = 10^4 N$ and averaged over 10 realizations. The parameter set for the green circles ($b_0 = b_1 = 0.8$) lies inside the stationary region, and the red squares ($b_0 = 0.3$, $b_1 = 0.7$) and blue triangles ($b_0 = 0.7$, $b_1 = 0.3$) are outside the stationary region. The lines indicate the analytical predictions.

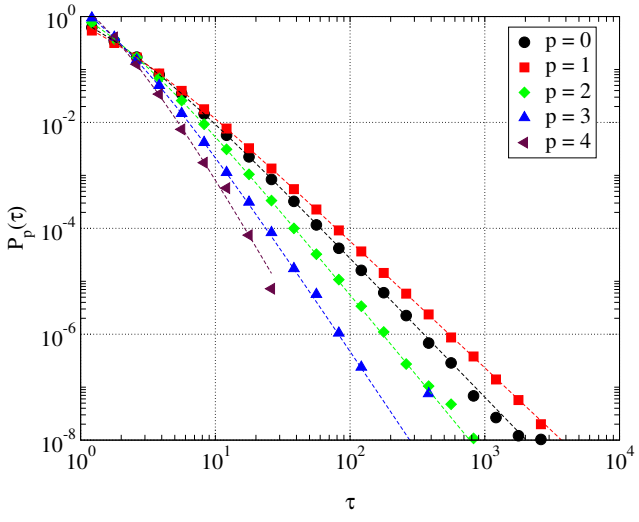


Figure 4.3: Distribution $P_p(\tau)$ of times spent in a given state p. Numerical simulations are done with $N = 1000$, $b_0 = b_1 = 0.7$, $\lambda = 0.8$ and the simulation is run for $T = 10^6 N$ time steps. The lines are the analytical predictions.

4.2.4.1 In the stationary phase region

In the situation where $f(\tau) = \Pi(\tau) = (1+\tau)^{-1}$, we compute the distribution $P_p(\tau)$ of time spent in a given state p. Figure 4.3 shows those distributions for $p \in [\![0, 4]\!]$, computed numerically and logbinned (see appendix in [Pastor-Satorras 2004] for more details on plotting heavy-tailed distributions). This figure suggests a good agreement with analytical results predicting shifted power laws.

Figure 4.4 shows the average state $\langle p \rangle$ for different parameter sets. We recover the results predicted in equation (4.14). When approaching the boundaries defined by the parameters b_0 and b_1, points obtained with longer computations ($T_{\max} = 10^4 N$) are slightly closer to the analytical line than those obtained with $T_{\max} = 10^3 N$. It can be due to a longer convergence time to the stationary state. For $\lambda \to 0.5^+$, the divergence of the average state is observed numerically.

In figure 4.5, we show in the left panel the distribution of contact durations between two agents, which differs from $P_1(\tau)$ when $\lambda \neq 1$ because a state $p = 1$ can be left either by splitting the pair or by introducing an other agent to the group, making the state variable change while the contact goes on. In the middle panel, we have represented the distribution of time elapsed between the beginnings of successive contacts of an agent A with possibly two different other agents B and C. This quantity is of great interest in the context of contagion processes because it is fully related to the time scale of the diffusion. In the right panel, the distribution of triads durations is plotted, which is again different from $P_2(\tau)$ because triads exist as well in larger cliques. All these quantities exhibit heavy tailed distributions, similarly to empirical observations as noticed in section 2.2.1.

4.2.4.2 Out of the stationary phase region

The system exhibits two different non-stationary behaviors depending on the value of λ.

For ($\lambda > 1/2$, $b_0 < (2\lambda - 1)/(3\lambda - 1)$ and/or $b_1 < 0.5$), transition rates decay with time as power-laws as shown in Figure 4.2.

Figure 4.6 displays the distributions of lifetimes of groups $P_p(\tau)$ and of intercontact times $P_0(\tau)$ given by numerical simulations. These remain stationary and are in agreement with theoretical results. The average coordination number remains small. Theoretical analysis indeed predicted it to reach a limit of 0 or 1 in the limit $t \to \infty$, depending on the maximum value between $1 - 2b_1$ and $1 - b_0 \frac{3\lambda-1}{2\lambda-1}$ (see equation (4.21)). Figure 4.7 shows the agreement of this predicted behavior with simulation results for several parameter values in this nonstationary region.

If $\lambda < 0.5$, i.e., if a group size is more likely to increase than to decrease, a large cluster appears with a size of $\mathcal{O}(N)$, lasting on a diverging time scale. Interestingly, it seems that the distributions of $\{P_p(\tau)\}_{p\in\mathbb{N}}$ remain stationary. This might be an interesting feature because most empirical data are obtained in non-stationary environments but exhibit stationary distribution.

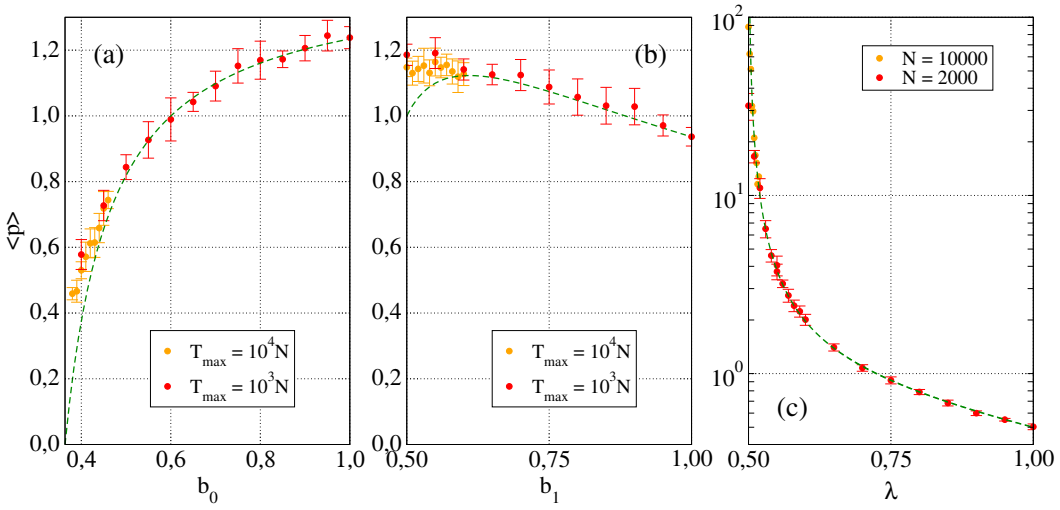


Figure 4.4: Average state $\langle p \rangle$ for different sets of parameters and population sizes. (a) $b_1 = \lambda = 0.7$, $N = 2000$ and two different computation durations $T_{\max}$, (b) $b_0 = \lambda = 0.7$, $N = 2000$ and two different computation durations $T_{\max}$, (c) $b_0 = b_1 = 0.7$, two different population sizes N and a computation duration $T_{\max} = 10^3 N$. The green dashed lines indicate the theoretical prediction given in Equation (4.14).

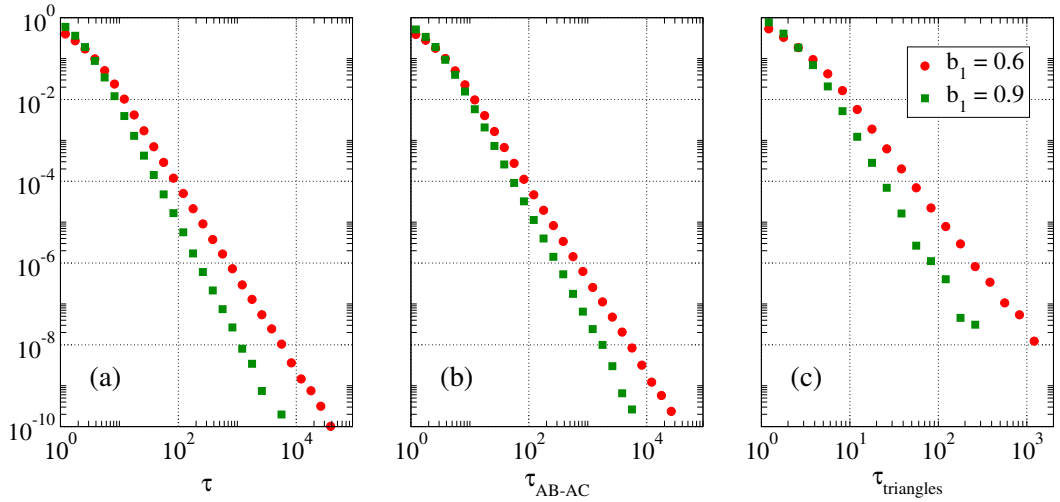


Figure 4.5: For different values of b_1, distribution of (a) the duration of a contact between two agents, (b) the time intervals between the beginning of two successive contacts of an agent A with two different or same agents B and C, (c) duration of a triad ($b_0 = 0.7$, $\lambda = 0.8$, $N = 1000$ and $T_{\max} = 10^5 N$).

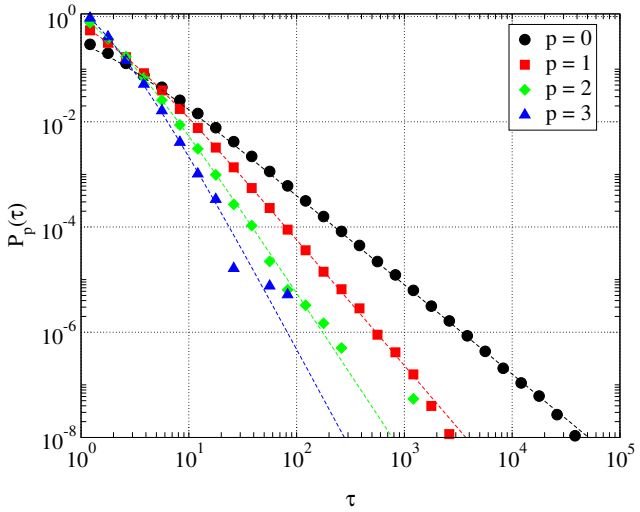


Figure 4.6: Distribution $P_p(\tau)$ of times spent in a given state p. Numerical simulations are done with $N = 1000$, $b_0 = 0.3$, $b_1 = 0.7$, $\lambda = 0.8$ and the simulation is run for $T = 10^5 N$ time steps. The lines are given by equations 4.12.

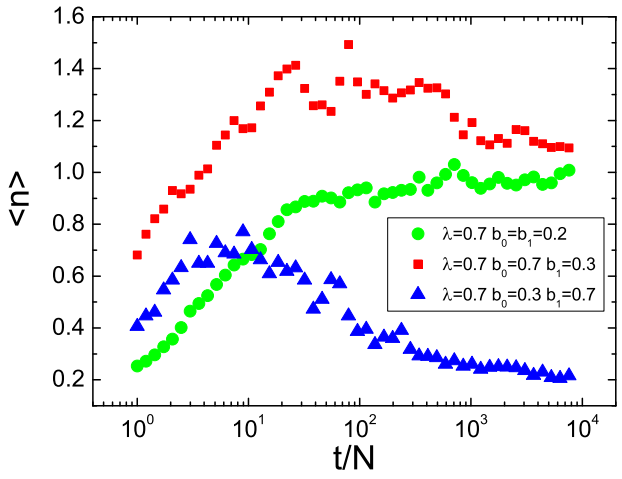


Figure 4.7: Average coordination number $\langle n \rangle$ as a function of time, when $\lambda > 1/2$, $b_0 < (2\lambda - 1)/(3\lambda - 1)$ and/or $b_1 < 0.5$. Simulations are performed with $N = 1000$ agents for a number of time steps $T_{\max} = 10^4 N$. Data are averaged over 10 realizations.

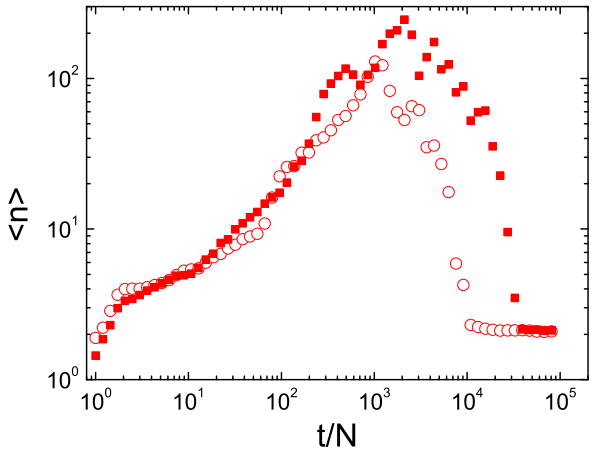


Figure 4.8: Average coordination number $\langle n \rangle$ as a function of time, for $\lambda = 0.2$, $b_0 = b_1 = 0.7$. Simulations are performed with $N = 250$ (open circles) and $N = 500$ (filled squares) agents for a number of time steps $T_{\max} = 10^5 N$.

4.2.5 Aggregated networks

In the previous paragraphs we have shown how our modeling framework produces dynamical properties of the interactions between agents that yield broad distributions of contact and intercontact times. In order to understand the structure of the resulting interaction networks at coarser temporal resolutions, it is as well interesting to investigate the properties of the aggregated networks constructed as in Chapter 2.

Given a starting time t_0 and a temporal window ΔT the nodes of these networks are the agents and a link is drawn between two agents whenever they have been in contact between t_0 and $t_0 + \Delta T$, with a link weight given by the total time during which they have interacted in $[t_0, t_0 + \Delta T]$. As in Chapter 2, the degree k_i of an agent i is given by the number of distinct agents with whom i has been in contact in $[t_0, t_0 + \Delta T]$, while its strength s_i is the sum of the interaction times with other agents, and the participation ratio $Y_2(i)$ quantifies the heterogeneity of the times spent by i with these other agents.

As an exhaustive exploration of the aggregated networks and of how their properties depend on the model's parameter would be tedious, we simply report in Figure 4.9 the properties of aggregated networks for increasing window lengths ΔT and for two sets of parameters. Some properties are qualitatively similar to the empirically observed networks. In particular, the degree distributions are peaked around an average value that increases with ΔT . As time passes each agent encounters more and more distinct other agents, and the distribution $P(k)$ globally shifts towards larger degrees. The links weights distributions are broad and extend

to larger values as ΔT increases. Some other properties seem to depend strongly on the model's parameters. In particular, the average strength of nodes of degree k, and the average participation ratio of nodes of degree k, can have shapes rather different from the empirical ones. Moreover, the time window lengths ΔT on which the aggregated network remains sparse are rather restricted.

4.3 Variation on the model

To illustrate the model's versatility, two examples of realistic extension of the model are provided here. In the first, agents can have heterogeneous propensities to form groups, in the second, the population is allowed to vary with time for example to account for entrance and exit fluxes.

4.3.1 Heterogeneous population

In the previous section, agents were considered to have the same tendency to form a group or to leave a group. Real social systems display however additional complexity since the social behavior of individuals may significantly vary across the population.

A natural extension of the model presented above consists therefore in making the probabilities to form or to leave groups dependent on the agent who is updating its state. For that purpose, we assign to each agent i a parameter η_i that characterizes its propensity to form social interactions. In real networks this propensity will depend on the features of the agents. In the model we assume that this propensity, that we call *sociability*, is a quenched random variable, which is assigned to each agent at the start of the dynamical evolution and remains constant, and we assume for simplicity that it is uniformly distributed in $[0, 1]$. In this modified model, the probability $p_p^i(t, t')$ that an agent i with coordination number p since time t' changes his/her coordination number at time t is given by

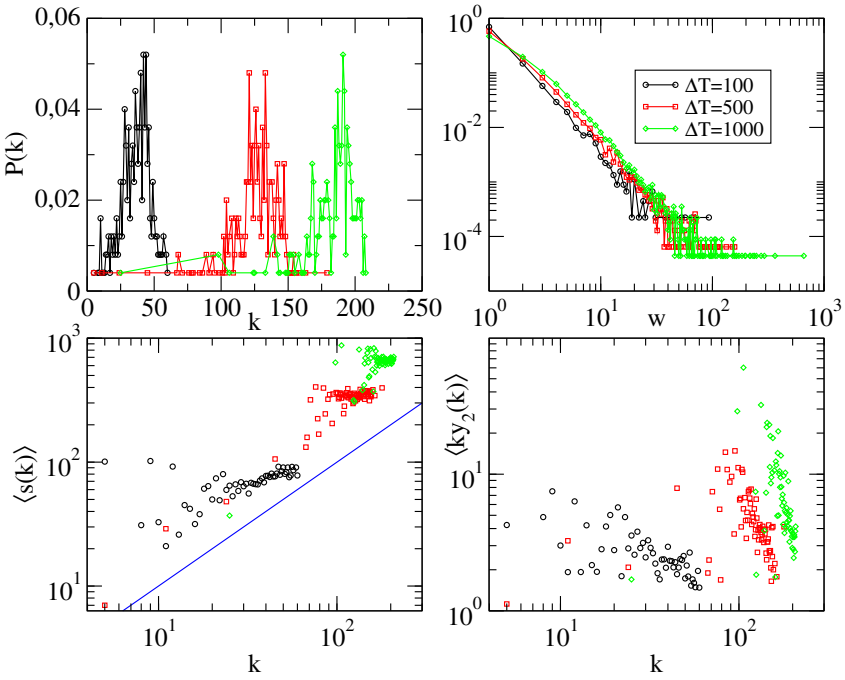
$$\begin{cases} p_0^i(t, t') = \dfrac{\eta_i}{1 + (t - t')/N} \\ p_p^i(t, t') = \dfrac{1 - \eta_i}{1 + (t - t')/N}, \text{ for } p \geq 1. \end{cases} \tag{4.21}$$

In this setup, the parameters (b_0, b_1), which did not depend on i in the original version of the model, are replaced for each agent i by the values $(\eta_i, 1 - \eta_i)$: a large η_i corresponds to an agent who prefers not to be isolated.

The agents' heterogeneity adds a significant amount of complexity to the problem, and we have reached an analytical solution of the evolution equations only in the case of pairwise interactions ($\lambda = 1$).

Let us denote by $N_0(t, t', \eta)$ the number of isolated agents with parameter $\eta_i \in [\eta, \eta + \Delta\eta]$ who have not changed their state since time t'. Similarly, we indicate by $N_1(t, t', \eta, \eta')$ the number of agents in a pair joining two agents i and j with $\eta_i \in [\eta, \eta + \Delta\eta], \eta_j \in [\eta', \eta' + \Delta\eta]$, who have been interacting since time t'. The

(a) $b_0 = 0.55$, $b_1 = 0.8$ and $\lambda = 0.9$



(b) $b_0 = 0.7$, $b_1 = 0.7$ and $\lambda = 0.8$

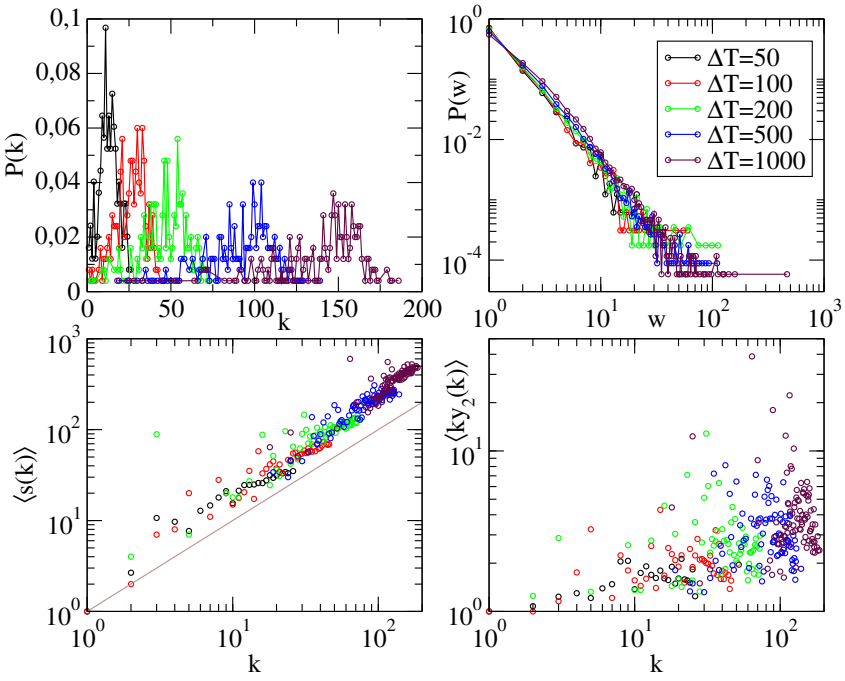


Figure 4.9: Aggregated networks' characteristics for the model with constant number of agents ($N = 250$) for time windows of increasing lengths ΔT and two parameter sets.

mean-field equations for the model are then given by

$$\begin{cases} \dfrac{\partial N_0(t,t',\eta)}{\partial t} = -2\dfrac{N_0(t,t',\eta)}{N} p_0(t,t',\eta) + \pi_{10}^{\eta}(t)\delta_{tt'} \\ \dfrac{\partial N_1(t,t',\eta,\eta')}{\partial t} = -\dfrac{N_1(t,t',\eta,\eta')}{N}\left[p_1(t,t',\eta) + p_1(t,t',\eta')\right] + \pi_{01}^{\eta\eta'}(t)\delta_{tt'} \end{cases} \tag{4.22}$$

With the expression for $p_n(t,t',\eta)$ given by equations (4.21) we find

$$\begin{cases} N_0(t,t',\eta) = \pi_{10}^{\eta}(t')\Big(1+\dfrac{t-t'}{N}\Big)^{-2\eta} \\ N_1(t,t',\eta,\eta') = \pi_{01}^{\eta\eta'}(t')\Big(1+\dfrac{t-t'}{N}\Big)^{-2+\eta+\eta'}. \end{cases} \tag{4.23}$$

The transition rate π_{10}^{η} gives the rate at which agents with $\eta_i \in [\eta, \eta+\Delta\eta]$ become isolated, and $\pi_{01}^{\eta\eta'}$ is the rate at which pairs ij with $\eta_i \in [\eta,\eta+\Delta\eta], \eta_j \in [\eta',\eta'+\Delta\eta]$ are formed. These rates can be expressed as a function of $N_0(t,t',\eta)$ and $N_1(t,t',\eta,\eta')$ according to

$$\begin{cases} \pi_{10}^{\eta}(t) = \displaystyle\sum_{t',\eta'} \frac{N_1(t,t',\eta,\eta')}{N}\left[p_1(t,t',\eta) + p_1(t,t',\eta')\right] \\ \pi_{01}^{\eta\eta'}(t) = 2\displaystyle\sum_{t',t''} \frac{N_0(t,t',\eta)N_0(t,t'',\eta')}{C(t)N} p_0(t,t',\eta)p_0(t,t'',\eta') \end{cases} \tag{4.24}$$

where $C(t)$ is a normalization factor given by

$$C(t) = \sum_{t'=1}^{t}\sum_{\eta} N_0(t,t',\eta)p_0(t,t',\eta). \tag{4.25}$$

To solve this problem with the same strategy used for the model without heterogeneity we make the self-consistent assumption that the transition rates are either constant or decaying as a power-law with time:

$$\pi_{10}^{\eta}(t) = \Delta\eta\tilde{\pi}_{10}^{\eta}\Big(\frac{t}{N}\Big)^{-\beta(\eta)} \tag{4.26}$$

$$\pi_{01}^{\eta\eta'}(t) = \Delta\eta\Delta\eta'\tilde{\pi}_{01}^{\eta\eta'}\Big(\frac{t}{N}\Big)^{-\beta(\eta,\eta')}. \tag{4.27}$$

By using these expressions in equations (4.23) and (4.24), we find the following analytical prediction

$$\begin{aligned} \beta(\eta) &= \max(1-2\eta, \eta-1/2) \\ \beta(\eta,\eta') &= \beta(\eta)+\beta(\eta') \end{aligned} \tag{4.28}$$

and the value of $\tilde{\pi}_{10}^{\eta}$ is given by

$$\tilde{\pi}_{10}^{\eta} = \begin{cases} \dfrac{\Delta\eta}{B(1-2\eta,2\eta)} & \text{if } \eta \le 1/2 \\ \dfrac{\Delta\eta}{B(\eta-1/2,1)} & \text{if } \eta \ge 1/2 \end{cases} \tag{4.29}$$

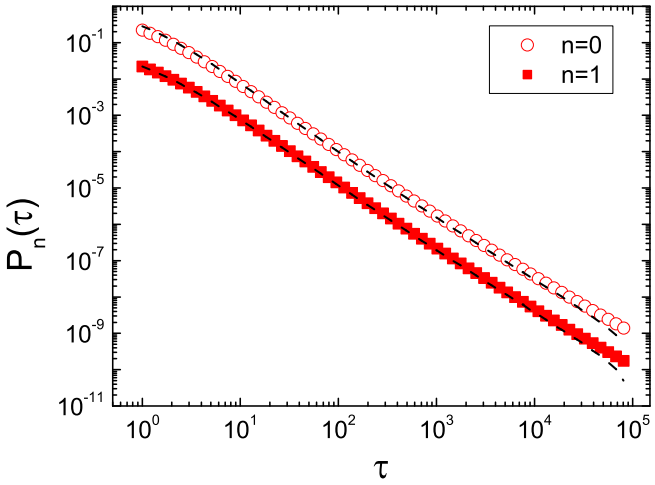


Figure 4.10: Distributions of times spent in state 0 and 1 for the heterogeneous model. The simulation is performed with $N = 10^4$ for a number of time steps $T_{max} = N \times 10^5$. The data are averaged over 10 realizations. The symbols represent the simulation results (circles for $n = 0$ and squares for $n = 1$). The dashed lines represent the analytical prediction. In order to improve the readability of the figure we have multiplied $P_1(\tau)$ by a factor of 10^{-1}.

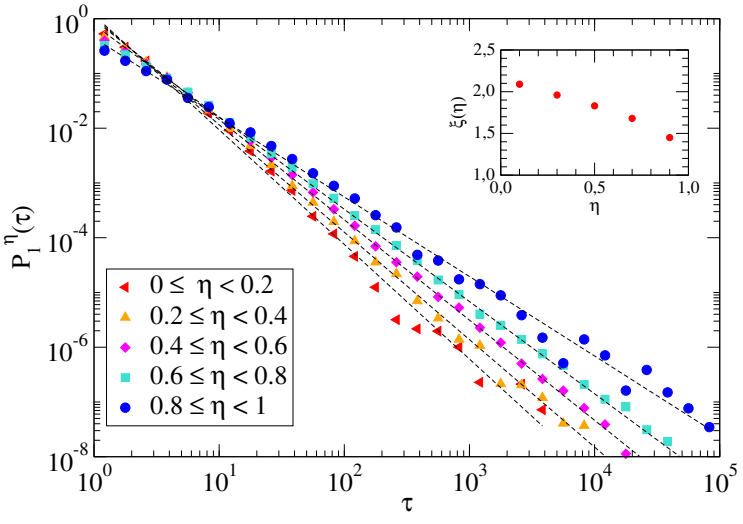


Figure 4.11: Distribution $P_1^{\eta}(\tau)$ of contact durations of individuals with sociability η in the pairwise model (i.e. $\lambda = 1$). Numerical simulations are done with $N = 500$ agents, during $T = 10^5 N$ time steps. Lines are non linear fits $P_1^{\eta(\tau)} \propto (1+\tau)^{-\xi(\eta)}$, with exponent $\xi(\eta)$ reported as a function of η reported in the inset.

In order to check the validity of the mean-field calculation, we study the probability distribution $P_0(\tau)$ of the durations of inter-contact periods and the distribution $P_1(\tau)$ of the durations of pairwise contacts, which are given, when averaged for a total simulation time T_{max}, by

$$P_0(\tau) \propto \int_0^{T_{max}-N\tau} dt \int_0^1 d\eta\, \pi_{10}^{\eta}(t)\, \eta\, (1+\tau)^{-2\eta-1} \tag{4.30}$$

$$P_1(\tau) \propto \int_0^{T_{max}-N\tau} dt \int_0^1 d\eta \int_0^1 d\eta' (2-\eta-\eta')(1+\tau)^{\eta+\eta'-3} \tag{4.31}$$

In Figure 4.10 we compare the probabilities of intercontact time $P_0(\tau)$ and contact time $P_1(\tau)$ averaged over the full population together with the numerical solution of the stochastic model, showing a perfect agreement. In Figure 4.11, moreover, we show the distributions $P_1^{\eta}(\tau)$ of the contact durations of agents with $\eta_i \in (\eta, \eta+\Delta\eta)$. Power-law behaviors are obtained even at fixed sociability, and the broadness of the contact duration distribution of an agent increases with the *sociability* of the agent under consideration.

As previously mentioned, the model can be extended by allowing the formation of large groups, by setting $\lambda < 1$. Power law distributions of the lifetime of groups are again found and, as in the basic model without heterogeneity of the agents, larger groups are more unstable than smaller groups, as $P_n(\tau)$ decays faster if the coordination number n is larger. As the parameter $\lambda \to 0.5$ there is a phase transition and the average coordination number diverges. Overall, the main features of the model are therefore robust with respect to the introduction of heterogeneity in the agents' individual behavior.

4.3.2 Fluxes in the population

A second realistic extension of the model consists in considering a time-varying population size. In real situations, the number of individuals present on a premise fluctuates, reflecting activity patterns (e.g. coffee breaks in conferences, breaks in playground and lunch in a school), circadian rhythm (e.g. day/night alternation) and simply entrances and exits on a premise (e.g. visit flux in a museum). This population fluctuation is for instance well captured by the protocol measuring face-to-face proximity, as presented in chapter 2.

In the model, time-variation of the population size can be adequately modeled by a supplementary individual variable, that can take two values: *present* or *absent*. In the former case, the interaction dynamics of this individual would be the same as described previously. In the latter case, the individual would simply be considered as isolated, initiating no interaction and being forbidden to join any group. Instead of considering a fixed number of agents N, the system is composed at a given time t of $N(t)$ present agents. There are mainly two approaches to model a fluctuating population size:

- in a supervised manner, we can impose a time series for the number of present individuals, for example by taking an empirical time series in a dataset, and

then each time the number $N(t)$ changes, randomly selecting (or with any other criterion) present agents become absent (and stop all their interactions) or conversely, absent agents become present,

- or in a rather unsupervised way, an additional evolution rule can be added, for example with random rates of entrance and exit of individuals, inducing random fluctuations in the population size.

The present agents becoming absent can either become present again, or not, depending on the modeler wishes. It is then possible to obtain by numerical simulations the temporal statistics described previously, such as the time spent by an agent in a given state, or of the duration of contact and inter-contact times, by now considering only *present* agents.

I will briefly report hereafter the implementation of this extension to the model. The followed approach is to impose the number of agents present in the system at each time according to the empirical ESWC09 time series, described in section 2.2. Given the temporal resolution of the dataset of $20s$, and considering that agents act independently, and make choices in a simultaneous way, the empirical time step is identified with the model time step τ, with on average one attempted status update per present agent. After each series of $N(t)$ attempted status updates corresponding to a duration τ in the model description, the manipulation of the number of present agents is considered. If $N(t+1) > N(t)$, some random agents are removed (i.e., put in the *absent* state) in order to match the desired $N(t+1)$. If $N(t+1) < N(t)$, absent agents are introduced into the system, and put into the isolated state. The system evolves in this way for the number of time steps of the empirical dataset.

Figure 4.12 compares the resulting activity patterns between the empirical dataset ESWC09 and the numerical simulations of the model, for two values of the parameter set (b_0, b_1, λ), when $N(t)$ is taken from the empirical ESWC time series. Although only $N(t)$ is imposed to be exactly the same in the model and in the data, the model parameters can be tuned so that other measures (number of isolated nodes, of links, of triangles) remain simultaneously similar to real data. Their highly non-stationary dynamics, similar to those empirically observed, are highly correlated to the node number's. However, distributions of contact durations and of time spent by agents in each state, obtained with the numerical simulation of the model, remain much more stationary, as observed empirically (see section 2.2.1). These distributions are broad, as in the original model with constant number of agents, do not depend strongly on the imposed $N(t)$, and can be superimposed from one time window to the next. Similarly, basic quantities constructed on aggregated contact networks over different time windows, such as node degrees k, the average strength $\langle s(k) \rangle$ and the average participation ratio $\langle kY_2(k) \rangle$ of the nodes of degree k, behave comparably to empirical quantities (confront figures 4.13 and 4.14 to figures 2.5 and 2.6 of chapter 2).

Finally, the versatility of the modeling framework is illustrated by considering the case in which a present becoming absent agent cannot become present anymore. This can adequately model an environment with a stream of coming and leaving

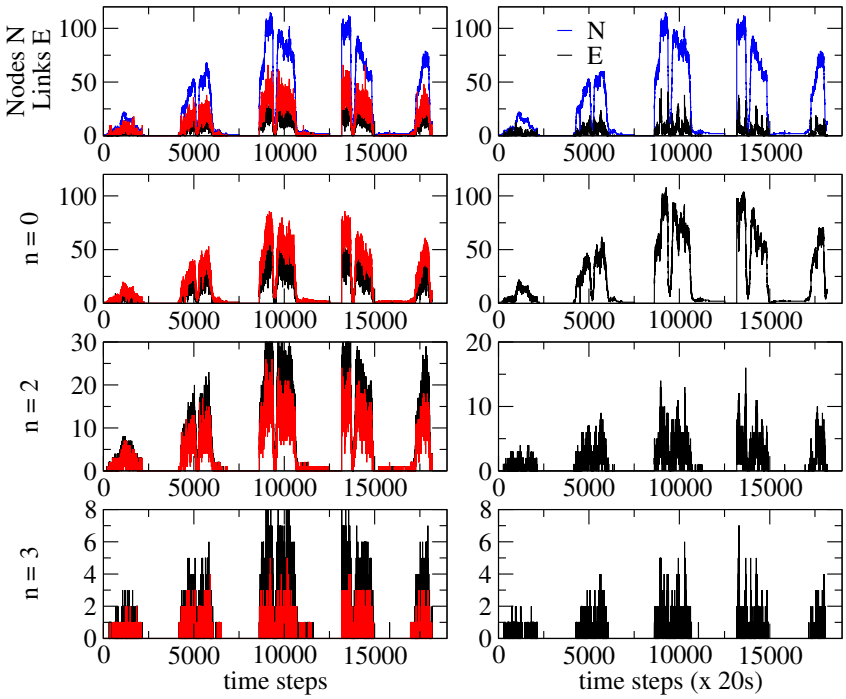


Figure 4.12: Timelines of the number of (from top to bottom): nodes and links in the instantaneous network (top), isolated nodes, groups of 2 nodes, groups of 3 nodes. The left column corresponds to the model with $N(t)$ imposed from the ESWC09 dataset and two sets of parameters, namely $(b_0, b_1, \lambda) = (0.55, 0.8, 0.9)$ (black curves) and $(0.7, 0.7, 0.8)$ (red curves), the right column to the real ESWC09 dataset.

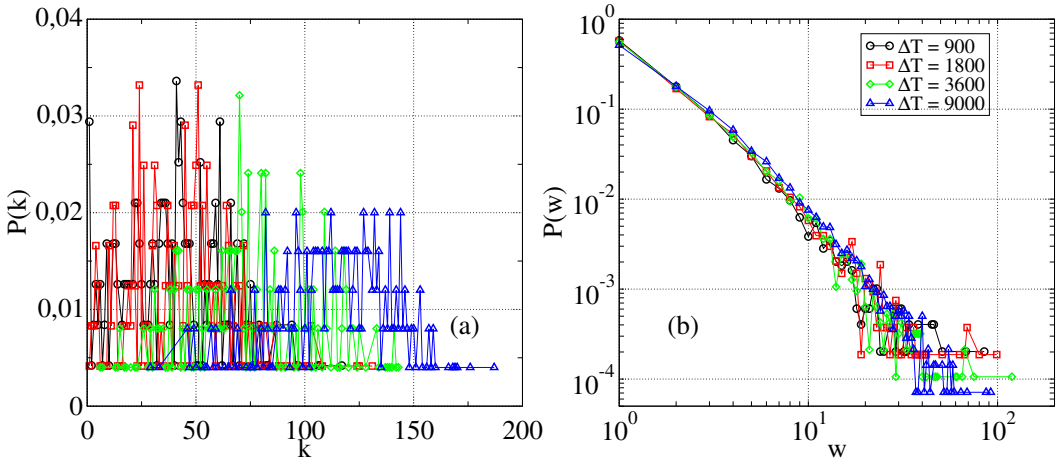


Figure 4.13: Distribution of (a) degree and (b) weight on aggregated networks on various time windows, given by the ESWC09 time series, and $b_0 = b_1 = 0.7$, $\lambda = 0.8$.

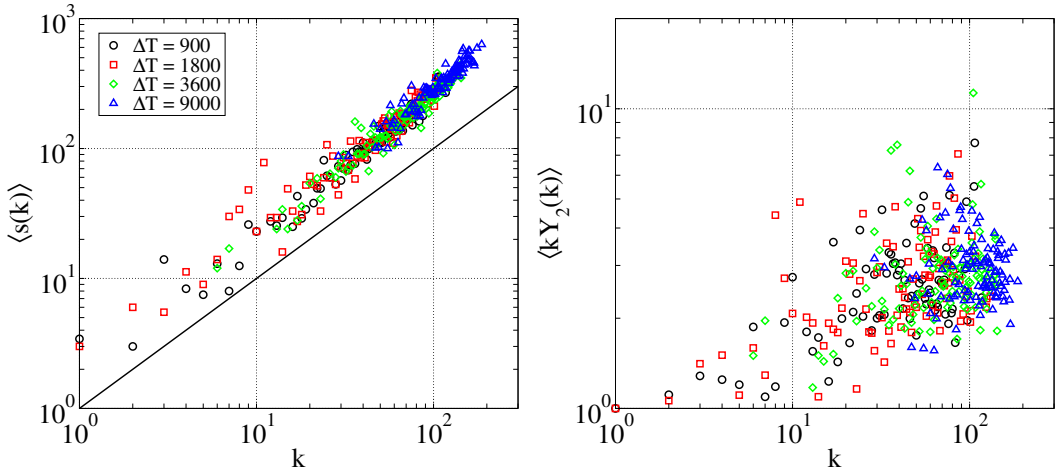


Figure 4.14: Distribution of the average strength of nodes of degree k versus k (left) and (b) average participation ratio of nodes of degree k vs k for aggregated networks on various time windows , given by the ESWC09 time series, and $b_0 = b_1 = 0.7$, $\lambda = 0.8$.

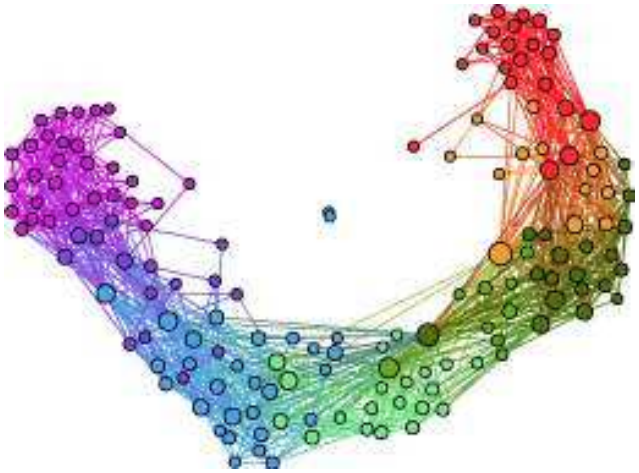


Figure 4.15: Example of aggregated network of 157 nodes obtained by imposing the timeline of the number of agents present at each time during a given day of data gathered during a SocioPatterns deployment at the Science gallery in Dublin described in chapter 2. Here $(b_0, b_1, \lambda) = (0.55, 0.8, 0.9)$, and an agent who leaves the network cannot re-enter it. The nodes are colored according to the entry time of the corresponding visitor, as in chapter 2 (from red to green to blue to violet). The elongated shape of the network is similar to the one observed empirically.

persons, such as a museum. Considering this assumption and the time series of the number of persons in the museum dataset over one day, the model is able to reproduce the elongated shape of the aggregated network whose topology is dictated by the timeline of the visits described in chapter 2 (see figure 4.15).

4.4 Partial conclusions and perspectives

We have presented a model framework of interacting agents that is flexible enough to reproduce the dynamics phenomenology observed during social gatherings collected with wearable sensors presented in chapter 2. The model is treated both analytically and numerically. Its major interest is to propose micro-mechanisms that are able to produce the macroscopic phenomenology of contact dynamics. More precisely, broad distributions in contact durations and intercontact times are explained here with self-reinforcement rules that increase the propensity of agents to remain in the same state (same coordination number) with the time already spent in this state. These rules, which are reminiscent of the preferential attachment model introduced by Barabási and Albert [Barabási 1999] already applied to dynamical processes in [Barabási 2005], exhibit a rich behavior with nonequilibrium transitions between stationary and nonstationary phases. Conversely to some network models such as in [Scherrer 2008, Rocha 2012] in which the distribution of contact durations is determined *a priori*, distributions are here an output of the micro-rules of the model and therefore, these rules can be considered as a plausible explanation of the contact duration features.

We have also shown the versatility of the model. First, rules are flexible enough to produce different kinds of contact and intercontact distributions, such as simple, compressed and stretched exponentials. Second, heterogeneity among agent rules can be easily implemented, and in the case we studied, it does not change the overall phenomenology of the model (broad distributions of contact, intercontact times, lifetimes of groups, nonequilibrium transitions). Third, the model assumes a fixed population size but it can be modified to model more realistic environments in which the population size may fluctuate with time. In this case, the population size can be taken as an input of the model, as given by an empirical time series. We have then shown that the model produces nonstationary dynamical networks whose features are close to the empirically observed ones.

Several research directions can be envisioned at this point. Microscopic rules may be thought of and implemented in order to model more realistically social interactions in various contexts or even animal interactions. For example, merging and splitting of groups such as those exposed in [Johnson 2009] could easily be introduced. It would also be possible to impose individual characteristics, for example the time intervals an agent is present. Although this would happen at the cost of a large input of empirical information, it would produce dynamic networks that retain the details of the presence properties of empirical data set at an individual level. Another interesting outcome of the model with varying population size is that it

makes it possible, starting from an empirical data set that is by definition limited in time, to create a dynamical network on arbitrarily long timescales, with the same properties as the real one, by simply repeating the time-series $N(t)$ as many times as required. This corresponds to an interesting way of creating a non-stationary dynamical network, without having to repeat the *real* sequence of contacts: on each new repetition of $N(t)$, the model will generate a new sequence of contacts.

The most interesting perspective of such a model would be to use it as a support for the simulation of dynamical processes such as information or infectious spreading or synchronizations. The possibility to tune dynamics network properties allows one to investigate finely the effect of these characteristics on dynamic processes, instead of using reshuffling techniques that are difficult to manipulate without any risk of spurious inference due to correlations [Vázquez 2007, Miritello 2011, Rocha 2011, Karsai 2011]. Generating artificial data sets that are based on empirical ones, preserve a certain number of their properties, modify others in a tunable way, and can be extended to large sizes and long times, represent a very important step in such studies.

Chapter 5

Disease spreading

Contents

5.1 Motivation 97
5.1.1 Dynamic processes on networks 97
5.1.2 Modeling disease spreading 98
5.1.3 Toward more realism in contact patterns 103
5.1.4 The need of data on contact patterns 108
5.1.5 Why does contact dynamics matter? 109
5.2 Simulation of an SEIR model on empirical data 112
5.2.1 Data collection 112
5.2.2 Description of the model 113
5.2.3 Results 119
5.2.4 Limitations 124
5.3 Partial conclusion and perspectives 127

5.1 Motivation

In this chapter, we investigate a dynamical process unfolding on a human proximity network. The dynamic process we consider can represent viral or bacterial infections transmitted through close proximity and physical contacts, such as an influenza. In this section, the perspective of dynamical process from a physicist point of view will be outlined. Some historical insight on mathematical modeling of disease spreading is provided and the need of data and more precisely of dynamical data on human contacts that motivates our study is discussed.

5.1.1 Dynamic processes on networks

A dynamical process is given by a system and a set of fixed rules that determines its temporal evolution. Probably one of the most natural spaces in which physicists have studied dynamical processes is the ordinary 3-dimensional Euclidean space and dynamical processes were mainly given by Newtonian mechanics. Different spaces and different dynamical processes can be considered. For example, in electronics, the geometrical space in which the system dynamics is described is often a complex 1-dimensional space ; regular lattices are often used for studying synchronization of

a collection of oscillators. In this perspective, networks too can be considered as a support of dynamic processes : signal may only pass from node to node through edges. The book written by Barrat, Barthélémy and Vespignani presents some dynamic processes that have been studied extensively on networks [Barrat 2008]. A famous example of dynamic process that has been transposed onto networks is the Kuramoto model, describing synchronization between oscillators. In this model, nodes are harmonic oscillators of various pulsation coupled with each others through edges. Depending on the model parameters, the system may either be synchronized or completely unsynchronized. This full dependence of the global behavior on local mechanisms is what characterizes many of the beloved phase transitions of physicists.

An other kind of dynamic processes that have been widely investigated on networks are diffusive processes such as random walks, or reaction-diffusion models. These models are widely studied, principally because they offer a way to analyze real processes such as the propagation of diseases, the spreading of fads, of computer viruses or viral marketing.

Simple disease spreading models such as the SIR model in which individuals are either susceptible, infectious or recovered and in which an infectious individual may transmit a disease to susceptible individuals which whom he/she is in contact with, and recover, can be viewed as reaction-diffusion models. Originally, reaction-diffusion models describe the evolution of a chemical system in which a local chemical reaction such as a spontaneous chemical transformation and a spatial diffusion allowing a substance to spread co-evolve. In a similar manner, the fact that individuals become infected and recover can be considered as local reactions and their moves as a diffusion, as the peer-to-peer interactions make the direct transmission channels change. Such a model is often characterized by the existence of a threshold separating a phase in which epidemics can reach a sizable proportion of the population and an other in which an infection almost surely fades out after few transmission steps without contaminating a significant part of the system.

The importance of the network's topology on the system behavior must be outlined. It happens that for some network models the threshold separating two very different macroscopic behaviors vanishes, as it is the case in scale-free network models. In [Pastor-Satorras 2001] the authors give an analytical argument for this feature, and point out the relevance of such a result for the internet and the web, both networks having a scale-free like distribution of degree. As outlined by Lloyd and May, this feature is less likely to occur in social networks because the degree distribution is expected to be much narrower [Lloyd 2001], with the exception of sexual networks. Those have been found to have an heavy tailed degree distribution [Liljeros 2001].

5.1.2 Modeling disease spreading

According to Serfling and Anderson [Serfling 1952, Anderson 1991], the mathematical description of epidemics dates back to Daniel Bernoulli. In 1760, he published a study of variolation techniques against smallpox [Bernoulli 1760], i.e., the deliberate

inoculation of the virus to prevent epidemics. Mathematical epidemiology has then seen a second development with William Farr in the 19th century. He was indeed amongst the first to think that regularities of epidemic patterns can be described by a mathematical approach and more precisely, as it was the case in other sciences, that a mechanistic explanation could account for observed regularities. Farr thought that this was supported by his analysis in 1840 of the English smallpox epidemic of 1837-1839, in which he showed that the epidemic curve (i.e., the number of new cases versus time) could be well fitted by a Gaussian function. Two decades later, in 1866, he even attempted to forecast the spread of the rinderpest that struck English cattle severely at that time. He thus fitted again the past and actual number of cases versus time (with only 4 records of the monthly number of cases), and made predictions for the successive months. This statistical analysis of historical dataset is recognized as a major step towards the modern mathematical epidemiology.

Whilst Farr's point of view was clearly based on the study of empirical dataset with only the underlying idea that some mechanistic rules were responsible for the observed regularity, the increasing knowledge on bacteriology and epidemiology became gradually incorporated into mathematical epidemiology. This led to two major developments. In 1906, W. H. Hamer introduced the first elements of the well known *mass action principle*. He assumed that the number of new cases would be proportional to the number of infectious individuals, the number of people susceptible to contract the disease and to some constant depending on factors influencing the contagion from an infectious carrier to a susceptible. In 1911, Sir Ronald Ross presented the first differential equation to describe the evolution of a malaria outbreak. This equation takes into account the proportion of infected carriers, the ratio between the number of mosquitoes carrying the malaria, the proportion of mosquito bites leading to an infection and the recovery rate. Ross developed his model in 1915 for more general situations and built a *theory of happenings* relying on a set of differential equations. It is interesting to note that the models developed for contagious diseases were already applied in other scientific fields such as economics and sociology, as it was the case later on.

The third major development of mathematical epidemiology is the introduction of probabilities. In 1928, Lowell J. Reed and Wade H. Frost presented, but never published, a model that includes the possibility for a susceptible person to have contacts with infective persons and not to necessarily become infected. This was the beginning of the stochastic approach. More precisely, in the version of the Reed-Frost model exposed by Helen Abbey [Abbey 1952], the dynamics is discrete. At each time interval, *each individual has a fixed probability p of coming into adequate contact with any other specified individual in the group [. . .], and this probability is the same for every member of the group.* A so called adequate contact would be named an infectious contact today. In the simplest version of the Reed-Frost model, where individuals can only be susceptible or infective, the dynamic equation is given

by

$$\begin{cases} S_{t+1} \sim \text{Bin}(S_t, (1-p)^{I_t}) \\ I_{t+1} = N - S_{t+1} \end{cases} \tag{5.1}$$

where S_t and I_t are respectively the number of susceptible and infective individuals at time t, N is the total number of individuals and $\text{Bin}(a, b)$ is the binomial law with parameters a and b.

This discrete time model has later been extended to a continuous time version, in which the contamination of a susceptible individual is described by a Poisson process of rate proportional to the number of infectious carriers (often called the standard SI model). Moreover, discrete time models can be used as an approximation of continuous time models, when time steps become infinitely small, or more precisely, very small compared to the typical time scales of the spreading dynamics. Some links can be made between deterministic and stochastic models. For example in [Andersson 2000], a demonstration shows that, in the case of large populations, the expected course of a standard SIR stochastic model (in continuous time) may be approximated by the deterministic model presented in [Kermack 1927] as given below in equation (5.2). The finite size case is studied in [Fierro 2010].

The field of epidemiology faces two major difficulties that have been long recognized and relentlessly outlined. The first is concerned with the quality of data to test and estimate models. In 1952, Serfling expressed the following wish, that is very illustrative for a physicist [Serfling 1952]: he wished an epidemiological Tycho Brahe who would collect data of unquestionable accuracy. The second difficulty is to find an adequate representation of reality. Can one be satisfied with the oversimplification of contact patterns by a single parameter? What about cultural and social varieties? Seasonal effects? Varying environmental conditions? The pure understanding of mechanisms is clearly not the first objective in the field. The main one is to be able to make predictions and to design effective public health policies.

One can distinguish three rather separate groups in the literature of mathematical epidemiology that deals with three separate stages in the scientific approach. The first is model design, often based on phenomenology. The second is more theoretical and deals with the understanding of mathematical properties of models. The third kind of literature concerns estimation of models and is composed of sophisticated statistical methods to evaluate various models and make predictions.

5.1.2.1 Compartmental models

Most of the epidemiological models are compartmental models, i.e. they rely on the assumption that the epidemiological status of individuals can be classified in a finite set of categories, called compartments. These categories exchange incoming and outgoing fluxes with each other and in the case of births, deaths and travels, incoming and outgoing fluxes can feed the system.

In general, at least two categories exist. One is the infectious category that comprises of individuals able to transmit the disease. The other is the susceptible category and comprises of individuals susceptible to contract the disease. Because

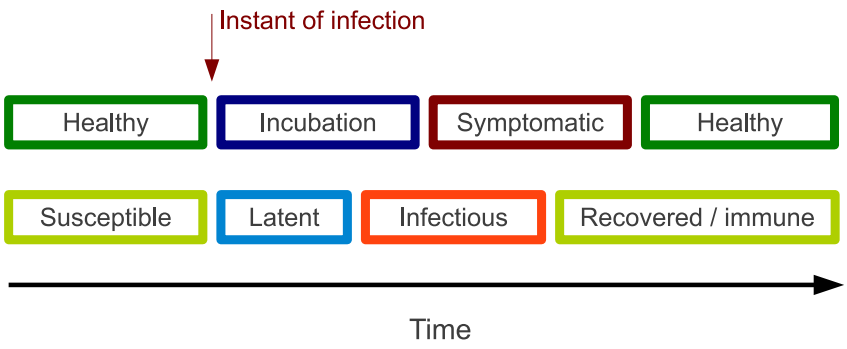


Figure 5.1: Schematic view of the difference between the symptoms (above) and the infectious state (below). Note that the infectious period and the symptomatic period are not necessarily synchronous.

one may want to include heterogeneity in the population concerning the immune system for example depending on age, these two categories may be split into subcategories. Besides, when one is interested in the statistical estimation of epidemiological parameters of specific diseases, it may be worth splitting the infectious category into symptomatic carriers, that can be detected, and asymptomatic carriers, that cannot be detected at first sight. In general, data based on reports of medical doctors (sentinel surveillance data) and self reported data cannot unveil the prevalence of asymptomatic cases. Only so called serosurveys, that consists in measuring the antibodies of a specific infection, can be used to estimate the overall prevalence, and a fortiori the prevalence of asymptomatic. As pointed out in [Dowse 2011], the clinical attack rates extracted from sentinel surveillance data and the overall prevalence measured by serosurveys differ considerably in the case of the H1N1 influenza. It indicates that there is a substantial proportion of asymptomatic cases or mild infections. Some experiments have been run in order to estimate properly the asymptomatic prevalence [Carrat 2008]. Moreover, the infectious period and the symptomatic period are not necessarily synchronous, as illustrated by the figure 5.1, that represents the difference between the infectious state and the symptoms. A well-known example is the case of HIV-AIDS: while individuals may become infectious rapidly after the infection, it can take years for symptoms to appear. The table 5.1 gives some estimation of these different periods for various infectious diseases. For most of them, the incubation period lasts longer than the latent period, indicating that an infectious person can contaminate whilst symptoms are not already visible. In some instances such as the whooping cough and the diphtheria, the opposite occurs.

5.1.2.2 Quantities of interest

Three major quantities are often looked at when analyzing empirical, simulated or theoretical outbreaks.

Infectious disease	Incubation period	Latent period	Infectious period
Measles	8–13	6–9	6–7
Mumps	12–26	12–18	4–8
Whooping cough	6–10	21–23	7–10
Rubella	14–21	7–14	11–12
Diphtheria	2–5	14–21	2–5
Chicken pox	13–17	8–12	10–11
Hepatitis B	30–80	13–17	19–22
Poliomyelitis	7–12	1–3	14–20
Influenza	1–3	1–3	2–3
Smallpox	10–15	8–11	2–3
Scarlet fever	2–3	1–2	14–21

Table 5.1: Incubation, latent and infectious periods for various infectious diseases. Table from [Anderson 1991].

The first is the **basic reproductive number** (also called the basic reproductive rate). It is defined as the average number of individuals a single infectious individual would directly infect in a fully susceptible population before recovering [Anderson 1991]. Intuitively, this quantity, denoted by R_0, gives an idea of the capacity of a disease to spread. For example if this quantity is much larger than 1, then from a single initially infected individual, a generation of R_0 new infected persons is expected on average. If the *neighborhood* (whose precise definition depends on the model, but can be understood in general as the set of persons an individual can directly infect) of this second generation is still composed of a high proportion of susceptible persons, then the third generation will be on average larger than the second (and on average equal to R_0^2 if the number of susceptible neighbors of each member of the second generation is larger than $R_0 + 1$ and if these neighborhoods do not overlap with each others). On the contrary, if R_0 is much lower than 1, it is rather unlikely that the first infected person will contaminate anyone else, and may it be the case, the probability that a second generation of infective exists and contaminates itself a third generation is even lower. In that case, the spreading will stop very rapidly. This basic reproductive number is of major interest for epidemiologists because it is a synthetic indicator of the virulence of an outbreak and in some models, it defines two regions in the phase space in which the system behaves very differently. Two diseases, one that is very contagious but from which one recovers very fast and an other that is not very contagious but whose period of infectiousness lasts longer, may have the same basic reproductive number value, meaning that they propagate with a similar amplitude (but not the same pace).

A second quantity of interest is the **size of the outbreak** which corresponds to the final number of cases when the outbreak is over. The proportion given by the final number of cases divided by the population size is called the attack rate

(AR). For a past epidemics, it is given by the proportion of persons who have caught the disease. In a disease model, which does not necessarily have an empirical counterpart, it is generally given by a random variable, whose distribution and average measure its risk. For example, Dowse et al. have estimated that about 25% of preschool children and 40% of school-aged children have been infected by the (H1N1) influenza in Western Australia during winter 2009. If the mortality had been comparable to the one of the 1918-1920 pandemic, the number of mortal cases would have been much higher. Some researchers have estimated that the 1918-1920 influenza virus would have killed approximately 62 millions of people in 2004 [Murray 2007].

The last major quantity of interest is the **time of the epidemic peak**, i.e., the instant the number of new cases reaches its maximum. It gives an indication on the temporal scale of the spread and may be of interest for planning public health policies. For example, it may be worth trying to delay the peak time in order to implement a mass vaccination. Balcan et al. have investigated the effect of antivirals on the peak time for the A(H1N1) influenza outbreak in 2009. For instance, they estimated that 5 and 10 millions of antivirals could respectively delay the peak time of about 4 weeks for Spain and Germany [Balcan 2009].

Empirically, these quantities are estimated from data and according to an underlying spreading model, but technically, the task is far from trivial [Anderson 1991, Andersson 2000, Diekmann 1990, Heffernan 2005, Breban 2007]. For example the estimation of the basic reproductive number can be determined in the early stage from the growth rate of the number of infected individuals (see subsection 5.1.3.1 for more details). On the contrary, in numerical simulations, these quantities of interest can be directly computed with successive Monte Carlo runs.

5.1.3 Toward more realism in contact patterns

According to the terminology Abbey used in the description of the Reed-Frost model, infectious diseases propagate through *adequate contacts* [Abbey 1952]. The precise definition of an adequate contact depends on the studied disease. It may be a sexual intercourse (e.g. HIV/AIDS, hepatitis B or syphilis), a direct physical contact (e.g. syphilis, chickenpox) or simply a physical proximity, as many diseases are transmitted by droplets emitted while coughing, sneezing or even speaking (e.g. mumps, influenza, smallpox, chickenpox, rubella, tuberculosis, SARS, measles, common cold). Some other ways may exist, such as oral transmission (for example through kissing), fecal oral transmission usually through contaminated water, by injection or transplantation of contaminated material, or trough vectors such as some types of mosquitoes (e.g. malaria, chikungunya). Models applied to a particular type of disease must take the specificity of the transmission route(s) into account.

As the reader may have noticed in the definition of the basic reproductive number (see previous subsection 5.1.2.2), contact structure is crucial for the spreading of infectious diseases, but the variety of situations makes the modeling task difficult. On the one hand, the modeling procedure consists in making appropriate simplifi-

cations in order to extract relevant mechanisms but on the other hand, too much simplification may over-reduce the natural variety of situations and create a gap between theory and reality which prevents the models to have predictive properties. This explains why a wide range of models of contact patterns exist in epidemiology, from the simplest to the more realistic, as illustrated by figure 5.2. The families of models are presented hereafter. They can adequately describe the dynamics, depending on the geographic scale and the type of disease that is considered, as shown in Riley's review on the large-scale models for four different diseases [Riley 2007].

5.1.3.1 Homogeneous mixing

The simplest way to model contacts in a population is to consider that any individual is in permanent contact with all others and that the transmission between an infectious individual and a susceptible one is described by a unique and independent process. In other words, any individual can potentially be infected by any infectious individual, with the same strength. This hypothesis is known under the name of **homogeneous mixing** [Anderson 1991].

This assumption has been widely used and allows to obtain informative results on the behavior of spreading with very few parameters. For example with a simple SIR Reed-Frost model in continuous time, in which the contamination of a susceptible individual by an infectious individual is described by a Poisson process of rate β and the transition from susceptible to recovered is described by a Poisson process of rate ν, Williams has shown that in the large population limit, the behavior of the system is determined by the value of the basic reproductive number R_0 [Williams 1971]. If N is the number of initially susceptible individuals in this model and a is the number of initially infected individuals, R_0 has a simple expression $R_0 = \beta N/\nu$. In the limit of $N \to \infty$, if $R_0 \leq 1$ the probability that a *true epidemic* occur is null but if $R_0 > 1$, a true epidemic occurs with probability $1 - R_0^{-a}$. A true epidemic is for Williams an outbreak in which the number of infected people becomes infinitely large (in this infinitely large population). The value of the ba-

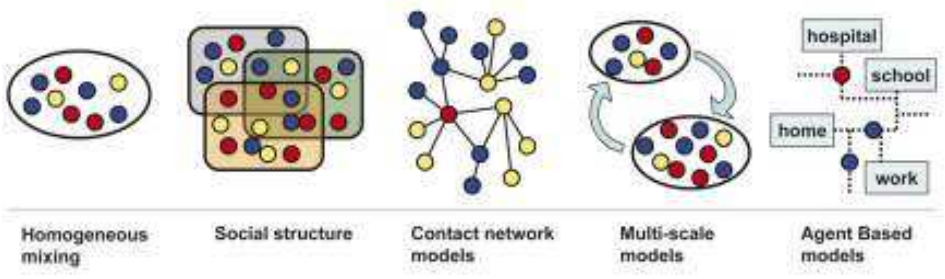


Figure 5.2: Different models of contact patterns, with various levels of realism. Individuals are represented by circles of different colors depending on their epidemiological state. This figure comes from [Colizza 2007b].

sic reproductive number defines an epidemic threshold. The reader can find more details on stochastic epidemiological models in [Andersson 2000].

Interestingly, the same result on the existence of this epidemic threshold remains true for deterministic models. If $S(t)$, $I(t)$ and $R(t)$ are respectively the number of susceptible, infectious and recovered individuals, a deterministic continuous time model, under the homogeneous mixing hypothesis, is described by the following set of differential equations [Anderson 1991, Kermack 1927]:

$$\begin{cases} \dfrac{\mathrm{d}S(t)}{\mathrm{d}t} = -\beta S(t)I(t) \\ \dfrac{\mathrm{d}I(t)}{\mathrm{d}t} = \beta S(t)I(t) \quad -\nu I(t) \\ \dfrac{\mathrm{d}R(t)}{\mathrm{d}t} = \quad \nu I(t) \end{cases} \tag{5.2}$$

In this model, it can be shown that if $R_0 = \beta N/\nu < 1$, the infection dies out, but if $R_0 > 1$, a spread occurs, infecting a sizable proportion of the population, and that the early stages are approximately described by an exponential increase of the number of infectious individuals at rate $\nu(R_0 - 1)$. This result is sometimes used in order to estimate the basic reproductive number, even if the number of cases must be high enough to make inference on the proportion (the discrepancy due to individual fluctuations is often a difficulty) and low enough to be in the early stages of the spread where the exponential growth approximation holds [Heffernan 2005].

5.1.3.2 Social structure

The evident limitation in the homogeneous mixing hypothesis assumption for contact patterns is to consider that any pair of persons in the population can be in contact and have the same probability to transmit a disease. No heterogeneity in the population is considered: probabilities of being contaminated and recovering could depend on age for example, and the probability of different persons having an adequate contact could depend on individual characteristics such as their age or their occupation.

These heterogeneities are empirically observed. For example, the first evidence of an age structure in contact patterns is reported in a preliminary questionnaire-based study of Edmunds et al. with 92 adults [Edmunds 1997]. This was validated in a further study with 7 290 participants of eight European countries who answered surveys about their face-to-face and physical contacts in everyday life: people interact most with other individuals of similar age, especially for children and young adults [Mossong 2008].

A common way to include structure between groups in contact patterns is to use *who acquires infection from whom* matrix (noted after WAIFW matrix). For example in [Anderson 1991], an age structure can be defined with n age groups and the WAIFW matrix is an $n \times n$ matrix (β_{ij}) where β_{ij} is the rate of infection of a susceptible in age class i by an infective in class j. Generally this WAIFW matrix

is symmetric. This model of contact structure can induce different vaccination strategies according to the risk of each group. This method can be applied to various contact structures. For example in the case of sexually transmitted diseases, it may be a way to distinguish between people having no sexual activity, people with a moderate activity and people with a very-high activity, because they do not face the same risk. Similarly the distinction between homosexual, bisexual and heterosexual patients is quite common in the study of HIV/AIDS.

5.1.3.3 Network-based models

The WAIFW matrix method can be applied to define heterogeneities due to geographical constraints or groups inside the population, but it assumes that inside a group, all individuals have homogeneous contact patterns. This is a simplifying assumption, because in reality, such groups are not homogeneous. Some members may have very few contacts while others may have many contacts, even if they share the same category attribute (age, origin, gender, sexual orientation). The number of groups can be increased in order to reduce the heterogeneity inside each group, but this does not solve the methodological drawback.

One way to deal with heterogeneities in the duration and frequency of contacts is to adopt a network perspective. Unlike the homogeneous mixing model assuming that everyone has the same probability to catch the disease if there is at least one infected person in the population, the contact matrix approach in which this probability depends only on the categories, the network approach assumes that an individual can be infected only if one of its neighbors (persons it has an adequate contact with) has the disease [Keeling 1999]. The network structure is defined *a priori*, before considering any disease transmission (the reconstruction *a posteriori* of the infection path between infected persons is known as the contact tracing method). This constrains a disease to propagate from individual to individual only through links corresponding to the existence of an adequate contact between these two specific individuals. Heterogeneities in the number of adequate contacts can then be introduced in the degree distribution of the network. This source of heterogeneity is known to be important for the course of an epidemic. For example, Eames and Keeling [Eames 2002] showed that a theoretic static network model which contains degree heterogeneity and small world features can not be reducible to a WAIFW contact matrix with degree categories. The degree heterogeneity is such an important quantity that even the existence of an epidemic threshold depends on it. Theoretically, if a SIR spreading occurs on a network with a degree distribution following a power-law of exponent between 2 and 3, the epidemic threshold vanishes [Pastor-Satorras 2001]. Networks with degrees following such a distribution are found in sexual networks [Liljeros 2001] even though the statistical estimation is subject to criticisms [Clauset 2009, Jones 2003, Stumpf 2012].

Degree heterogeneity is not the only feature that makes the network assumption relevant for contact patterns (see [Keeling 2005] for a review). The clustering and presence of well-identified communities of individuals is also known to impact

the course of an epidemic, as it reduces the number of susceptible from which an individual can acquire directly the disease [Rocha 2011, Yoneki 2008, Zaric 2002, Szendrői 2004, Smieszek 2009b, Eames 2008].

5.1.3.4 Mixed multi-scale models

The relevance of taking air transportation into account for studying large-scale pandemics has motivated mixed multi-scale model [Rvachev 1985, Hufnagel 2004, Colizza 2006b, Colizza 2006a]. In these models, the world is divided into small cells of a given population size, given by census data. In each of these cells, a homogeneous random mixing models gives the local evolution of the disease spreading. In some recent models, the population of each cell can travel in neighboring cells, mimicking commuting travels, and produce an infection from one region to the adjacent ones. Infection between non adjacent regions is allowed by the air transportation network. This modeling approach allows scientists to test the potential effectiveness of public policies, such as a reduction of air transportation flows or specific allocations of antivirals. These mitigation strategies have such economic and human consequences that a global trial is of course not feasible. On the other hand, numerical simulations allow for these tests and it was shown that for example, reduction of air transportation flows is not very efficient [Colizza 2007a, Bajardi 2011].

5.1.3.5 Agent based models

More and more details can be incorporated in models. The most informed way is to introduce all knowledge on human contacts at the individual level. This method is called agent based modeling. It is not specific of epidemiology and has been widely used in other disciplines, such as in social sciences [Schelling 1971]. In the context of disease spreading, each individual is simulated separately, and its behavior is defined according to its individual characteristics (e.g. age, gender).

At the most detailed level, the behavior consists in a time schedule of places to visit (school, office, household). A disease can spread from an infectious individual to a susceptible, only if they both visit the same place at the same time. In general, behavioral rules specific to each individual define the contact patterns in the population. These models are often simulated with a synthetic population whose demographic characteristics reflect census data. They sometimes describe a wide geographical area with a possibly varying population density, including commuting information relying on air travel data. Even though they can include very detailed information on age structure mixing, on school sizes or on workplaces, some informed approximation still needs to be done (for example for the school allocation, the individual workplace choices).

This type of modeling for disease spreading has been applied at the scale of cities [Eubank 2004], countries [Ciofi degli Atti 2008, Germann 2006, Ferguson 2006] or even at a more global scale [Longini 2005, Merler 2010]. At a global scale, a recent side-by-side comparison between this type of approach and a global multi-scale

model showed an excellent agreement between these two methods [Ajelli 2010]. This is relevant for policy makers, because very sophisticated models do not necessarily add a significant amount of precision compared to simpler models, provided that the latter still incorporate all key elements.

5.1.4 The need of data on contact patterns

As Serfling advocated for an epidemiological Tycho Brahe [Serfling 1952], he pointed out to one of the major difficulties in the modeling of disease spreading, i.e. the lack of knowledge on contact patterns. This requires knowing about the interactions of any individual, the variations between individuals, the correlations with individual characteristics such as gender or age, with the type of location where the interaction takes place (e.g. office, household). As Keeling and Eames underline in their review on networks in epidemiology, this is an *impractically time-consuming task* [Keeling 2005].

Epidemiologists designed indirect and direct methods to tackle this issue. The former are generally based on the estimation of each element of the WAIFW matrix using observed seroprevalence data [Anderson 1991]. So called *time-use data* analysis as in [Zagheni 2008] would be classified as an indirect method as well because it consists in asking respondents about the chronological sequence and duration of their daily activities and to ask how many participants take part to these activities. From this information, a co-presence matrix can be inferred, which can be considered as a fair approximation of the WAIFW matrix.

Direct methods cover mainly three techniques: infection tracing, complete contact tracing and diary-based studies. Infection tracing consists in identifying for each case the person who transmitted the disease [Haydon 2003, Riley 2003]. From all possible contacts, only those who lead to an infection are identified. All others contacts that may have been able to transmit a disease are left out. This method is very helpful to estimate relevant quantities such as the basic reproductive number for observed outbreaks.

Complete contact tracing methods are common for sexually transmitted diseases. For example in [Liljeros 2001], volunteers declared the number of sexual partners up to the time of the interview. Some studies have also investigated on the type of sexual intercourse and on individual characteristics on the partner(s). As sexual intercourses are a sensitive subject, these studies may suffer from a bias in volunteering. It can be pointed out that this contact tracing method may be a powerful health-care policy: it is used in Cuba for decades against HIV/AIDS and among other factors, it may explain the relatively lower prevalence of disease in this country [Hsieh 2002, Hsieh 2010].

In the context of airborne transmitted diseases, the diary based method is certainly the most used [Read 2008, Mossong 2008, Wallinga 2006, Edmunds 1997, Beutels 2006, Hens 2009]. It consists in asking people on the contacts they had on a limited number of snapshots in time, usually 1 day or typical week [Wallinga 2006, Beutels 2006]. Contacts are often very precisely defined: skin to

skin contact such as a kiss or a handshake, or a two-way conversation with three words or more in the physical presence of another person but no skin-to-skin contact. Additional information may be asked such as the age, the sex of each partner, the location of the contact, the total duration spent together or the frequency of contacts with some individuals (daily or almost daily, about once or twice a week, about once or twice a month, less than once a month or for the first time) [Mossong 2008]. Generally, as these informations are provided via self-reported diaries and are time-consuming, they are subject to an uncontrolled bias and a lack of representativeness due to cognitive limits (see a dedicated chapter on this subject in [Knoke 2008]). This problem starts being quantified through the comparison of declarations of both members of the interactions [Smieszek 2011] but it would require an entire cross-method study to accurately identify the difference between what individuals declare and what happens. This is especially important because even random contacts of very short duration (for example in public transportation) may transmit an infection. Other limits of these self-reported studies include the often limited number of participants, except few large-scale studies such as in [Mossong 2008] to which 7290 participants across different European countries took part, the absence of longitudinal analysis and the relative short period on which respondents are interviewed. For example, Eames et al. showed how different mixing patterns are for children between school time and holidays [Eames 2011].

New technologies such as those described in section 1.2.2 are promising with respect to the limits described above. Some studies have been designed with epidemiological objectives for tracking proximity between individuals [Salathé 2010, Isella 2011, Stehlé 2011b]. Not only they are not as limited as diary-based studies in terms of the number of participants and do not suffer the cognitive problem of recalling past interactions, but they also allow to analyze with a high temporal and spatial resolution the dynamics of encounters, such as the variations in the durations and frequencies of the contacts and the existence of causality constraints in the possible chains of transmission.

5.1.5 Why does contact dynamics matter?

As pointed out in [Smieszek 2009a], who used numerical simulation of disease spreading on diary-based data on contact patterns, and in a theoretical study in [Smieszek 2009b], variations in the durations and frequencies of the contacts may affect the course of epidemic spreading. These heterogeneities, even if clustering may dampen the effects, may create preferential contagion paths, while others may less likely exist, which can produce differences in the prevalence compared to a situation where encounters are considered as equal (which is often the only possible way to consider encounters in self-reported analysis that do not provide information on the duration and the frequency of contacts).

Beside the contact duration heterogeneity, Vazquez et al. pointed out that another type of dynamic heterogeneity may affect spreading processes [Vázquez 2007]. They showed with randomization procedures that the long-tailed distribution of in-

tervals between contacts induces a prevalence decay time significantly larger than predicted with standard Poisson distribution. In [Miritello 2011], it was shown with a similar randomization procedure that the burstiness of relay time intervals (time elapsed between two successive contacts with different partners) and interevent time intervals may produce larger or smaller outbreaks than if they followed an exponential distribution, depending on the transmission parameter value. Rocha et al. showed that temporal correlations in their sexual intercourse dataset generally increased outbreak sizes [Rocha 2011]. Karsai et al. [Karsai 2011] pointed out with sophisticated randomization methods but in the case of a simplified deterministic disease model (with an infection occurring systematically at any contact between an infective and a susceptible), that burstiness in interevent time intervals slows down the propagation while all other correlations accelerate the spreading. These randomization methods are tricky because many correlations may be lost in a the reshuffling process and inferring which one of those is responsible of a faster or slower spreading requires to have two randomization procedures differing only for this specific correlation. Nonetheless, they outline the possible impact of temporal correlations on contagious process via the dynamics and the prevalence amplitude. An online large-scale experiment described in [Iribarren 2009] indeed showed the role played by the large heterogeneity found in the response time for the information spreading pace, and in some aspects, information spreading shares many similarities with disease spreading (it can be argued that information spreading differs from disease spreading in the intentionality of agents).

A last aspect of contact dynamics is the existence of temporal constraints that may prevent propagation paths which would be allowed in static aggregated networks. The following example with three nodes may illustrate this issue. Consider the following sequence of events: an individual A interacts first with an individual B who then interacts with a third individual C who never interacts back with A. In that case, a disease initially infecting A can spread from A to B and from B to C. Here we consider a disease, but it may be a piece of information such as a gossip. If on the contrary, individuals B and C interact first, and then A interacts with B, then the disease infecting A cannot reach C. If one considers a time aggregated contact network which disregards the information on the contact order, the transmission from A to C is always possible because B is in contact with both A and C.

We have examined this feature in [Isella 2010] within a deterministic snowball SI model on daily networks collected in a museum and at a conference (this dataset is described in detail in section 2.2) and taking into account, or not, the time order of contacts (with a daily aggregated contact network versus a 20 sec time resolved contact network). All individuals are considered as susceptible at the beginning. An infected seed is chosen at random. Every contact between a susceptible individual and an infected one, no matter how short, results in a transmission event in which the susceptible becomes infected and never recovers. By varying the choice of the seed over individuals, a distribution of the number of infected individuals is obtained at the end of each day. The transmission events can be used to define the network

along which the infection spreads (i.e. the network whose edges are those who have led to an infection), also called the *transmission network*.

Due to causality, the infection can only reach individuals present at the venue after the entry of the seed. As a consequence, in the following we will use the term *partially aggregated network* to indicate the network aggregated from the time the seed enters the museum/conference to the end of the day. We note that the partially aggregated network defined in this way can be dramatically different from (much smaller than) the network aggregated along the whole day.

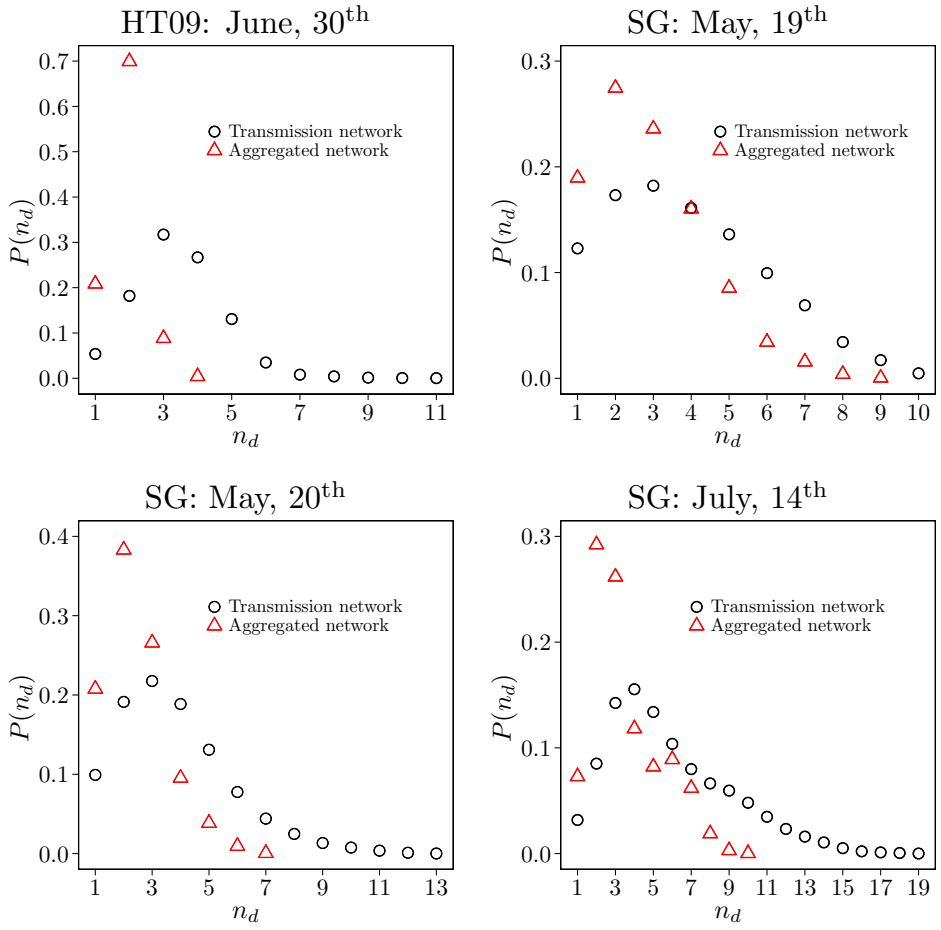


Figure 5.3: Distribution of the path length n_d from the seed to all the infected individuals calculated over the transmission network (circles) and the partially aggregated networks (triangles) for 3 days of the museum dataset and one day of the HT conference. The distribution are computed, for each day, by varying the choice of the seed over the individuals.

Effects of causality constraints are examined through the comparison between snowball SI spreading along the transmission network and on the partially aggregated network. Figure 5.3 reports the distribution of the network distances n_d

between the seed and every other infected individual. When calculated on the partially aggregated network, n_d measures the length of the *shortest* seed-to-infected-individual path, whereas it yields the length of the *fastest* seed-to-infected-individual path when calculated on the transmission network. We observe that the length distribution of fastest paths, i.e., the $P(n_d)$ distribution for the transmission network, always turns out to be broader and shifted toward higher values of n_d than the corresponding shortest path distribution, i.e., $P(n_d)$ for the partially aggregated network. The difference is particularly noticeable in the case of May 20$^{\text{th}}$ and July 14$^{\text{th}}$ for the SG deployment, and June 30$^{\text{th}}$ for the HT09 conference, where the longest paths on the transmission network are about twice as long as the longest paths along the partially aggregated network.

This result underlines that in order to understand realistic dynamical processes on contact networks, information about the time ordering of the contact events may in some cases be essential: the information carried by the aggregated network may lead to erroneous conclusions on the spreading paths.

5.2 Simulation of an SEIR model on empirical data

A crucial point in the mathematical modeling of disease spreading concerns the level of detail that should be incorporated in models. The trade-off between very simple contact models, such as the homogeneous mixing model, and very detailed models such as in agent-based models depends on the studied question. Very detailed models often lack transparency and this prevents discrimination between different effects, while on the other hand, too simple models may not capture relevant features.

In [Stehle 2011a] we have looked at the role of temporal aspects (heterogeneities and temporal constraints) with a 2-day conference dataset obtained with the protocol described in section 1.2.3. Three contact models that capture different amounts of available knowledge on the dynamics allow one to assess the effect of heterogeneities in contact durations and of temporal constraints on the prevalence and on the spreading dynamics, for a rapidly spreading but realistic disease model. This may be relevant for identifying the level of detail needed for contact data to adequately and realistically inform modeling approaches applied to public health problems, as underlined in a commentary on our study [Blower 2011].

5.2.1 Data collection

Contact data for this study was recorded following the protocol described above in section 1.2.3 at a French conference about Hospital Hygiene in Nice (France), called the SFHH conference. 405 volunteers out of roughly 1200 conference attendees participated to the deployment. An ethic committee of Lyon University Hospital approved the protocol and volunteers signed an informed consent when accepting to carry the sensors. Data was treated anonymously. The radio range of the order of 1.5 to 2 meters corresponds to adequate contacts for airborne transmitted diseases such as influenza or whooping cough.

The deployment lasted 2 days, from 9 am to 9 pm the first day and from 8:30 pm to 4:30 pm the second day. Contacts outside these time periods (denoted as *nights* hereafter) and outside the conference premises were not recorded.

The dataset consists of 28,540 contacts of an average duration of 49 seconds and a standard deviation of 112 seconds. As the ratio between the average duration and the standard deviation already seem to indicate, figure 5.4 shows that contact durations are broadly distributed, without any typical scale emerging from the distribution. This feature is observed in all other datasets (see section 2.2.1).

5.2.2 Description of the model

From the contact sequence between individuals that includes temporal information, three different contact pattern models are defined and described in detail below. They correspond to different levels of information. On top of each of these contact pattern models, a compartmental SEIR model is simulated. A so-called compartmental model is analogous to a SIR model, described in paragraph 5.1.3.1, with the exception that susceptible individuals contaminated by infectious ones enter in an *exposed* state in which it stays for a time period before becoming infectious and be able to transmit the disease. This exposed state corresponds to the latent period illustrated in figure 5.1. Such a model is preferred to SI, SIR or SIS models because it gives a realistic but simple enough description of an influenza-like disease. The objective is to study very general properties of contact pattern models and not to provide predictions for real diseases: the inclusion of more compartments, such as asymptomatic individuals, is then not necessary. The three stages of the epidemic process are stochastic. First the infection process of a susceptible by an infective is described by a Poisson process of rate β. After being infected, the susceptible enters a latency period of an exponentially distributed duration, of parameter σ. It then becomes infectious during another exponentially distributed duration of parameter ν. This compartmental model is sketched in figure 5.5.

5.2.2.1 Homogeneous network model

With the method described in subsection 2.2.2, we construct a daily aggregated network of behavioral relations from the contact sequence. Nodes of this network represent individuals, and pairs of nodes are connected by an edge if the corresponding participants have been in face-to-face proximity at least once over the day. No information is retained on the time order and the duration of these contacts. This defines an homogeneous network in which a disease can be transmitted from a node to its neighbors, at a constant rate β_{HOM} during the time period contacts are considered as active. It is denoted as HOM hereafter.

Some standard statistics are summarized hereafter (definitions of quantities are given in 2.1). On average, a node has 30 neighbors with a distribution decaying exponentially for large numbers. The average clustering coefficient is 0.28 to be compared to an average value of 0.07 in a random Erdös Rényi graph of the same

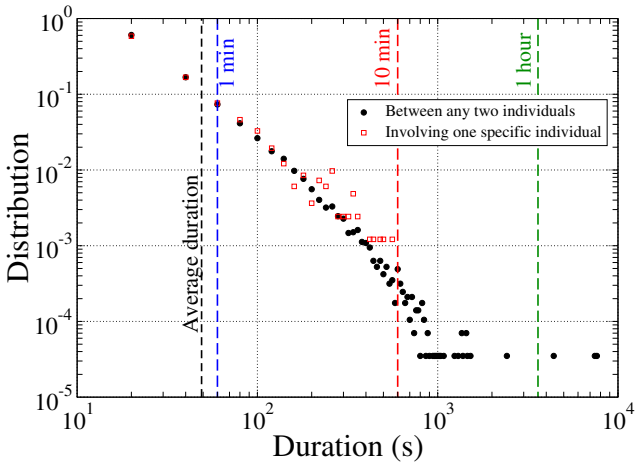


Figure 5.4: Distribution of the contact duration between any two individuals on a log-log scale.

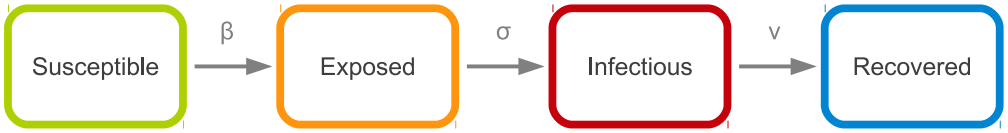


Figure 5.5: Schematic view of the SEIR compartment model. The transmission from an infective to a susceptible is described by a Poisson process of rate β, the latent period and infectious period follow respectively exponential laws of average duration σ^{-1} and ν^{-1}.

size (i.e. same number of edges and of nodes). This difference indicates a clustered network, as generally observed in behavioral interactions [Rocha 2011, Yoneki 2008]. The network also exhibits a small-world feature as the average shortest path is equal to 2.2.

As the deployment lasted two days, two daily aggregated networks are constructed. While the quantities described above are very similar between both days, these two networks differ considerably, as the fraction of repeated contacts in the second day with respect to the first reaches only 12%. The Pearson correlation of degree between the first and the second days is equal to 0.37 which is significantly positive at the 1% threshold, indicating that people interact with a large number of participants the first day are likely to do the same the second day, and conversely people interacting with very few persons are likely to do the same the second day, but not exactly in the same proportion (which would give a Pearson correlation of 1).

The contact model entailed by this non-weighted aggregated network contains individual information about who has met whom, but it neglects all knowledge about contact duration heterogeneity. All pairs of individuals are only considered to be connected or not, as a binary variable. It disregards whether pairs have spent or not most of their time together or met very briefly once, in contrast with the next two contact models.

5.2.2.2 Heterogeneous network model

The second model incorporates the information on contact durations. The same network structure as in HOM is considered, but a weight is associated to each edge. This weight is equal to the cumulated duration of contacts between the corresponding pair of individuals. By this method, we define two daily aggregated and weighted networks which are referred to as the HET model. In this model, a disease can spread from one node to its neighbors, but at a rate that is proportional to the weight. More precisely, a disease spread from an infective node i to a susceptible node j is described by a Poisson process of rate $\beta_{\mathrm{HET}} W_{ij}/\langle W \rangle$, where W_{ij} is the weight between nodes i and j and $\langle W \rangle$ is the average weight (averaged over connected pairs, i.e. when $W_{ij} \neq 0$).

As already observed in other deployments and described in section 2.2.2, weights are broadly distributed with an average duration of interaction of 2 minutes per day and a standard deviation of 7 minutes. The average strength is equal to 75 minutes and corresponds in the epidemiological terminology to the daily exposure duration. The Pearson correlation of strength between the first and the second day is equal to 0.52, which is much higher than the Pearson correlation of degree (0.37), indicating that strength is a more stable quantity than degree between both days.

Both HOM and HET could have been constructed from daily diaries of contacts, in which individuals report with whom they have been in contact during the day. In the HET case, the cumulated duration of contacts has to be given for each pair of individuals, while in the HOM case, as it will be explained below in subsection 5.2.2.5,

only the average duration of contacts has to be estimated.

5.2.2.3 Dynamic network model

The third model of contact patterns consists of the dynamic sequence of contacts directly collected by the SocioPatterns setup and is referred to as DYN. The dynamic network is given at each time step of 20 seconds from the beginning to the end of the deployment, by the list of interacting pairs of badges, which define the instantaneous list of adequate contacts. A disease can spread between two connected nodes at a constant rate β_{DYN}.

The very simple example of a deterministic SI model, presented in paragraph 5.1.5, informed us that the time constraints, that a dynamic network induces, can not be captured by the weighted aggregated network. Comparing simulations on this dynamic contact model to simulations on the weighted static networks can allow us to assess the possible importance of these dynamical constraints.

5.2.2.4 Extension to longer timescales

As the dataset lasts only two days and two days are too short for studying the spreading of any real disease, we replicate the dataset several times to obtain the desired duration (ca 50 times for the slower scenario that will be described bellow).

The simplest procedure, called REP, consists of the simple repetition of the same dataset. In the case of the DYN model, the sequence of contacts is exactly the same, individuals meeting each other at precisely the same time. For the aggregated networks, HOM and HET, the first and second networks model alternately uneven and even days, with the assumption that no contact is possible during *nights*, i.e. from 9 pm to 8:30 am then from 4:30 pm to 9 am. This is represented in figure 5.6. The assumption that contacts are inactive during these periods is made in order to respect the circadian rhythm of the conference dynamics.

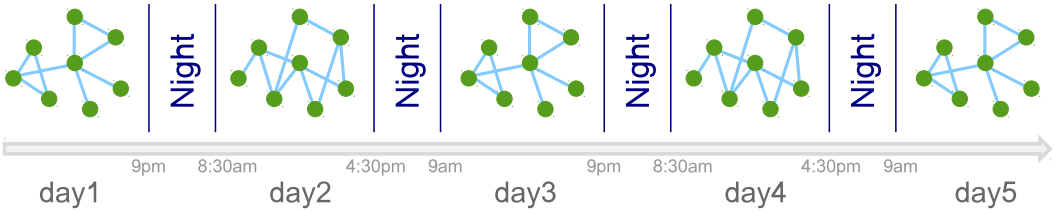


Figure 5.6: Schematic view of the extension to longer time scales for aggregated networks. Contacts are assumed to be inactive during *nights* in order to keep the circadian rhythm of the conference dynamics.

As this repetition procedure is relatively arbitrary and may affect the outcome, two other methods are introduced to check the robustness of the results. The second procedure, called RAND-SH, consists in the repetition of the contact pattern as well but node labels (i.e. tag IDs) are completely randomly reshuffled. In the case

of DYN, the contact sequence remains the same, but the identities are randomly reassigned.

On the one hand, the first method yields a total correlation of contacts between day couples. On the other hand, the second method erases all correlations. From one day to the next, the people met by one individual are not exactly the same, but they are not completely different either. A third repetition procedure, called CONSTR-SH, lies in between. We generate random reshuffling of tag IDs that preserve the fraction of repeated contacts during successive days and the attendees' social activity. A simulated annealing method is used to find a permutation of node labels that respect the fraction of repeated contacts in the second day with respect to the first day f_{emp} (i.e. 12 %). Permutations are generated by a succession of node label reversals. A reversal that creates a new fraction of repeated contacts f is accepted with a probability decreasing with $b\,(f - f_{\text{emp}})^2$, where b is a parameter.

As analyzed theoretically in [Smieszek 2009b], even if clustering should dampen this effect, the highest outbreak size is expected in the RAND-SH method because more direct connections between individuals are introduced by the reshuffling. The more repeated the contacts, the smaller the outbreak size.

5.2.2.5 Parameter equivalence between the different models

A compartmental SEIR model is simulated on top of each contact model. The only difference between the three models relies in the infection process. In order to make comparisons, the following scaling between the infectious parameters β is performed.

The relation between β_{HOM} and β_{HET} is given by the following constraint: the rate of infection averaged over all edges is the same in both models. It can be translated in the mathematical constraint:

$$\sum_{(i-j)\in E} \beta_{\text{HOM}} = \sum_{(i-j)\in E} W_{ij}\beta_{\text{HET}}/\langle W\rangle \tag{5.3}$$

where E is the set of edges. It gives the following relation: $\beta_{\text{HOM}} = \beta_{\text{HET}}$.

The relation between the DYN and the HET models is given by the second constraint: the probability of a disease transmission over one day from an infective node i to a susceptible node j, given that node i has not recovered by the end of the day, should be the same. In mathematical terms, it corresponds to the following relation:

$$\begin{aligned} \forall (i-j)\in E \qquad & 1 - \exp\left(-\beta_{\text{DYN}} W_{ij}\right) = 1 - \exp\left(-\beta_{\text{HET}} \frac{W_{ij}}{\langle W\rangle}\Delta T\right) \\ \text{i.e.} \qquad & \beta_{\text{HET}} \frac{W_{ij}}{\langle W\rangle} = \beta_{\text{DYN}} \frac{W_{ij}}{\Delta T} \end{aligned} \tag{5.4}$$

These two constraints give the needed relations between the rates to make comparisons and assess the effect of the various contact model assumptions. They are summarized in table 5.2, taking $\beta_{\text{DYN}} = \beta$ as a reference.

Two scenarios are considered.

Network model	Homogeneous	Heterogeneous	Dynamic
S + I $\rightarrow$ E + I	$\beta\langle W\rangle/\Delta T$	$\beta W_{ij}/\Delta T$	β
E $\rightarrow$ I	σ	σ	σ
I $\rightarrow$ R	ν	ν	ν

Table 5.2: Rate of the transmission, latency and recovery Poisson processes for each contact pattern model. W_{ij} corresponds to the cumulative duration of the contacts between nodes i and j, $\langle W\rangle$ is the average cumulative duration over all edges (i, j) and ΔT is the total duration during which the links of the static networks are considered as active.

- a rapid scenario: $\sigma^{-1} = 1$ day, $\nu^{-1} = 2$ days, $\beta = 3.10^{-4}s^{-1}$,
- a slower scenario: $\sigma^{-1} = 2$ day, $\nu^{-1} = 4$ days, $\beta = 15.10^{-5}s^{-1}$.

These parameter values are chosen in order to keep the ratio β/ν constant between the two situations, which means that the biological factors responsible for the rate of increase of cases are the same in both cases, but dynamics is different. These two scenarios are still rather too fast to be realistic, even though the slower scenario could correspond to a very virulent influenza outbreak, but they were chosen in order to limit the effect of the time extension procedure.

Numerical simulations are performed in the following manner. First, the permutations used to extend the duration of the contact models are computed once for all for the RAND-SH and the CONSTR-SH methods. This gives a unique version of each of the extended contact models. Then 5000 simulations runs are performed, each of them with a uniformly selected individual among the total population to be the first infectious individual, also called the seed. The dynamics is computed under a discrete time approximation: at each time step of length δt, a Poisson event of rate α (possibly equal to β, ν or σ) occurs with probability $\alpha\delta t$. This approximation holds if δt is small enough. I took $\delta t = 20$ s for the dynamic contact network and $\delta t = 1800$ s for the aggregated networks. In spite of this difference, as the number of neighbors is much higher in the aggregated networks than in the dynamic network, the computing time is longer in the former case. A more efficient way would have been to sample the recovering period from an exponential distribution each time a new infection occurred, and to sample the time of infection for each neighbor of the newly infected individual. If the time of infection precedes the recovering time, then the infection should occur. This way, initial draws are made only once for each infectious individual and not at each time step. Results on the basic reproductive number, the prevalence and the temporal evolution are analyzed on the 5000 trajectories obtained with 5000 random selection of seeds and one realization for each.

5.2.3 Results

5.2.3.1 Basic reproductive number

Several methods can be used to compute the basic reproductive number R_0 [Diekmann 1990, Heffernan 2005], possibly yielding different results [Breban 2007]. We estimate its value as the mean over different realizations of the number of secondary cases infected by the single randomly chosen initial infectious individual.

Figure 5.7 reports the distributions of R_0 for the three network models, for the REP time extension procedure. In all cases, the number of secondary cases from the initial seed of the single infectious individual ranges from 0, corresponding to the most probable event of no outbreak, to around 20-25 individuals.

Figure 5.8 gives the boxplot representation of the estimated distribution of R_0 depending on the scenarios (slow and fast), the network models (HOM, HET and DYN) and the time extension procedure (REP, RAND-CONSTR and RAND-SH). In all scenarios and all time extension procedures, higher values of R_0, together with larger variances, are observed in the HOM network compared to the HET and DYN network models, which both give very similar distributions.

5.2.3.2 Final size of the epidemic

Figure 5.9 shows the distribution of the final number of cases for the three network models and the REP data extension procedure. A high probability of rapid extinction of the pathogen spread is observed, corresponding to a small number of individuals who become infected. This is slightly smaller in the HOM case compared to the HET and DYN networks. On the contrary, when the epidemic starts, the final number of cases is high, and it is larger in the HOM case with respect to the HET and DYN networks. Intermediate cases with limited propagation are rare.

Table 5.3 summarizes the distribution of the final number of cases for the three networks for the various parameters of the SEIR model and in the various data extension scenarios. For all cases, and independently from the procedure adopted for extending the two-days data set, the probability of extinction is lower for the HOM cases with respect to the HET and DYN networks. In case of large outbreaks, the final size is higher in the HOM network compared to the HET and DYN networks. Propagation over HET and DYN networks leads to similar extinction probability and final number of cases. The final number of cases for both disease scenarios (i.e. slow and fast spreading parameter sets) are also fairly close.

This result implies that heterogeneity in the contact durations between individuals is associated with a lower spread of transmission, suggesting that the unequal sharing of time spent by an individual with its contact partners effectively reduces the routes of disease spread. Disregarding the heterogeneity of contact durations can lead to large differences in the estimated number of cases, suggesting that information on the daily cumulated contact time between individuals gives crucial information for correct modeling of disease spread.

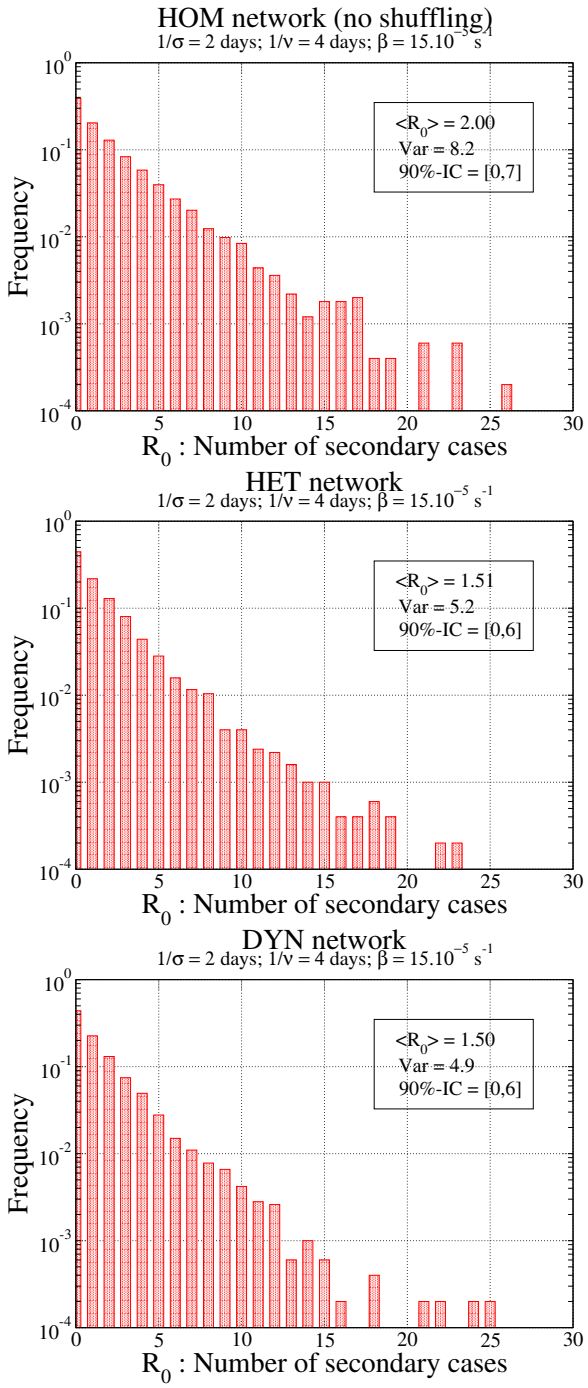


Figure 5.7: Distribution of the basic reproductive number R_0 for the homogeneous (HOM), heterogeneous (HET) and dynamic (DYN) contact pattern models in the repetition (REP) procedure.

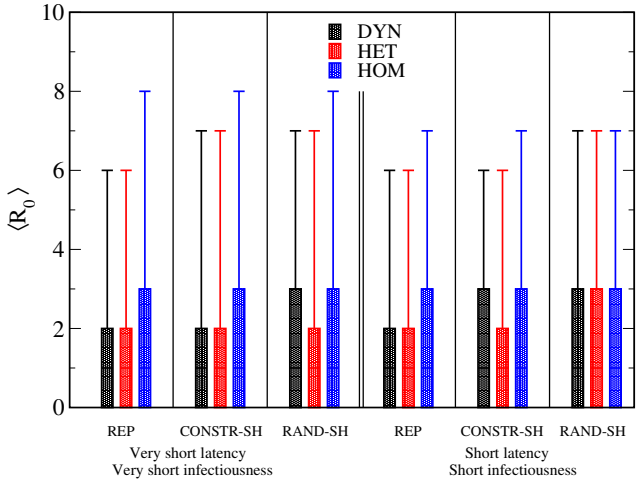


Figure 5.8: Boxplot of the distribution of R_0 according to the different methods to extend to longer timescales (REP, CONSTR-SH and RAND-SH), the different network models (DYN, HET and HOM) and for a slow and a rapid scenario. The bottom, middle and top lines of the rectangular box correspond respectively to the 25th, 50th and 75th percentiles of the distribution, the bottom and top whiskers give the 5th and 95th percentiles.

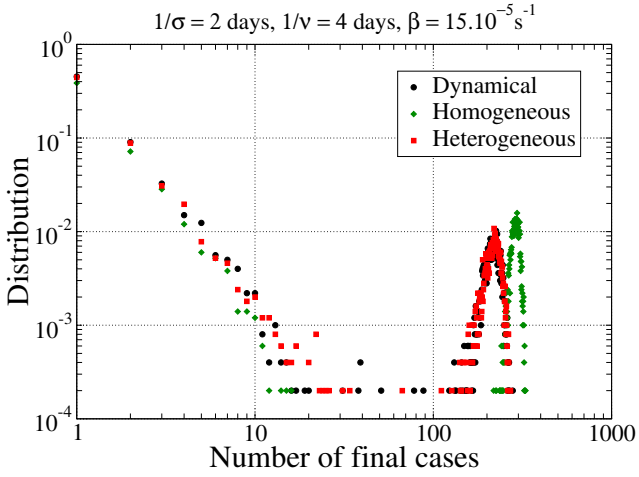


Figure 5.9: Distribution of the final number of cases for the three network models (DYN, HOM and HET) with the parameter of the slower scenario ($\sigma^{-1} = 2$ days, $\nu^{-1} = 4$ days and $\beta = 15.10^{-3}$ sec^{-1}) and the REP procedure.

				No secondary case	1 to 10 final cases (AR $\leq$ 2.5%)			11 to 40 final cases (2.5% $\leq$ AR $\leq$ 10%)			More than 40 final cases (AR $>$ 10%)		
Scenario	Parameters	Network	Runs	% run	% run	Mean cases	90% CI	% run	Mean cases	90% CI	% run	Mean cases	90% CI
REP													
Very short latency	$\sigma^{-1} = 1$ day	DYN	5000	47.3	18.2	2.3	[1,6]	0.7	15.9	[11,22]	33.8	208	[169,242]
Very short infectiousness	$\nu^{-1} = 2$ day	HET	5000	46.4	17.7	2.4	[1,7]	0.8	17.9	[11,32]	35.2	210	[171,243]
Transmission rate	$\beta = 3.10^{-4}s^{-1}$	HOM	5000	41.7	11.7	2.2	[1,6]	0.2	16.6	[11,30]	46.3	285	[257,310]
Short latency	$\sigma^{-1} = 2$ day	DYN	5000	45.3	17.0	2.2	[1,7]	0.4	18.3	[11,38]	37.3	214	[178,246]
Short infectiousness	$\nu^{-1} = 4$ day	HET	5000	44.4	16.4	2.2	[1,6]	0.6	16.8	[11,27]	38.6	216	[178,248]
Transmission rate	$\beta = 15.10^{-5}s^{-1}$	HOM	5000	38.7	13.2	2.1	[1,6]	0.1	13.2	[11,15]	48.1	288	[262,310]
RAND-SH													
Very short latency	$\sigma^{-1} = 1$ day	DYN	5000	44.8	19.4	2.8	[1,8]	2.2	17.9	[11,31]	33.6	278	[223,319]
Very short infectiousness	$\nu^{-1} = 2$ day	HET	5000	45.4	18.5	2.6	[1,7]	1.6	17.6	[11,30]	34.5	284	[241,322]
Transmission rate	$\beta = 3.10^{-4}s^{-1}$	HOM	5000	39.9	14.3	2.6	[1,7]	0.8	15.7	[11,28]	45.0	324	[291,350]
Short latency	$\sigma^{-1} = 2$ day	DYN	5000	40.6	18.6	2.7	[1,8]	1.4	19.2	[11,31]	39.4	297	[254,331]
Short infectiousness	$\nu^{-1} = 4$ day	HET	5000	39.5	18.0	2.7	[1,8]	1.3	16.7	[11,30]	41.2	300	[259,333]
Transmission rate	$\beta = 15.10^{-5}s^{-1}$	HOM	5000	35.9	15.7	2.5	[1,7]	0.9	17.0	[11,31]	47.5	325	[293,352]
CONSTR-SH													
Very short latency	$\sigma^{-1} = 1$ day	DYN	5000	45.4	17.7	2.4	[1,7]	1.0	17.0	[11,28]	35.8	240	[194,278]
Very short infectiousness	$\nu^{-1} = 2$ day	HET	5000	46.8	16.5	2.4	[1,7]	0.8	19.0	[11,33]	35.9	245	[202,282]
Transmission rate	$\beta = 3.10^{-4}s^{-1}$	HOM	5000	39.8	13.3	2.3	[1,6]	0.7	15.4	[11,21]	46.2	308	[278,334]
Short latency	$\sigma^{-1} = 2$ day	DYN	5000	40.9	18.2	2.3	[1,6]	0.8	16.8	[11,34]	40.2	258	[215,292]
Short infectiousness	$\nu^{-1} = 4$ day	HET	5000	41.3	16.8	2.3	[1,7]	0.5	14.0	[11,25]	41.4	257	[213,292]
Transmission rate	$\beta = 15.10^{-5}s^{-1}$	HOM	5000	35.7	14.8	2.4	[1,7]	0.4	15.2	[11,21]	49.2	314	[284,339]

Table 5.3: Final number of cases for the three network models (DYN, HET and HOM) according to the different methods to extend to longer timescales (REP, CONSTR-SH and RAND-SH) and for a slow and a rapid scenario. The attack rate (AR) gives the proportion of final cases in the population.

5.2.3.3 Temporal evolution

Regarding the peak times of disease spread in the various cases (see figure 5.10), we found that in most cases, the peak of the epidemic was reached first on average within the HOM network model. However, the differences between the peak times were small, and even the simulations on the network model with the least information (i.e. the HOM model) gave a good estimate of the peak time obtained when the full information on the contact patterns was included (i.e. the DYN network model). Interestingly, while the HOM network model yielded a much higher prevalence, the peak time is, however, only slightly changed, showing that even rather limited information can yield good estimates of the epidemic timescales.

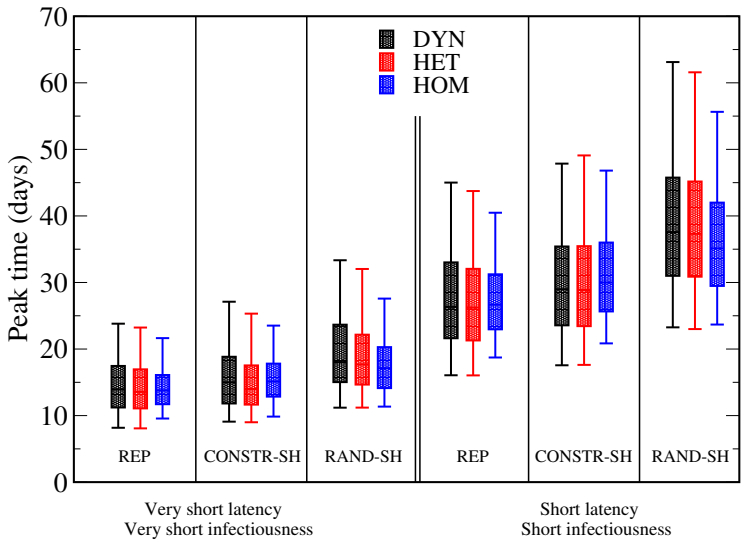


Figure 5.10: Boxplot of the distribution of the prevalence peak time t_{peak} according to the different methods to extend to longer timescales (REP, CONSTR-SH and RAND-SH), the different network models (DYN, HET and HOM) and for a slow and a rapid scenario. Boxplot conventions are the same as in figure 5.8. Only runs with an attack rate over 10% are taken into account.

Using the evolution in time of the number of infectious and recovered individuals for the different data-extension procedures and for the two sets of SEIR parameters, the temporal behavior of disease spread was analyzed (see figures 5.11 and 5.12). Symbols represent the median values, and lines represent the fifth and ninety-fifth percentiles of the number of infectious and recovered individuals. In all cases, disease spread on the HOM network evolved slightly faster and reached a significantly larger number of individuals, compared with the HET and DYN, which had very similar characteristics to each other. The comparison between disease spread in the HET and DYN networks provides insights into whether temporal constraints due to the precise sequence of the contacts might affect the propagation of disease. The time constraints on the paths that the infectious agent can follow between individuals may

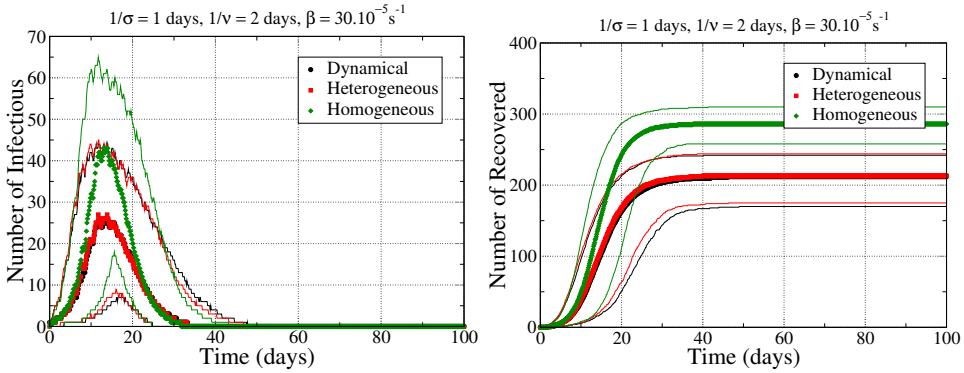
slow down disease spread on the DYN network compared with the HET network. However, this slowing down of infection and the differences in the final number of cases between the HET and DYN networks are too small to be relevant for the simulations investigated here. The similarity between the spreading behaviors in the HET and DYN networks are independent of the different procedures used to extend the initial 2-day dataset. The robustness of the comparison between HET and DYN therefore indicates that the observed similarity between the spreading on the HET and DYN networks is due to the discrepancy between the timescales considered for propagation (of the order of days), and the temporal resolution and the contact durations (of 20 seconds and of the order of minutes up to a few hours, respectively). The total time spent in contact by each pair of individuals is in this context sufficient to describe precisely the propagation pattern, as shown by the peak time and the final number of cases. Therefore, for the simulation of diseases such as those considered in this study, contact information at a daily resolution might be enough to characterize disease spreading, and the precise order of the sequence of contacts might not be needed. A similar result was found with a different setting in a primary school [Potter 2012]. However, this would not be the case for extremely fast-spreading processes, as shown in [Isella 2010]. This implies that there is a crossover between the two regimes, which could be the subject of future investigations.

Interesting differences were seen in the results of simulations on datasets extended with different procedures (see figures 5.10, 5.11 and 5.12). The spread was slightly slower in the RAND-SH case, but lasted longer, and consequently the final number of cases R_∞ was larger. In fact, we systematically found $R_\infty(\text{REP}) < R_\infty(\text{CONSTR-SH}) < R_\infty(\text{RAND-SH})$, and the more the identities of the tags were shuffled, the more efficient was the spread. Repeated encounters favor propagation, so that the REP procedure led to an initially faster spread, but contacts between different individuals from one day to the next favor propagation across the network, so that the RAND-SH procedure led in the end to a larger attack rate. This lies in agreement with other studies [Smieszek 2009b, Smieszek 2009a, Read 2008] showing the importance of knowledge of the respective fractions of repeated and new contacts between successive days. In [Smieszek 2009b] it is shown that, keeping constant the daily number of encounters, the final size of the outbreak is higher in the case of a fully random repetition of contacts than in the case where the same contacts are repeated over several days. It is worth noting that the fully random repetition of contacts does not exactly correspond to the uniform randomization of node labels used in the RAND-SH time extension procedure, because we keep the contact network constant while no network structure is assumed in [Smieszek 2009b].

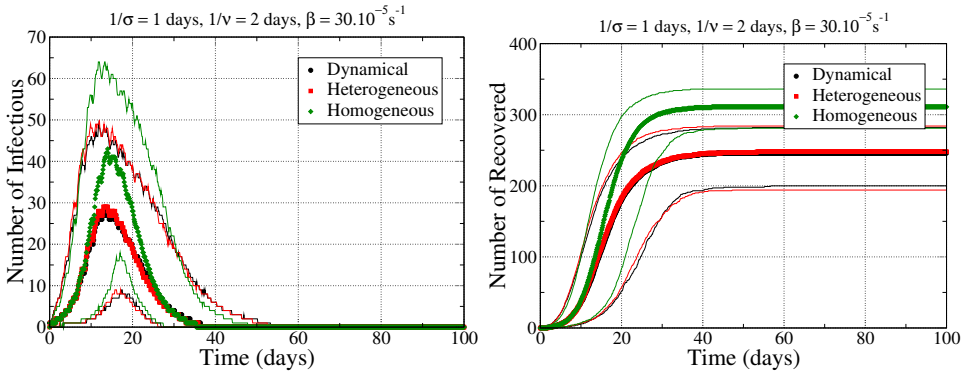
5.2.4 Limitations

This study based on numerical simulations on data collected with RFID sensors carry some caveats that need to be discussed. First, as already pointed out in the protocol description in section 1.2.3, individuals are not followed outside of the zone covered by RFID readers, so that contacts between participants that occur during

(a) REP



(b) CONSTR-SH



(c) RAND-SH

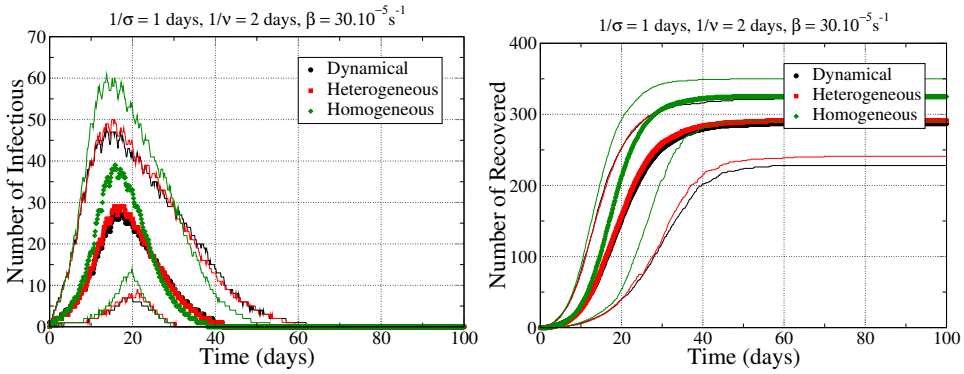
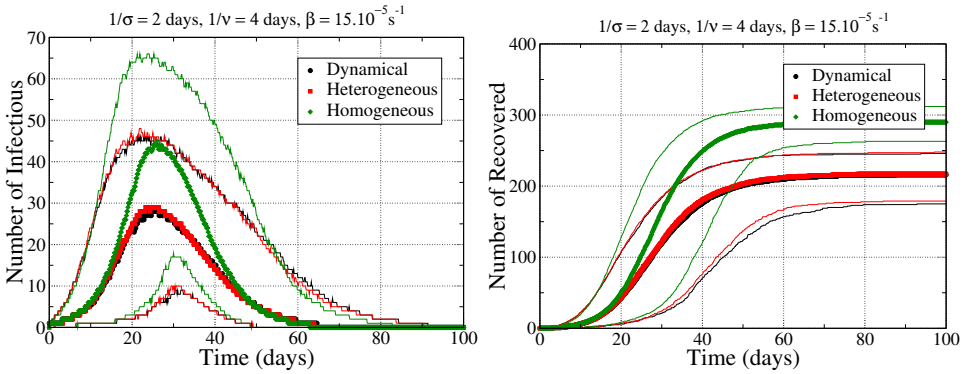
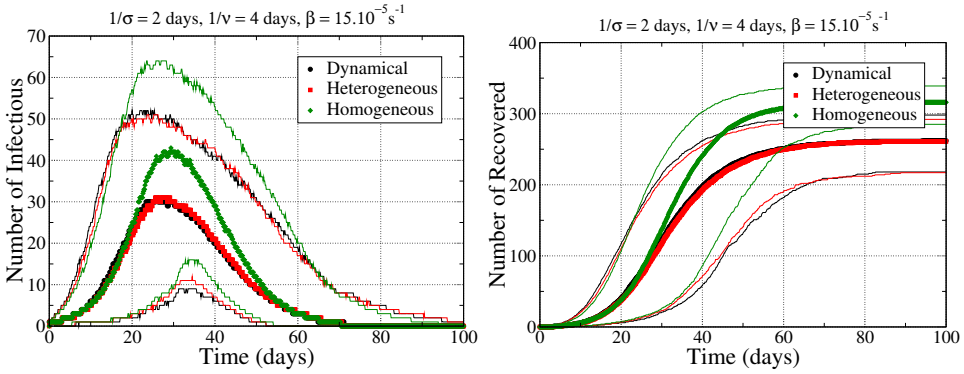


Figure 5.11: Temporal evolution of the spreading process for the three network models and for the three time extension procedures in the fast scenario. Left is the prevalence (i.e. the number of infectious individuals) and right is the number of recovered individuals. Only runs with an attack rate over 10% are taken into account. Symbols represent the median values and lines represent the fifth and ninety-fifth percentiles of the number of infectious and recovered individuals.

(a) REP



(b) CONSTR-SH



(c) RAND-SH

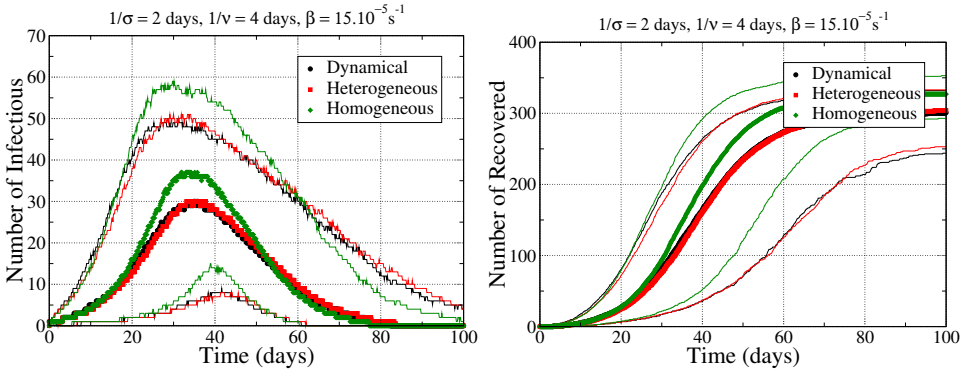


Figure 5.12: Temporal evolution of the spreading process for the three network models and for the three time extension procedures in the slow scenario. Left is the prevalence (i.e. the number of infectious individuals) and right is the number of recovered individuals. Only runs with an attack rate over 10% are taken into account. Notations are the same as in the previous figure.

the day outside of the area covered by the RFID readers are not monitored. This results in an underestimation of the number of contacts, and therefore of the possibilities for disease spread. Moreover, in this study, the periods of *nights* represented a proportion of 56% of the 24-hour period, during which individuals were assumed to be isolated. This may artificially increase the probability of extinction if the infectious period of an infected individual ends during these periods, precluding further transmission. This issue may be solved by upcoming technological improvements that will allow operation of the RFID sensing layer in a fully distributed fashion with on-board storage on the devices themselves; that is, such RFID tags will register and store contacts even if they are not close to RFID readers.

Another issue, well known in the field of social networks, is due to the partial sampling of the population [Handcock 2010]. Of the 1,200 attendees at this conference, 405 (34%) participated in the data collection. Consequently, only these attendees were taken into account in the model of disease spread, whereas they were in fact also in contact with the non-participating attendees. The analysis presented in chapter 2 showed that for a wide variety of real-world deployments of the RFID proximity-sensing platform used in this study, the behavior of the statistical distributions of quantities such as contact durations is not altered by unbiased sampling of individuals. However, paths of disease spread between sampled attendees that also involved unsampled attendees may have existed, but were not taken into account. This effect may lead to an underestimation of disease spread. In addition, it is possible that the volunteering participants themselves introduced a systematic bias into the sampled population concerning their interaction behavior, as they self-selected to participate to the experiment.

Finally, the limited period (2 days) of data collection made it necessary to generate artificially longer datasets by different procedures in order to model the spread of pathogens on realistic timescales. Deployment of the measuring infrastructure on much longer timescales would be needed to validate such generation procedures and to measure their effect.

5.3 Partial conclusion and perspectives

We used data collected in a 2-day conference involving 405 volunteers to compare the simulated spread of communicable diseases on the dynamic network of contacts (DYN) and on two other networks, one heterogeneous (HET) and one homogeneous (HOM), obtained by aggregating the dynamic network at two distinct levels of precision. To compensate for the relatively short duration of the observation period (2 days), we designed three different models to construct dynamical contact networks spanning an extended time period during which the spread of an infectious disease could be simulated. In the three networks, disease extinction occurred as frequently as large outbreaks and these latter tended to be explosive.

Despite the limitations described above, this study emphasizes the effects of contact duration heterogeneity on the dynamics of communicable diseases. On the one

hand, the small differences between simulated spread on both the HET and DYN networks shows that taking into account the very detailed actual time ordering of the contacts between individuals, with a time resolution of minutes, does not seem to be essential to describe disease spread on a timescale of several days or weeks. On the other hand, the large differences in disease spread in the HOM network emphasize the need to include detailed information about the heterogeneity of contact duration (compared with an assumption of homogeneity) to model disease spread, as also found previously [Read 2008, Smieszek 2009a] for simulations of disease spread dynamics based on diary-based survey data. Results from the different procedures for data extension also showed how the rate of new contacts is a very important parameter [Smieszek 2009b, Read 2008]. Overall, the combined comparison of the spreading processes simulated on the HET, DYN and HOM networks and using the different data-extension procedures gave an important assessment of the level of detail concerning the contact patterns of individuals that is needed to inform modeling frameworks of epidemic spread.

In this context, a data collection infrastructure such as the one developed in the SocioPatterns collaboration seems to be very effective, as it gives access to the level of information needed, and also allows the simulation of very fast-spreading processes characterized by timescales comparable with those intrinsic to social dynamics, where even the precise ordering of contact events becomes crucial. These measurements should be also extended to other contexts in which individuals interact closely in different ways, such as workplaces, schools or hospitals [Polgreen 2010, Isella 2011]. More experimental work is needed to collect data over longer time periods, on various populations and in various locations. This would help to better understand how to artificially extend datasets limited in time in order to yield realistic datasets. The results of these approaches could be helpful to anticipate the effect of preventive measures, and contribute to decisions about the best strategies to control the spread of known or emerging infections [Polgreen 2010, Masuda 2009, Kitsak 2010].

Chapter 6

Conclusion

This thesis hinges on the study of behavioral dynamic networks collected by means of wearable sensors. These sensors, based on radio communication, allow one to measure face-to-face proximity between individuals in a closed environment with an unprecedented time resolution of 20 seconds.

Compared to classic pen-and-paper approaches, this methodology allows researchers to collect detailed information while avoiding difficulties due to cognitive limits. Indeed, the systematic record of encounters over several days, especially if they last less than 1 minute, is not easily obtained by classical methods such as personal interviewing and self-reported diaries. The use of a technological infrastructure overcomes this limitation, at the cost of the reduction of interactions to one dimension, i.e. face-to-face proximity. To be more specific, only highly resolved behavioral data is obtained, whilst oral or written information provided by participants, that could help us to contextualize social interactions, are not available. This has consequences for the interpretation of results, because we can not allege any intentionality in the interaction patterns. In the same manner network survey data are likely to suffer from technical problems (interviewer effect, misinterpretation, lost and incomplete forms, interruption of interviews, answer inconsistencies, non-response bias), these sensor based network datasets have their share of problems: defective badges or antennas, lost badges, unloaded batteries, badge swaps between participants, mishandling of badges (gathering of switched-on badges in a box at the beginning or end of the day). While some of these problems can be automatically treated (for example, the mishandling of badges is easily removed with an upper filter on the instantaneous degree), others are, likely only partially, solved by a careful postprocessing of the data (for instance, the reallocation of the badge identification numbers).

The relatively new nature of such datasets led me to think on the appropriateness of analytical tools and measures. The formalism to process highly resolved temporal networks is not an old established basis on which I could have relied on to build a systematic descriptive analysis of proximity. I had during my thesis to think always twice (or more) on the quantities to compute, and I found a balance between an analysis in terms of dynamical networks and the construction of weighted static networks based on the aggregation of contact durations over the entire span of the deployments.

I analyzed various datasets in diverse contexts, namely in scientific conferences, in a primary school and in a museum, showing differences and similarities by specifically designed quantities and distributions. Typical network measures, such as the degree and strength distributions, the analysis in terms of groups, of distances and the number of connected components, quantify the interaction patterns of each event, and are generally well explained by the knowledge of the contexts in which these interactions take place. On the other hand, measures on the dynamic of interactions, such as the distribution of the contact durations, of the cumulated contact durations, of the time elapsed between two contacts, are revealed to be robust across the studied contexts. These similarities have awaken our interest for generic mechanisms that could produce such patterns.

The model developed in this thesis takes place within this framework. Treated both analytically and numerically, it successfully produces the same generic interaction patterns based on simple micro-mechanisms, while remaining flexible enough to be implemented with additional features. More precisely, broad distributions in contact durations and intercontact times take their origin here in self-reinforcement rules that increase the propensity of agents to remain in the same state (same coordination number) with the time already spent in this state. The versatility of the model is exemplified first with its ability to produce different kinds of contact and intercontact distributions, such as simple, compressed and stretched exponentials. Second, heterogeneity among individuals' behaviors can be implemented in the micro-rules, and in the case we studied, it does not change the overall phenomenology of the model. Third, a fluctuating population size can be introduced to model more realistic environments in which individuals enter and leave through time. An interesting perspective for this model would be to use for the simulation of dynamical processes such as information or infectious spreading or synchronizations, to investigate the effect of its tunable parameters on the temporal evolution of the studied dynamical process.

I presented two main lines for the use of such datasets. First, the epidemiological community working on infectious diseases is directly interested in highly detailed data on contacts. On a ground level, the simple quantification of contact rates between the various population stratas is needed to inform epidemiological models. These models may then be used to assess the efficiency of different public policies, on the basis of numerical simulations. On a more theoretical level, it is desirable to know which details on contact dynamics are needed for these models. In terms of cost/benefit ratio, how much does 20 second resolved data stream on interactions improve the quality of predictions compared to other data that would be obtained at a lower cost, for example with mobility traces obtained with mobile phone data? The analysis I presented on different contact pattern models, including different levels of information, is only a small step in this direction, which could be further explored. More precisely, I compared the simulated spread of communicable diseases on the dynamic network of contacts and on two other networks, one heterogeneous and one homogeneous, obtained by aggregating the

dynamic network at two distinct levels of precision. This comparison emphasized the effects of contact duration heterogeneity on the dynamics of communicable diseases. On the one hand, the small differences between simulated spread on both the heterogeneous and dynamic networks shows that taking into account the very detailed actual time ordering of the contacts between individuals, with a time resolution of minutes, does not seem to be essential to describe disease spread on a timescale of several days or weeks. On the other hand, the large differences in disease spread in the homogeneous network emphasize the need to include detailed information about the heterogeneity of contact duration (compared with an assumption of homogeneity) to model disease spread. Additional measurement campaigns in different settings and covering different schools, countries, and age groups, on longer timescales, are much needed to, first, validate our preliminary results on the need of information on the heterogeneity of contact durations but not necessarily on the contact sequence order and, second, to obtain datasets which could be compared with other datasets obtain by different means.

The second direction line is oriented toward social studies. This methodology allows to access to very detailed information on behavior ties between persons. While these ties do not inform us on intrinsic preferences, which would be known by directly asking people on their social relationships, they provide us a detailed picture on what they actually do. Interestingly, the development of unsupervised methodologies, such as the one I worked on, allows social scientists to collect large-scale, high-resolution dynamical data on human behavior in a reproducible manner. Because of this valuable asset, such methodologies are likely to further develop.

In this thesis, I presented a study of gender homophily among children in primary school. Based on behavioral ties (*who people interact with*), we recover results that are generally obtained with social ties (*who people are friend with*), namely that, (1) gender homophily is statistically present in all grades, (2) same-gender preference reaches a higher level for boys than for girls in middle grades, (3) same-gender ties are more stable than mixed-gender ties, and (4) same-gender preference tends to increase with age for strong ties, at a higher rate for boys than for girls. I also brought an additional stone in the literature of gender homophily, with the analysis of the difference between boys and girls in their homophilous behavior when we consider their weak ties, i.e., the mates they spend little time with.

I also presented a study bridging the gap between social relationships, or more precisely virtual social ties, and face-to-face interaction. The coupling between the sensor-based infrastructure with the Live Social Semantics platform during a scientific conference highlights that (1) face-to-face contacts between attendees are more frequent and last longer if they share a virtual tie, such as a Facebook friendship, than if they do not, (2) these virtual friends also share a higher behavioral similarity (they interact more similarly with the rest of the population) and a higher trajectory similarity (they move from place to place in a similar manner), and (3) face-to-face interactions provide relevant information on the existence of virtual friendships, in terms of the link prediction. While these results shed light on the interrelation

between face-to-face proximity and virtual ties, it remains an open problem for classical social ties. While Mark Granovetter defined the amount of time spent together as an ingredient of tie strength [Granovetter 1973], Marsden and Campbell objected that interaction frequencies would poorly measure social affiliation (an unobservable concept), reflecting rather the effect of contingencies than being the outcome of a personal preference or choice [Marsden 1984]. Their analysis would benefit from being conducted again with data on contact frequencies and durations that are correctly recorded without the problem of the respondent limited recall. Furthermore, it would be interesting to know whether friendship closeness is a more relevant variable than contact duration depending on the sociological question that is addressed.

As outlined in the thesis, this methodology would provide an interesting alternative to repeated panel network surveys. The study of the coevolution of network and behavior is a hot topic in social sciences [Steglich 2010]. Longitudinal data are necessary for measuring peer effects, which are known to play a crucial role in education, in alcohol and drug consumption, in sexual risk behavior and more widely for any social behavior susceptible to be influenced by others' behaviors. Among other ingredients, peer influence and peer selection are the two principal mechanisms shaping the correlated behaviors of peers. The relative importance of these mechanisms can influence the efficiency of various public policies, when an organizational authority considers the concerned behavior as noxious or beneficial for the society.

A last opened problem concerns the effect of missing data and incomplete sampling on the properties of dynamical processes unfolding on networks. Whilst it is generally difficult to define the network boundaries for a specific question, it is even more difficult to obtain data on all nodes and ties. Generally, we only have at our disposal a node and tie sample. Usual statistical models such as ERGM, that rely on a complete network data assumption, are then used to analyze these sampled datasets. The bias and/or inaccuracy it generates is not sufficiently assessed, even though some recent works have been done to tackle the problem. The same difficulty exists for dynamical processes such as disease spreading. The imprecision done when ignoring unobserved ties that would play a role in real processes, is not well controlled. I hope that these simple and interesting questions will awake some interest in the statistician community.

To conclude, this thesis lies in a highly interdisciplinary context, at the frontiers between statistical physics, epidemiology and sociology. This fruitful confluence of approaches has stimulated my interest, which I hope to have shown in the diversity of my contributions. At present I am bifurcating toward the disjoint field of the monetary redistribution but I hope that the time I am allowed to dedicate to research activity will allow me to pursue in the field of network dynamics. This field is still in a growing phase and much remains to be done.

Appendix A

List of publications

Articles published in peer reviewed journals:

- J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina and P. Vanhems. *High-resolution measurements of face-to-face contact patterns in a primary school*, PLoS ONE, **6**(8):e23176, 2011.

- J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Régis, J.-F. Pinton, N. Khanafer, W. Van den Broeck and P. Vanhems. *Simulation of an SEIR infection disease model on the dynamic contact network of conference attendees*, BMC Medicine, **9**:87, 2011.

- K. Zhao, J. Stehlé, G. Bianconi and A. Barrat. *Social network dynamics of face-to-face interactions*, Physical Review E, **83**,056109 & arXiv:1102.2423, 2011.

- L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton and W. Van den Broeck. *What's in a crowd? Analysis of face-to-face behavioral networks*, Journal of Theoretical Biology, **271**, 166-180 & arXiv:1006.1260, 2011.

- J. Stehlé, A. Barrat and G. Bianconi. *Dynamical and bursty interactions in social networks*, Physical Review E, **81**, 035101(R) & arXiv:1002.4109, 2010.

Manuscripts submitted to peer reviewed journals:

- J. Stehlé, F. Charbonnier, T. Picard, C. Cattuto and A. Barrat. *Gender homophily in a primary school : a sociometric study using high-resolution networks of face-to-face proximity*, Submitted to Social Networks.

- A. Barrat, C. Cattuto, V. Colizza, F. Gesualdo, L. Isella, E. Pandolfi, J.-F. Pinton, L. Ravà, M. Romano, C. Rizzo, J. Stehlé, A. E. Tozzi and W. Van den Broeck. *Empirical temporal networks of face-to-face human interactions*, Submitted to European Physical Journal B.

Glossary

Notation	Description	
P_k	Degree distribution	18
R_0	Basic reproductive number	103
$Y_2(i)$	Participation ratio of node i	19
c_i	Clustering coefficient of node i	18
k_i	Degree of node i	17
s_i	Strength of node i	19
$sim_{i,j}$	Cosine similarity between nodes i and j	18
w_{ij}	Edge weight between nodes i and j	18
25C3	25^{th} Chaos communication congress, Berlin, Germany (2008)	14
ACC	Aggregated Conference Contact Network	63
AR	Attack rate	104
AUC	Area under the curve	70
CC	Connected components	20
CV	Coefficient of variation	33
DYN	Dynamic network model	117
ESWC09	6^{th} European semantic web conference, Heraklion, Greece (2009)	14
ESWC10	7^{th} Extended semantic web conference, Heraklion, Greece (2010)	14
FB	Facebook	63
FPR	False positive rate	70
HET	Heterogeneous network model	117
HOM	Homogeneous network model	115
HT2009	20^{th} ACM Conference on hypertext and hypermedia, Torino, Italy (2009)	14
LSS	Live Social Semantic	62

Notation	Description	
ROC	Receiver operating characteristic curve	69
SFHH	Conference of the French society for hospital hygiene (*Société française d'hygiène hospitalière*), Nice, France (2009)	14
SG	Science Gallery deployment, Dublin, Ireland (2009)	20
SIR	Susceptible Infected Recovered	106
WAIFW	The *who acquires infection from whom* matrix	107

Bibliography

[Abbey 1952] Helen Abbey. *An Examination of the Reed-Frost Theory of Epidemics.* Human Biology, vol. 24, no. 3, pages 201–233, Sept 1952. (Cited on pages 99 and 103.)

[Adamic 2003] Lada Adamic and Eytan Adar. *Friends and neighbors on the Web.* Social Networks, vol. 25, no. 3, pages 211 – 230, 2003. (Cited on page 47.)

[Aiello 2010] Luca Maria Aiello, Alain Barrat, Ciro Cattuto, Giancarlo Ruffo and Rossano Schifanella. *Link Creation and Profile Alignment in the aNobii Social Network.* In Proceedings of the 2010 IEEE Second International Conference on Social Computing, SOCIALCOM '10, pages 249–256, Washington, DC, USA, 2010. IEEE Computer Society. (Cited on page 48.)

[Ajelli 2010] Marco Ajelli, Bruno Goncalves, Duygu Balcan, Vittoria Colizza, Hao Hu, Jose Ramasco, Stefano Merler and Alessandro Vespignani. *Comparing large-scale computational approaches to epidemic modeling: Agent-based versus structured metapopulation models.* BMC Infectious Diseases, vol. 10, no. 1, page 190, 2010. (Cited on page 108.)

[Alani 2009] Harith Alani, Martin Szomszor, Ciro Cattuto, Wouter Van den Broeck, Gianluca Correndo and Alain Barrat. *Live Social Semantics.* In 8th International Semantic Web Conference (ISWC), October 2009. (Cited on page 62.)

[Albert 2000] Réka Albert, Hawoong Jeong and Albert-László Barabási. *Error and attack tolerance of complex networks.* Nature, vol. 406, pages 378–382, July 2000. (Cited on page 5.)

[Anderson 1991] Roy M. Anderson and Robert M. May. Infectious diseases of humans : dynamics and control. Oxford : Oxford University Press, 1991. (Cited on pages 45, 98, 102, 103, 104, 105 and 108.)

[Andersson 2000] Håkan Andersson and Tom Britton. Stochastic epidemic models and their statistical analysis. Lecture Notes in Statistics. Springer, 2000. (Cited on pages 100, 103 and 105.)

[Bajardi 2011] Paolo Bajardi, Chiara Poletto, Jose J. Ramasco, Michele Tizzoni, Vittoria Colizza and Alessandro Vespignani. *Human Mobility Networks, Travel Restrictions, and the Global Spread of 2009 H1N1 Pandemic.* PLoS ONE, vol. 6, no. 1, page e16591, 01 2011. (Cited on page 107.)

[Balcan 2009] Duygu Balcan, Hao Hu, Bruno Goncalves, Paolo Bajardi, Chiara Poletto, Jose Ramasco, Daniela Paolotti, Nicola Perra, Michele Tizzoni, Wouter Broeck, Vittoria Colizza and Alessandro Vespignani. *Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo*

likelihood analysis based on human mobility. BMC Medicine, vol. 7, no. 1, page 45, 2009. (Cited on page 103.)

[Barabási 1999] Albert-László Barabási and Réka Albert. *Emergence of Scaling in Random Networks.* Science, vol. 286, no. 5439, pages 509–512, 1999. (Cited on pages iii, 2, 3, 27, 48 and 95.)

[Barabási 2005] Albert-László Barabási. *The origin of bursts and heavy tails in human dynamics.* Nature, vol. 435, no. 7039, pages 207–211, May 2005. (Cited on pages ix, 74 and 95.)

[Barahona 2002] Mauricio Barahona and Louis M. Pecora. *Synchronization in Small-World Systems.* Phys. Rev. Lett., vol. 89, page 054101, Jul 2002. (Cited on page 5.)

[Barrat 2004] Alain Barrat, Marc Barthélemy, Romualdo Pastor-Satorras and Alessandro Vespignani. *The architecture of complex weighted networks.* Proceedings of the National Academy of Sciences of the United States of America, vol. 101, no. 11, pages 3747–3752, 2004. (Cited on page 20.)

[Barrat 2008] Alain Barrat, Marc Barthélemy and Alessandro Vespignani. Dynamical Processes on Complex Networks. Cambridge University Press, New York, NY, USA, 1st édition, 2008. (Cited on pages 5, 17 and 98.)

[Barrat 2010] 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 2010), 2010. (Cited on page 62.)

[Bernoulli 1760] Daniel Bernoulli. *De la mortalité causée par la petite Vérole et des avantages de l'inoculation pour la prévenir.* Mémoires de Mathématique et de Physique tirés des Registres de l'Académie Royale des Sciences, 1760. (Cited on page 98.)

[Beutels 2006] Philipp Beutels, Ziv Shkedy, Marc Aerts and Pierre Van Damme. *Social mixing patterns for transmission models of close contact infections: exploring self-evaluation and diary-based data collection through a web-based interface.* Epidemiology and Infection, vol. 134, pages 1158–1166, 2006. (Cited on page 108.)

[Bianconi 2002] Ginestra Bianconi. *Mean field solution of the Ising model on a Barabási-Albert network.* Physics Letters A, vol. 303, pages 166 – 168, 2002. (Cited on page 5.)

[Bisson 2012] Giacomo Bisson, Ginestra Bianconi and Vincent Torre. *The Dynamics of Group Formation Among Leeches.* Frontiers in Physiology, vol. 3, no. 133, 2012. (Cited on page 80.)

[Blower 2011] 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, vol. 9, no. 1, page 88, 2011. (Cited on page 112.)

[Boccaletti 2006] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D.-U. Hwang. *Complex networks: Structure and dynamics.* Physics Reports, vol. 424, pages 175 – 308, 2006. (Cited on pages ix and 3.)

[Brandes 2009] Ulrik Brandes, Jürgen Lerner and Tom A. B. Snijders. *Networks Evolving Step by Step: Statistical Analysis of Dyadic Event Data.* In Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining, ASONAM '09, pages 200–205, Washington, DC, USA, 2009. IEEE Computer Society. (Cited on page 47.)

[Breban 2007] Romulus Breban, Raffaele Vardavas and Sally Blower. *Theory versus Data: How to Calculate R_0?* PLoS ONE, vol. 2, no. 3, page e282, 03 2007. (Cited on pages 103 and 119.)

[Carrat 2008] F. Carrat, E. Vergu, N.M. Ferguson, M. Lemaitre, S. Cauchemez, S. Leach and A.J. Valleron. *Time lines of infection and disease in human influenza: a review of volunteer challenge studies.* American Journal of Epidemiology, vol. 167, no. 7, pages 775–785, Apr 2008. (Cited on page 101.)

[Castellano 2009] Claudio Castellano, Santo Fortunato and Vittorio Loreto. *Statistical physics of social dynamics.* Rev. Mod. Phys., vol. 81, pages 591–646, May 2009. (Cited on page 75.)

[Cattuto 2010] 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, vol. 5, no. 7, pages e11596+, July 2010. (Cited on pages 11, 12, 13, 14 and 45.)

[Cauchemez 2009] Simon Cauchemez, Neil M Ferguson, Claude Wachtel, Anders Tegnell, Guillaume Saour, Ben Duncan and Angus Nicoll. *Closure of schools during an influenza pandemic.* The Lancet Infectious Diseases, vol. 9, no. 8, pages 473 – 481, 2009. (Cited on page 44.)

[Chaintreau 2007] Augustin Chaintreau, Pan Hui, Jon Crowcroft, Christophe Diot, Richard Gass and James Scott. *Impact of Human Mobility on Opportunistic Forwarding Algorithms.* IEEE Transactions on Mobile Computing, vol. 6, pages 606–620, 2007. (Cited on pages 9, 10 and 11.)

[Ciofi degli Atti 2008] Marta Luisa Ciofi degli Atti, Stefano Merler, Caterina Rizzo, Marco Ajelli, Marco Massari, Piero Manfredi, Cesare Furlanello, Gianpaolo Scalia Tomba and Mimmo Iannelli. *Mitigation Measures for Pandemic Influenza in Italy: An Individual Based Model Considering Different Scenarios.* PLoS ONE, vol. 3, no. 3, page e1790, 03 2008. (Cited on page 107.)

[Clauset 2009] Aaron Clauset, Cosma Rohilla Shalizi and Mark E. J. Newman. *Power-Law Distributions in Empirical Data.* SIAM Review, vol. 51, no. 4, pages 661–703, 2009. (Cited on page 106.)

[Colizza 2006a] Vittoria Colizza, Alain Barrat, Marc Barthélemy and Alessandro Vespignani. *The Modeling of Global Epidemics: Stochastic Dynamics and Predictability.* Bulletin of Mathematical Biology, vol. 68, pages 1893–1921, 2006. (Cited on page 107.)

[Colizza 2006b] Vittoria Colizza, Alain Barrat, Marc Barthélemy and Alessandro Vespignani. *The role of the airline transportation network in the prediction and predictability of global epidemics.* Proceedings of the National Academy of Sciences of the United States of America, vol. 103, no. 7, pages 2015–2020, 2006. (Cited on page 107.)

[Colizza 2007a] Vittoria Colizza, Alain Barrat, Marc Barthelemy, Alain-Jacques Valleron and Alessandro Vespignani. *Modeling the Worldwide Spread of Pandemic Influenza: Baseline Case and Containment Interventions.* PLoS Medicine, vol. 4, no. 01, page e13, 2007. (Cited on page 107.)

[Colizza 2007b] Vittoria Colizza, Marc Barthélemy, Alain Barrat and Alessandro Vespignani. *Epidemic modeling in complex realities.* C. R. Biologies, vol. 330, pages 364–374, 2007. (Cited on page 104.)

[Conlan 2011] Andrew J. K. Conlan, Ken T. D. Eames, Jenny A. Gage, Johann C. von Kirchbach, Joshua V. Ross, Roberto A. Saenz and Julia R. Gog. *Measuring social networks in British primary schools through scientific engagement.* Proceedings of the Royal Society B: Biological Sciences, vol. 278, no. 1711, pages 1467–1475, 2011. (Cited on pages 38, 44 and 48.)

[Criswell 1939] Joan Henning Criswell. *A sociometric Study of Race Cleavage in the Classroom.* Archives of Psychology, vol. 235, January 1939. (Cited on page 55.)

[Davidsen 2002] Jörn Davidsen, Holger Ebel and Stefan Bornholdt. *Emergence of a Small World from Local Interactions: Modeling Acquaintance Networks.* Phys. Rev. Lett., vol. 88, page 128701, Mar 2002. (Cited on page 7.)

[Del Valle 2007] Sara Y. Del Valle, James Mac Hyman, Herbert W. Hethcote and Stephen G. Eubank. *Mixing patterns between age groups in social networks.* Social Networks, vol. 29, no. 4, pages 539 – 554, 2007. (Cited on pages 42 and 43.)

[Diekmann 1990] O. Diekmann, J. A. P. Heesterbeek and J. A. J. Metz. *On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations.* Journal of Mathematical Biology, vol. 28, pages 365–382, 1990. (Cited on pages 103 and 119.)

[Dodds 2003] Peter Sheridan Dodds, Roby Muhamad and Duncan J. Watts. *An Experimental Study of Search in Global Social Networks.* Science, vol. 301, no. 5634, pages 827–829, 2003. (Cited on page 20.)

[Doreian 1980] Patrick Doreian. *Linear Models with Spatially Distributed Data.* Sociological Methods & Research, vol. 9, no. 1, pages 29–60, 1980. (Cited on page 71.)

[Dorogovtsev 2008] S. N. Dorogovtsev, A. V. Goltsev and J. F. F. Mendes. *Critical phenomena in complex networks.* Reviews of Modern Physics, vol. 80, pages 1275–1335, Oct 2008. (Cited on page 5.)

[Dowse 2011] Gary K. Dowse, David W. Smith, Heath Kelly, Ian Barr, Karen L. Laurie, Anthony R. Jones, Anthony D. Keil and Paul Effler. *Incidence of pandemic (H1N1) 2009 influenza infection in children and pregnant women during the 2009 influenza season in Western Australia – a seroprevalence study.* Medical Journal of Australia, vol. 194, no. 2, pages 68–72, 2011. (Cited on page 101.)

[Eagle 2006] Nathan Eagle and Alex (Sandy) Pentland. *Reality mining: sensing complex social systems.* Personal Ubiquitous Comput., vol. 10, pages 255–268, March 2006. (Cited on pages 9, 10 and 11.)

[Eames 2002] Ken T. D. Eames and Matt J. Keeling. *Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases.* Proceedings of the National Academy of Sciences, vol. 99, no. 20, pages 13330–13335, 2002. (Cited on page 106.)

[Eames 2008] Ken T.D. Eames. *Modeling disease spread through random and regular contacts in clustered populations.* Theoretical Population Biology, vol. 73, no. 1, pages 104 – 111, 2008. (Cited on page 107.)

[Eames 2011] Ken T.D. Eames, Natasha L. Tilston and W. John Edmunds. *The impact of school holidays on the social mixing patterns of school children.* Epidemics, vol. 3, no. 2, pages 103 – 108, 2011. (Cited on pages 45 and 109.)

[Ebel 2002] Holger Ebel, Lutz-Ingo Mielsch and Stefan Bornholdt. *Scale-free topology of e-mail networks.* Phys. Rev. E, vol. 66, page 035103, Sep 2002. (Cited on page 3.)

[Eckmann 2004] Jean-Pierre Eckmann, Elisha Moses and Danilo Sergi. *Entropy of dialogues creates coherent structures in e-mail traffic.* Proceedings of the National Academy of Sciences of the United States of America, vol. 101, no. 40, pages 14333–14337, 2004. (Cited on pages 31, 45 and 73.)

[Edmunds 1997] W. John 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. Series B: Biological Sciences, vol. 264, no. 1384, pages 949–957, 1997. (Cited on pages 105 and 108.)

[Erdős 1959] Paul Erdős and Alfréd Rényi. *On random graphs.* Publicationes Mathematicae, vol. 6, pages 290–297, 1959. (Cited on pages 2 and 18.)

[Eubank 2004] Stephen Eubank, Hasan Guclu, V. S. Anil Kumar, Madhav V. Marathe, Aravind Srinivasan, Zoltán Toroczkai and Nan Wang. *Modelling disease outbreaks in realistic urban social networks.* Nature, vol. 429, pages 180–184, May 2004. (Cited on page 107.)

[Euler 1741] Leinhard Euler. *Solutio problematis ad geometriam situs pertinentis.* Commentarii academiae scientiarum Petropolitanae, vol. 8, pages 128–140, 1741. (Cited on page 2.)

[Ferguson 2006] Neil M. Ferguson, Derek A. T. Cummings, Christophe Fraser, James C. Cajka, Philip C. Cooley and Donald S. Burke. *Strategies for mitigating an influenza pandemic.* Nature, vol. 442, no. 7101, pages 448 – 452, July 2006. (Cited on page 107.)

[Ferrand 1997] Alexis Ferrand. *La structure des systÃ¨mes de relations.* L'année sociologique, vol. 47, no. 1, pages 37–54, 1997. (Cited on page 46.)

[Fierro 2010] Raùl Fierro. *A class of stochastic epidemic models and its deterministic counterpart.* Journal of the Korean Statistical Society, vol. 39, no. 4, pages 397 – 407, 2010. (Cited on page 100.)

[Fraboulet 2007] Antoine Fraboulet, Guillaume Chelius and Eric Fleury. *Worldsens: development and prototyping tools for application specific wireless sensors networks.* In Proceedings of the 6th international conference on Information processing in sensor networks, IPSN '07, pages 176–185. ACM, 2007. (Cited on page 9.)

[Freeman 1996] Linton C. Freeman. *Some Antecedents of Social Network Analysis.* Concetions, vol. 19, no. 1, pages 39–42, 1996. (Cited on page 49.)

[Friggeri 2011] Adrien Friggeri, Guillaume Chelius, Eric Fleury, Antoine Fraboulet, France Mentré and Jean-Christophe Lucet. *Reconstructing Social Interactions Using an unreliable Wireless Sensor Network.* Computer Communications, vol. 34, no. 5, pages 609–618, April 2011. (Cited on pages 9, 10 and 11.)

[Gautreau 2009] Aurélien Gautreau, Alain Barrat and Marc Barthélemy. *Microdynamics in stationary complex networks.* Proceedings of the National Academy of Sciences, vol. 106, no. 22, pages 8847–8852, 2009. (Cited on page 6.)

[Germann 2006] Timothy C. Germann, Kai Kadau, Ira M. Longini and Catherine A. Macken. *Mitigation strategies for pandemic influenza in the United States.* Proceedings of the National Academy of Sciences, vol. 103, no. 15, pages 5935–5940, 2006. (Cited on page 107.)

[Gest 2003] Scott D. Gest, Thomas W. Farmer, Beverley D. Cairns and Hongling Xie. *Identifying Children's Peer Social Networks in School Classrooms: Links Between Peer Reports and Observed Interactions.* Social Development, vol. 12, no. 4, pages 513–529, 2003. (Cited on page 16.)

[Glass 2008] Laura Glass and Robert Glass. *Social contact networks for the spread of pandemic influenza in children and teenagers.* BMC Public Health, vol. 8, no. 61, pages 1–15, 2008. (Cited on pages 42 and 43.)

[Goldenberg 2009] Anna Goldenberg, Alice X. Zheng, Stephen E. Fienberg and Edoardo M. Airoldi. *A Survey of Statistical Network Models.* Foundations and Trends in Machine Learning, vol. 2, no. 2, pages 129–233, 2009. (Cited on page 7.)

[Golder 2007] Scott A. Golder, Dennis M. Wilkinson and Bernardo A. Huberman. *Rhythms of Social Interaction: Messaging Within a Massive Online Network.* In Charles Steinfield, Brian T. Pentland, Mark Ackerman and Noshir Contractor, editeurs, Communities and Technologies 2007, pages 41–66. Springer London, 2007. (Cited on page 6.)

[González 2008] M. C. González, C. A. Hidalgo and A. L. Barabasi. *Understanding individual human mobility patterns.* Nature, vol. 453, page 479, 2008. (Cited on pages 6 and 72.)

[Granovetter 1973] Mark S. Granovetter. *The Strength of Weak Ties.* American Journal of Sociology, vol. 78, no. 6, pages 1360–1380, 1973. (Cited on pages 14, 52, 61 and 132.)

[Gräser 2009] O. Gräser, C. Xu and Pan M. Hui. *Disconnected-connected network transitions and phase separation driven by co-evolving dynamics.* Europhysics Letters, vol. 87, no. 3, page 38003, 2009. (Cited on page 7.)

[Greeno 1989] Catherine G. Greeno. Gender differences in children's proximity to adults. Doctoral dissertation at Stanford University, CA, 1989. (Cited on pages 45 and 48.)

[Gross 2009] Thilo Gross and Hiroki Sayama. *Adaptive Networks.* In Thilo Gross and Hiroki Sayama, editeurs, Adaptive Networks, volume 51 of *Understanding Complex Systems*, pages 1–8. Springer Berlin / Heidelberg, 2009. (Cited on page 7.)

[Gueron 1995] Shay Gueron and Simon A. Levin. *The dynamics of group formation.* Mathematical Biosciences, vol. 128, no. 1–2, pages 243 – 264, 1995. (Cited on page 80.)

[Handcock 2010] Mark S. Handcock and Krista J. Gile. *Modeling social networks from sampled data.* Annals of Applied Statistics, vol. 4, no. 1, pages 5–25, 2010. (Cited on page 127.)

[Hayden-Thomson 1987] Laura Hayden-Thomson, Kenneth H. Rubin and Shelley Hymel. *Sex Preferences in Sociometric Choices.* Developmental Psychology, vol. 23, no. 4, pages 558 –562, 1987. (Cited on page 50.)

[Haydon 2003] Daniel T. Haydon, Margo Chase-Topping, D. J. Shaw, L. Matthews, J. K. Friar, J. Wilesmith and M. E. J. Woolhouse. *The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak.* Proceedings of the Royal Society of London. Series B: Biological Sciences, vol. 270, no. 1511, pages 121–127, 2003. (Cited on page 108.)

[Hechter 1997] Michael Hechter and Satoshi Kanazawa. *Sociological Rational Choice Theory.* Annual Review of Sociology, vol. 23, no. 1, pages 191–214, 1997. (Cited on page 75.)

[Heffernan 2005] J.M Heffernan, R.J Smith and L.M Wahl. *Perspectives on the basic reproductive ratio.* Journal of The Royal Society Interface, vol. 2, no. 4, pages 281–293, 2005. (Cited on pages 103, 105 and 119.)

[Henderson 2004] Tristan Henderson, David Kotz and Ilya Abyzov. *The changing usage of a mature campus-wide wireless network.* In Proceedings of the 10th annual international conference on Mobile computing and networking, MobiCom '04, pages 187–201, New York, NY, USA, 2004. ACM. (Cited on page 9.)

[Hens 2009] Niel Hens, Nele Goeyvaerts, Marc Aerts, Ziv Shkedy, Pierre Van Damme and Philippe Beutels. *Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium.* BMC Infectious Diseases, vol. 9, no. 1, page 5, 2009. (Cited on page 108.)

[Herfindahl 1959] Oris C Herfindahl. Copper costs and prices: 1870-1957. Baltimore: Johns Hopkins University Press, 1959. (Cited on page 19.)

[Hill 2010] Scott A. Hill and Dan Braha. *Dynamic model of time-dependent complex networks.* Phys. Rev. E, vol. 82, page 046105, Oct 2010. (Cited on page 7.)

[Hirschman 1964] Albert O. Hirschman. *The Paternity of an Index.* The American Economic Review, vol. 54, no. 5, page 761, Sep 1964. (Cited on page 19.)

[Hogan 2007] Bernie Hogan, Juan Antonio Carrasco and Barry Wellman. *Visualizing Personal Networks: Working with Participant-aided Sociograms.* Field Methods, vol. 19, no. 2, pages 116–144, 2007. (Cited on page 8.)

[Holland 1977] Paul W. Holland and Samuel Leinhardt. *A dynamic model for social networks.* The Journal of Mathematical Sociology, vol. 5, no. 1, pages 5–20, 1977. (Cited on page 7.)

[Holme 2006] Petter Holme and Mark E. J. Newman. *Nonequilibrium phase transition in the coevolution of networks and opinions.* Phys. Rev. E, vol. 74, page 056108, Nov 2006. (Cited on page 5.)

[Holme 2012] Petter Holme and Jari Saramäki. *Temporal networks.* Physics Reports, 2012. (Cited on pages 6, 7 and 20.)

[Hsieh 2002] Ying-Hen Hsieh, Hector de Arazoza, Shen-Ming Lee and Cathy WS Chen. *Estimating the number of Cubans infected sexually by human immunodeficiency virus using contact tracing data.* International Journal of Epidemiology, vol. 31, no. 3, pages 679–683, 2002. (Cited on page 108.)

[Hsieh 2010] Ying-Hen Hsieh, Yun-Shih Wang, Hector de Arazoza and Rachid Lounes. *Modeling secondary level of HIV contact tracing: its impact on HIV intervention in Cuba.* BMC Infectious Diseases, vol. 10, no. 1, page 194, 2010. (Cited on page 108.)

[Hufnagel 2004] Lars Hufnagel, Dirk Brockmann and Theo Geisel. *Forecast and control of epidemics in a globalized world.* Proceedings of the National Academy of Sciences of the United States of America, vol. 101, no. 42, pages 15124–15129, 2004. (Cited on page 107.)

[Hui 2005] Pan Hui, Augustin Chaintreau, James Scott, Richard Gass, Jon Crowcroft and Christophe Diot. *Pocket switched networks and human mobility in conference environments.* In Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking, WDTN'05, pages 244–251, New York, NY, USA, 2005. ACM. (Cited on pages v, 9, 45 and 73.)

[Iribarren 2009] José Luis Iribarren and Esteban Moro. *Impact of Human Activity Patterns on the Dynamics of Information Diffusion.* Phys. Rev. Lett., vol. 103, page 038702, Jul 2009. (Cited on page 110.)

[Isella 2010] 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, vol. 271, pages 166–180, 2010. (Cited on pages 14, 21, 110 and 124.)

[Isella 2011] 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, vol. 6, no. 2, page e17144, 02 2011. (Cited on pages 32, 45, 109 and 128.)

[Jackson 2011] Charlotte Jackson, Punam Mangtani, Emilia Vynnycky, Katherine Fielding, Aileen Kitching, Huda Mohamed, Anita Roche and Helen Maguire. *School closures and student contact patterns.* Emerging infectious diseases, vol. 17, no. 2, pages 245–247, 2011. (Cited on page 45.)

[Johnson 2009] Neil F. Johnson, Chen Xu, Zhenyuan Zhao, Nicolas Ducheneaut, Nicholas Yee, George Tita and Pak Ming Hui. *Human group formation in online guilds and offline gangs driven by a common team dynamic.* Phys. Rev. E, vol. 79, page 066117, Jun 2009. (Cited on pages 75 and 95.)

[Jones 2003] James H. Jones and Mark S. Handcock. *Sexual contacts and epidemic thresholds.* Nature, vol. 423, pages 605–606, June 2003. (Cited on page 106.)

[Karsai 2011] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A. L. Barabási and J. Saramäki. *Small but slow world: How network topology and burstiness slow down spreading.* Physical Review E, vol. 83, no. 2, pages 025102+, 2011. (Cited on pages 96 and 110.)

[Kazandjieva 2010] Maria A. Kazandjieva, Jung Woo Lee, Marcel Salathé, Marcus W. Feldman, James H. Jones and Philip Levis. *Experiences in measuring a human contact network for epidemiology research.* In Proceedings of the 6th Workshop on Hot Topics in Embedded Networked Sensors, HotEmNets '10, pages 7:1–7:5, New York, NY, USA, 2010. ACM. (Cited on page 9.)

[Keeling 1999] Matt J. Keeling. *The effects of local spatial structure on epidemiological invasions.* Proceedings of the Royal Society of London. Series B: Biological Sciences, vol. 266, no. 1421, pages 859–867, 1999. (Cited on page 106.)

[Keeling 2005] Matt J Keeling and Ken T.D Eames. *Networks and epidemic models.* Journal of The Royal Society Interface, vol. 2, no. 4, pages 295–307, 2005. (Cited on pages 106 and 108.)

[Kermack 1927] William Ogilvy Kermack and Anderson Gray McKendrick. *Contribution to the mathematical theory of epidemics.* Proceedings of the Royal Society of London A, vol. 115, pages 700–721, 1927. (Cited on pages 100 and 105.)

[Kitsak 2010] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley and H. A. Makse. *Identification of influential spreaders in complex networks.* Nature Physics, vol. 6, pages 888–893, November 2010. (Cited on page 128.)

[Knoke 2008] David Knoke and Song Yang. Social network analysis. Numeéro 07-154 de Quantitative applications in the social sciences. Sage Publications, second édition, 2008. (Cited on pages 14 and 109.)

[Kolaczyk 2009] Eric D. Kolaczyk. Statistical analysis of network data: Methods and models. Springer Publishing Company, Incorporated, 1st édition, 2009. (Cited on page 18.)

[Kossinets 2006] Gueorgi Kossinets and Duncan J. Watts. *Empirical Analysis of an Evolving Social Network.* Science, vol. 311, no. 5757, pages 88–90, 2006. (Cited on page 6.)

[Kossinets 2008] Gueorgi Kossinets, Jon Kleinberg and Duncan Watts. *The structure of information pathways in a social communication network.* In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '08, pages 435–443, New York, NY, USA, 2008. ACM. (Cited on page 6.)

[Kossinets 2009] Gueorgi Kossinets and Duncan J. Watts. *Origins of homophily in an evolving social network.* American Journal of Sociology, vol. 115, pages 405–450, 2009. (Cited on page 47.)

[Kostakos 2009] Vassilis Kostakos. *Temporal graphs.* Physica A, vol. 388, no. 6, pages 1007–1023, 2009. (Cited on page 6.)

[Kostakos 2010] Vassilis Kostakos, Eamonn O'Neill, Alan Penn, George Roussos and Dikaios Papadongonas. *Brief encounters: Sensing, modeling and visualizing urban mobility and copresence networks.* ACM Trans. Comput.-Hum. Interact., vol. 17, pages 2:1–2:38, April 2010. (Cited on pages 9, 10 and 11.)

[Kovanen 2011] Lauri Kovanen, Jari Saramaki and Kimmo Kaski. *Reciprocity of mobile phone calls.* Dynamics of Socio-Economic Systems, vol. 2, no. 2, pages 138–151, 2011. (Cited on page 48.)

[Kozma 2008] Balazs Kozma and Alain Barrat. *Consensus formation on adaptive networks.* Phys. Rev. E, vol. 77, page 016102, Jan 2008. (Cited on page 5.)

[Kumar 2010] Ravi Kumar, Jasmine Novak and Andrew Tomkins. *Structure and Evolution of Online Social Networks.* In Philip S. S. Yu, Jiawei Han and Christos Faloutsos, editeurs, Link Mining: Models, Algorithms, and Applications, pages 337–357. Springer New York, 2010. (Cited on page 48.)

[La Freniere 1984] Peter La Freniere, F. F. Strayer and Roger Gauthier. *The emergence of same-sex affiliative preference among preschool peers: a developmental/ethological perspective.* Child development, vol. 55, no. 5, pages 1958–1965, 1984. (Cited on pages 9 and 49.)

[Lazer 2009] 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, vol. 323, no. 5915, pages 721–723, 2009. (Cited on page 3.)

[Lee 2007] Linda Lee, Carollee Howes and Brandt Chamberlain. *Ethnic Heterogeneity of Social Networks and Cross-Ethnic Friendships of Elementary School Boys and Girls.* Merril-Palmer Quaterly, vol. 53, no. 3, pages 325 – 346, July 2007. (Cited on page 49.)

[Leguay 2006] Jérémie Leguay, Anders Lindgren, James Scott, Timur Friedman and Jon Crowcroft. *Opportunistic content distribution in an urban setting.* In Proceedings of the 2006 SIGCOMM workshop on Challenged networks, CHANTS '06, pages 205–212, 2006. (Cited on pages 10 and 11.)

[Leskovec 2008] Jure Leskovec and Eric Horvitz. *Planetary-scale views on a large instant-messaging network.* In Proceedings of the 17th international conference on World Wide Web, WWW '08, pages 915–924, New York, NY, USA, 2008. ACM. (Cited on pages 6, 20, 31, 45 and 48.)

[Lewis 2008] Kevin Lewis, Jason Kaufman, Marco Gonzalez, Andreas Wimmer and Nicholas Christakis. *Tastes, ties, and time: A new social network dataset using Facebook.com.* Social Networks, vol. 30, no. 4, pages 330 – 342, 2008. (Cited on page 47.)

[Liben-Nowell 2005] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan and Andrew Tomkins. *Geographic routing in social networks.* Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 33, pages 11623–11628, 2005. (Cited on pages ix and 71.)

[Liljeros 2001] Fredrik Liljeros, Christofer R. Edling, Luis A. Nunes Amaral, H. Eugène Stanley and Yvonne Åberg. *The web of human sexual contacts.* Nature, vol. 411, pages 907–908, June 2001. (Cited on pages 3, 24, 98, 106 and 108.)

[Lloyd 2001] Alun L. Lloyd and Robert M. May. *How Viruses Spread Among Computers and People.* Science, vol. 292, no. 5520, pages 1316–1317, 2001. (Cited on page 98.)

[Longini 1982] Ira M. Longini, James S. Koopman, Arnold S Monto and John P. Fox. *Estimating Household and Community Transmission Parameters for Influenza.* American Journal of Epidemiology, vol. 115, no. 5, pages 736–751, 1982. (Cited on page 32.)

[Longini 2005] Ira M. Longini, Azhar Nizam, Shufu Xu, Kumnuan Ungchusak, Wanna Hanshaoworakul, Derek A. T. Cummings and M. Elizabeth Halloran. *Containing Pandemic Influenza at the Source.* Science, vol. 309, no. 5737, pages 1083–1087, 2005. (Cited on page 107.)

[Lubbers 2010] Miranda J. Lubbers, José Luis Molina, Jürgen Lerner, Ulrik Brandes, Javier Ávila and Christopher McCarty. *Longitudinal analysis of personal networks. The case of Argentinean migrants in Spain.* Social Networks, vol. 32, no. 1, pages 91 – 104, 2010. (Cited on page 8.)

[Maccoby 1990] Eleanor E. Maccoby. *Gender and Relationships: A Developmental Account.* American Psychologist, vol. 45, no. 4, pages 513 – 520, 1990. (Cited on pages 45 and 48.)

[Maccoby 2003] Eleanor E. Maccoby. The two sexes: Growing up apart, coming together. Belknap Press of Harvard University Press, 5 édition, 2003. (Cited on pages 49 and 58.)

[Malmgren 2009] R. Dean Malmgren, Daniel B. Stouffer, Andriana S. L. O. Campanharo and Luís A. Nunes Amaral. *On Universality in Human Correspondence Activity.* Science, vol. 325, no. 5948, pages 1696–1700, 2009. (Cited on page 6.)

[Mandelbrot 1953] Benoit Mandelbrot. *An Informational Theory of the Statistical Structure of Language.* In Willis Jackson, editeur, Communication Theory, The Second London Symposium, pages 486–504, London, 1953. Butterworth. (Cited on page 74.)

[Manski 1993] Charles F. Manski. *Identification of Endogenous Social Effects: The Reflection Problem.* The Review of Economic Studies, vol. 60, no. 3, pages 531–542, 1993. (Cited on pages 47 and 71.)

[Marsden 1984] Peter V. Marsden and Karen E. Campbell. *Measuring Tie Strength.* Social Forces, vol. 63, no. 2, pages 482–501, 1984. (Cited on page 132.)

[Marsden 1990] Peter V. Marsden. *Network Data and Measurement.* Annual Review of Sociology, vol. 16, no. 1, pages 435–463, 1990. (Cited on pages 9 and 14.)

[Marsden 2011] Peter V. Marsden. Survey methods for network data, pages 370–388. Sage Publications, London, 2011. (Cited on page 9.)

[Martin 2001] Carol Lynn Martin and Richard A. Fabes. *The Stability and Consequences of Young Children's Same-Sex Peer Interactions.* Developmental Psychology, vol. 37, no. 3, pages 431 – 446, 2001. (Cited on pages 9 and 49.)

[Maslov 2004] Sergei Maslov, Kim Sneppen and Alexei Zaliznyak. *Detection of topological patterns in complex networks: correlation profile of the internet.* Physica A, vol. 333, pages 529 – 540, 2004. (Cited on page 28.)

[Masuda 2009] Naoki Masuda. *Immunization of networks with community structure.* New Journal of Physics, vol. 11, no. 12, page 123018, 2009. (Cited on page 128.)

[McNett 2005] Marvin McNett and Geoffrey M. Voelker. *Access and mobility of wireless PDA users.* SIGMOBILE Mob. Comput. Commun. Rev., vol. 9, no. 2, April 2005. (Cited on page 9.)

[McPherson 2001] Miller McPherson, Lynn Smith-Lovin and James M. Cook. *Birds of a Feather: Homophily in Social Networks.* Annual Review of Sociology, vol. 27, pages 415–445, 2001. (Cited on page 46.)

[Mehta 2009] Clare M. Mehta and JoNell Strough. *Sex segregation in friendships and normative contexts across the life span.* Developmental Review, vol. 29, no. 3, pages 201 – 220, 2009. (Cited on pages 49 and 58.)

[Merler 2010] Stefano Merler and Marco Ajelli. *The role of population heterogeneity and human mobility in the spread of pandemic influenza.* Proceedings of the Royal Society B: Biological Sciences, vol. 277, no. 1681, pages 557–565, 2010. (Cited on page 107.)

[Mikolajczyk 2008] Rafael Mikolajczyk, Manas Akmatov, S Rastin and Mirjam Kretzschmar. *Social contacts of school children and the transmission of respiratory-spread pathogens.* Epidemiology and Infection, vol. 136, no. 6, pages 813–822, June 2008. (Cited on pages 42 and 44.)

[Milgram 1967] Stanley Milgram. *The Small World Problem.* Psychology Today, vol. 1, pages 61–67, May 1967. (Cited on pages 20 and 48.)

[Miritello 2011] Giovanna Miritello, Esteban Moro and Rubén Lara. *Dynamical strength of social ties in information spreading.* Phys. Rev. E, vol. 83, page 045102, Apr 2011. (Cited on pages 22, 96 and 110.)

[Mizruchi 1996] Mark S. Mizruchi. *What Do Interlocks Do? An Analysis, Critique, and Assessment of Research on Interlocking Directorates.* Annual Review of Sociology, vol. 22, pages 271–298, 1996. (Cited on page 8.)

[Moreno 1953] Jacob Levy Moreno. Who shall survive? foundations of sociometry, group psychotherapy and socio-drama. Oxford, England: Beacon House, 2nd ed. édition, 1953. (Cited on pages 2, 8, 49 and 58.)

[Morgan 1976] Byron J.T. Morgan. *Stochastic Models of Grouping Changes.* Advances in Applied Probability, vol. 8, no. 1, pages 30–57, March 1976. (Cited on pages 75 and 80.)

[Mossong 2008] 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, vol. 5, no. 3, page e74, 03 2008. (Cited on pages 42, 105, 108 and 109.)

[Murray 2007] Christopher JL Murray, Alan D Lopez, Brian Chin, Dennis Feehan and Kenneth H Hill. *Estimation of potential global pandemic influenza mortality on the basis of vital registry data from the 1918–20 pandemic: a*

quantitative analysis. The Lancet, vol. 368, no. 9554, pages 2211 – 2218, 2007. (Cited on page 103.)

[Nardini 2008] Cecilia Nardini, Balázs Kozma and Alain Barrat. *Who's Talking First? Consensus or Lack Thereof in Coevolving Opinion Formation Models*. Phys. Rev. Lett., vol. 100, page 158701, Apr 2008. (Cited on page 5.)

[Newman 2001] Mark E. J. Newman. *The structure of scientific collaboration networks*. Proceedings of the National Academy of Sciences, vol. 98, no. 2, pages 404–409, 2001. (Cited on page 24.)

[Newman 2003] Mark E. J. Newman. *The Structure and Function of Complex Networks*. SIAM Review, vol. 45, no. 2, pages 167–256, 2003. (Cited on page 3.)

[Noh 2004] Jae Dong Noh and Heiko Rieger. *Random Walks on Complex Networks*. Phys. Rev. Lett., vol. 92, page 118701, Mar 2004. (Cited on page 5.)

[O'Neill 2006] Eamonn O'Neill, Vassilis Kostakos, Tim Kindberg, Ava Schiek, Alan Penn, Danaë Fraser and Tim Jones. *Instrumenting the City: Developing Methods for Observing and Understanding the Digital Cityscape*. In Paul Dourish and Adrian Friday, editeurs, UbiComp 2006: Ubiquitous Computing, volume 4206 of *Lecture Notes in Computer Science*, pages 315–332. Springer Berlin / Heidelberg, 2006. (Cited on page 9.)

[Onnela 2007] Jukka-Pekka Onnela, Jari Saramäki, J. Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész and Albert-László Barabási. *Structure and tie strengths in mobile communication networks*. Proceedings of the National Academy of Sciences, vol. 104, no. 18, pages 7332–7336, 2007. (Cited on pages 6, 27, 31, 45 and 73.)

[Padgett 1993] John F. Padgett and Christopher K. Ansell. *Robust Action and the Rise of the Medici, 1400-1434*. American Journal of Sociology, vol. 98, no. 6, page 1259, 1993. (Cited on page 8.)

[Pan 2011] Raj Kumar Pan and Jari Saramäki. *Path lengths, correlations, and centrality in temporal networks*. Phys. Rev. E, vol. 84, page 016105, Jul 2011. (Cited on pages 6 and 20.)

[Pastor-Satorras 2001] Romualdo Pastor-Satorras and Alessandro Vespignani. *Epidemic Spreading in Scale-Free Networks*. Phys. Rev. Lett., vol. 86, pages 3200–3203, Apr 2001. (Cited on pages 5, 98 and 106.)

[Pastor-Satorras 2004] Romualdo Pastor-Satorras and Alessandro Vespignani. Evolution and structure of the internet: A statistical physics approach. Cambridge University Press, New York, NY, USA, 2004. (Cited on pages 22 and 83.)

[Polgreen 2010] Philip M. Polgreen, Troy Leo Tassier, Sriram Venkata Pemmaraju and Alberto Maria Segre. *Prioritizing Healthcare Worker Vaccinations on the Basis of Social Network Analysis.* Infection Control and Hospital Epidemiology, vol. 31, no. 9, pages pp. 893–900, 2010. (Cited on page 128.)

[Potter 2012] Gail E. Potter, Mark S. Handcock, Ira M. Longini and M. Elizabeth Halloran. *Estimating within-school contact networks to understand influenza transmission.* Annals of Applied Statistics, vol. 6, no. 1, pages 1–26, 2012. (Cited on page 124.)

[Poulin 2007] François Poulin and Sara Pedersen. *Developmental Changes in Gender Composition of Friendship Networks in Adolescent Girls and Boys.* Developmental Psychology, vol. 43, no. 6, pages 1484–1496, 2007. (Cited on pages 50 and 60.)

[Poulin 2010] François Poulin and Alessandra Chan. *Friendship stability and change in childhood and adolescence.* Developmental Review, vol. 30, pages 257 – 272, 2010. (Cited on pages 50 and 58.)

[Read 2008] 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, vol. 5, no. 26, pages 1001–1007, 2008. (Cited on pages 108, 124 and 128.)

[Richards 1998] Maryse H. Richards, Paul A. Crowe, Reed Larson and Amy Swarr. *Developmental Patterns and Gender Differences in the Experience of Peer Companionship during Adolescence.* Child Development, vol. 69, no. 1, pages 154–163, 1998. (Cited on pages 50 and 60.)

[Riley 2003] Steven Riley, Christophe Fraser, Christl A. Donnelly, Azra C. Ghani, Laith J. Abu-Raddad, Anthony J. Hedley, Gabriel M. Leung, Lai-Ming Ho, Tai-Hing Lam, Thuan Q. Thach, Patsy Chau, King-Pan Chan, Su-Vui Lo, Pak-Yin Leung, Thomas Tsang, William Ho, Koon-Hung Lee, Edith M. C. Lau, Neil M. Ferguson and Roy M. Anderson. *Transmission Dynamics of the Etiological Agent of SARS in Hong Kong: Impact of Public Health Interventions.* Science, vol. 300, no. 5627, pages 1961–1966, 2003. (Cited on page 108.)

[Riley 2007] Steven Riley. *Large-Scale Spatial-Transmission Models of Infectious Disease.* Science, vol. 316, no. 5829, pages 1298–1301, 2007. (Cited on page 104.)

[Ripley 2011] Ruth M. Ripley, Tom A.B. Snijders and Paulina Preciado. Manual for siena version 4.0. Oxford: University of Oxford, Department of Statistics; Nuffield College, 2011. (version January 17, 2012). (Cited on page 57.)

[Rocha 2011] Luis E. C. Rocha, Fredrik Liljeros and Petter Holme. *Simulated Epidemics in an Empirical Spatiotemporal Network of 50,185 Sexual Contacts.*

PLoS Comput Biol, vol. 7, no. 3, page e1001109, 03 2011. (Cited on pages 96, 107, 110 and 115.)

[Rocha 2012] Luis E. C. Rocha and V. D. Blondel. *Temporal Heterogeneities Increase the Prevalence of Epidemics on Evolving Networks.* ArXiv e-prints, June 2012. (Cited on pages ix and 95.)

[Rvachev 1985] Leonid A. Rvachev and Ira M. Longini Jr. *A mathematical model for the global spread of influenza.* Mathematical Biosciences, vol. 75, no. 1, pages 3 – 22, 1985. (Cited on page 107.)

[Rybski 2009] Diego Rybski, Sergey V. Buldyrev, Shlomo Havlin, Fredrik Liljeros and Hernán A. Makse. *Scaling laws of human interaction activity.* Proceedings of the National Academy of Sciences, vol. 106, no. 31, pages 12640–12645, 2009. (Cited on pages v, 31, 45 and 73.)

[Salathé 2010] 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.* Proceedings of the National Academy of Science, vol. 1072, pages 22020–22025, December 2010. (Cited on pages v, 9, 10, 11, 40, 42, 44, 45 and 109.)

[Schaefer 2010] David R. Schaefer, John M. Light, Richard A. Fabes, Laura D. Hanish and Carol Lynn Martin. *Fundamental principles of network formation among preschool children.* Social Networks, vol. 32, no. 1, pages 61 – 71, 2010. (Cited on pages 47 and 72.)

[Schelling 1971] Thomas C. Schelling. *Dynamic models of segregation.* The Journal of Mathematical Sociology, vol. 1, no. 2, pages 143–186, 1971. (Cited on page 107.)

[Scherrer 2008] A. Scherrer, P. Borgnat, E. Fleury, J.-L. Guillaume and C. Robardet. *Description and simulation of dynamic mobility networks.* Computer Networks, vol. 52, no. 15, pages 2842 – 2858, 2008. (Cited on pages 7 and 95.)

[Serfling 1952] Robert E. Serfling. *Historical Review of Epidemic Theory.* Human Biology, vol. 24, no. 3, page 145, Sept 1952. (Cited on pages 98, 100 and 108.)

[Shalizi 2010] Cosma Rohilla Shalizi and Andrew C Thomas. *Homophily and Contagion Are Generically Confounded in Observational Social Network Studies.* Sociological Methods Research, vol. 40, no. 2, page 27, 2010. (Cited on page 47.)

[Shrum 1988] Wesley Shrum, Neil H. Cheek and Saundra MacD. Hunter. *Friendship in School: Gender and Racial Homophily.* Sociology of Education, vol. 61, no. 4, pages 227–239, 1988. (Cited on page 50.)

[Simon 1955] Herbert A. Simon. *On a Class of Skew Distribution Functions.* Biometrika, vol. 42, no. 3/4, pages pp. 425–440, 1955. (Cited on page 74.)

[Smieszek 2009a] Timo Smieszek. *A mechanistic model of infection: why duration and intensity of contacts should be included in models of disease spread.* Theoretical Biology and Medical Modelling, vol. 6, no. 1, page 25, 2009. (Cited on pages 109, 124 and 128.)

[Smieszek 2009b] Timo Smieszek, Lena Fiebig and Roland Scholz. *Models of epidemics: when contact repetition and clustering should be included.* Theoretical Biology and Medical Modelling, vol. 6, pages 1–15, 2009. (Cited on pages 107, 109, 117, 124 and 128.)

[Smieszek 2011] Timo Smieszek, E.U. Burri, R. Scherzinger and Roland W. Scholz. *Collecting close-contact social mixing data with contact diaries: reporting errors and biases.* Epidemiol. Infect., vol. 21, pages 1–9, June 2011. (Cited on pages vi, 14 and 109.)

[Snijders 2007] T. Snijders, C. Steglich and M. Schweinberger. *Modeling the coevolution of networks and behavior.* In Han Oud Kees van Montfort and Albert Satorra, editeurs, Longitudinal models in the behavioral and related sciences, pages 41–71. Lawrence Erlbaum, 2007. (Cited on page 7.)

[Song 2010] Chaoming Song, Zehui Qu, Nicholas Blumm and Albert-László Barabási. *Limits of Predictability in Human Mobility.* Science, vol. 327, no. 5968, pages 1018–1021, 2010. (Cited on page 72.)

[Steglich 2010] Christian Steglich, Tom A. B. Snijders and Michael Pearson. *Dynamic networks and behavior: separating selection from influence.* Sociological methodology, pages 329–393, 2010. (Cited on pages 47, 71, 72 and 132.)

[Stehlé 2010] Juliette Stehlé, Alain Barrat and Ginestra Bianconi. *Dynamical and bursty interactions in social networks.* Phys. Rev. E, vol. 81, no. 3, page 035101, Mar 2010. (Cited on page 75.)

[Stehle 2011a] Juliette Stehle, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Vittoria Colizza, Lorenzo Isella, Corinne Regis, 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, vol. 9, no. 1, page 87, 2011. (Cited on page 112.)

[Stehlé 2011b] 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, vol. 6, no. 8, page e23176, 08 2011. (Cited on pages 32, 42 and 109.)

[Stehlé 2012] Juliette Stehlé, François Charbonnier, Tristan Picard, Alain Barrat and Ciro Cattuto. Gender homophily among children in a primary school: a sociometric study using high-resolution detection of face-to-face contact patterns. Unpublished, 2012. (Cited on page 48.)

[Stumpf 2012] Michael P. H. Stumpf and Mason A. Porter. *Critical Truths About Power Laws*. Science, vol. 335, no. 6069, pages 665–666, 2012. (Cited on page 106.)

[Szell 2010] Michael Szell, Renaud Lambiotte and Stefan Thurner. *Multirelational Organization of Large-scale Social Networks in an Online World*. Proceedings of the National Academy of Sciences, vol. 107, no. 31, pages 13636–13641, 2010. (Cited on page 48.)

[Szendrői 2004] Balázs Szendrői and Gábor Csányi. *Polynomial epidemics and clustering in contact networks*. Proceedings of the Royal Society of London. Series B: Biological Sciences, vol. 271, pages S364–S366, 2004. (Cited on page 107.)

[Szomszor 2010] Martin Szomszor, Ciro Cattuto, Wouter Van den Broeck, Alain Barrat and Harith Alani. *Semantics, Sensors, and the Social Web: The Live Social Semantics Experiments*. In Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral and Tania Tudorache, editeurs, The Semantic Web: Research and Applications, volume 6089 of *Lecture Notes in Computer Science*, pages 196–210. Springer Berlin / Heidelberg, 2010. (Cited on page 62.)

[Takaguchi 2011] Taro Takaguchi, Mitsuhiro Nakamura, Nobuo Sato, Kazuo Yano and Naoki Masuda. *Predictability of Conversation Partners*. Phys. Rev. X, vol. 1, page 011008, Sep 2011. (Cited on pages 10, 11 and 72.)

[Takhteyev 2012] Yuri Takhteyev, Anatoliy Gruzd and Barry Wellman. *Geography of Twitter networks*. Social Networks, vol. 34, no. 1, pages 73 – 81, 2012. (Cited on pages ix, 47 and 71.)

[Tang 2010] J. Tang, S. Scellato, M. Musolesi, C. Mascolo and V. Latora. *Small-world behavior in time-varying graphs*. Phys. Rev. E, vol. 81, page 055101, May 2010. (Cited on page 6.)

[Traud 2012] Amanda L. Traud, Peter J. Mucha and Mason A. Porter. *Social structure of Facebook networks*. Physica A: Statistical Mechanics and its Applications, vol. 391, no. 16, pages 4165 – 4180, 2012. (Cited on page 48.)

[Travers 1969] Jeffrey Travers and Stanley Milgram. *An Experimental Study of the Small World Problem*. Sociometry, vol. 32, no. 4, pages 425–443, 1969. (Cited on page 20.)

[Van der Broeck 2010] Wouter Van der Broeck, Ciro Cattuto, Alain Barrat, Martin Szomszor, Gianluca Correndo and Harith Alani. *The Live Social Semantics application: a platform for integrating face-to-face presence with online social networking*. In First International Workshop on Communication, Collaboration and Social Networking in Pervasive Computing Environments (PerCol 2010), April 2010. (Cited on page 62.)

[Vázquez 2006] Alexei Vázquez, João Gama Oliveira, Zoltán Dezsö, Kwang-Il Goh, Imre Kondor and Albert-László Barabási. *Modeling bursts and heavy tails in human dynamics.* Phys. Rev. E, vol. 73, page 036127, Mar 2006. (Cited on page 74.)

[Vázquez 2007] Alexei Vázquez, Balázs Rácz, András Lukács and Albert-László Barabási. *Impact of Non-Poissonian Activity Patterns on Spreading Processes.* Phys. Rev. Lett., vol. 98, page 158702, Apr 2007. (Cited on pages 96 and 109.)

[Vazquez 2008] Federico Vazquez, Víctor M. Eguíluz and Maxi San Miguel. *Generic Absorbing Transition in Coevolution Dynamics.* Phys. Rev. Lett., vol. 100, page 108702, Mar 2008. (Cited on page 5.)

[Viboud 2004] Cécile Viboud, Pierre-Yves Boëlle, Simon Cauchemez, Audrey Lavenu, Alain-Jacques Valleron, Antoine Flahault and Fabrice Carrat. *Risk factors of influenza transmission in households.* British Journal of General Practice, vol. 54, no. 506, pages 684–689, 2004. (Cited on page 32.)

[Vigil 2007] Jacob Vigil. *Asymmetries in the Friendship Preferences and Social Styles of Men and Women.* Human Nature, vol. 18, pages 143–161, 2007. (Cited on pages 49 and 52.)

[Wakisaka 2009] Yoshihiro Wakisaka, Koji Ara, Miki Hayakawa, Youichi Horry, Norihiko Moriwaki, Norio Ohkubo, Nobuo Sato, Satomi Tsuji and Kazuo Yano. *Beam-scan sensor node: reliable sensing of human interactions in organization.* In Proceedings of the 6th international conference on Networked sensing systems, INSS'09, pages 58–61, Piscataway, NJ, USA, 2009. IEEE Press. (Cited on page 11.)

[Wallinga 2006] Jacco Wallinga, Peter Teunis and Mirjam Kretzschmar. *Using Data on Social Contacts to Estimate Age-specific Transmission Parameters for Respiratory- spread Infectious Agents.* American Journal of Epidemiology, vol. 164, no. 10, pages 936–944, 15 November 2006. (Cited on pages 42 and 108.)

[Wang 2012] Chunyan Wang and Bernardo A. Huberman. *How Random are Online Social Interactions?* ArXiv, page http://arxiv.org/abs/1207.3837v2, 2012. (Cited on page 72.)

[Wasserman 1994] Stanley Wasserman and Katherine Faust. Social Network Analysis: Methods and Applications. Cambridge University Press, 1994. (Cited on page 9.)

[Watts 1998] Duncan J. Watts and Steven H. Strogatz. *Collective dynamics of "small-world" networks.* Nature, vol. 393, no. 6684, pages 440–442, June 1998. (Cited on pages iii and 2.)

[Watts 2007] Duncan J. Watts. *A twenty-first century science.* Nature, vol. 445, no. 7127, page 489, Feb 2007. (Cited on page 3.)

[Williams 1971] Trevor Williams. *An Algebraic Proof of the Threshold Theorem for the General Stochastic Epidemic.* Advances in Applied Probability, vol. 3, no. 2, page 223, 1971. (Cited on page 104.)

[Yoneki 2008] Eiko Yoneki, Pan Hui and Jon Crowcroft. *Wireless Epidemic Spread in Dynamic Human Networks.* Bio-Inspired Computing and Communication, pages 116–132, 2008. (Cited on pages 9, 10, 11, 107 and 115.)

[Zagheni 2008] Emilio Zagheni, Francesco C. Billari, Piero Manfredi, Alessia Melegaro, Joel Mossong and W. John Edmunds. *Using Time-Use Data to Parameterize Models for the Spread of Close-Contact Infectious Diseases.* American Journal of Epidemiology, vol. 168, no. 9, pages 1082–1090, 2008. (Cited on pages 42, 43 and 108.)

[Zaric 2002] Gregory S. Zaric. *Random vs. Nonrandom Mixing in Network Epidemic Models.* Health Care Management Science, vol. 5, pages 147–155, 2002. (Cited on page 107.)

[Zhao 2011] Kun Zhao, Juliette Stehlé, Ginestra Bianconi and Alain Barrat. *Social network dynamics of face-to-face interactions.* Phys. Rev. E, vol. 83, page 056109, May 2011. (Cited on pages 21 and 75.)

Résumé

Les technologies modernes permettent d'avoir des renseignements toujours plus précis sur les interactions entre individus. Dans ce contexte, la collaboration SocioPatterns a permis de développer une infrastructure mesurant, avec une très grande résolution temporelle, la proximité face-à-face d'individus volontaires, portant des badges de radio-identification. Cette infrastructure a été déployée dans divers contextes, tels que des conférences scientifiques, un musée, une école ou encore un service hospitalier. La simple analyse de ces données représente un enjeu majeur pour l'étude de la dynamique humaine et soulève des questions aussi fondamentales que la recherche d'outils et de techniques d'analyse adaptés. Cette thèse présente la caractérisation statistique de la dynamique de proximité physique, mise en relation avec le contexte et les autres métadonnées disponibles, telles que l'âge, le sexe des individus, ou bien la structure de leurs réseaux sociaux virtuels. Si la structure des contacts diffère considérablement selon le contexte, les distributions empiriques des durées des interactions et entre interactions sont très similaires. Un modèle individu-centré, présenté dans cette thèse, propose des règles d'interactions microscopiques simples susceptibles de donner lieu à cette structure macroscopique complexe des temps d'interaction. Enfin, la caractérisation de la dynamique des contacts entre individus constitue une étape cruciale pour comprendre les mécanismes de propagation de maladies telles que la grippe dans une population. Les données de proximité humaine ont permis d'étudier la quantité d'informations nécessaires sur la dynamique des contacts pour la construction de modèles épidémiologiques de contagion. De tels modèles permettent de mieux estimer a priori l'impact de stratégies de santé publique telles que la fermeture de classes et les vaccinations ciblées.

Abstract

Modern technologies allow to access to more and more detailed information on human interactions. In this context, the SocioPatterns collaboration has allowed to develop an infrastructure based on radio-identification devices, that records human proximity patterns at a fine grained resolution, among voluntary individuals. This infrastructure has been deployed in diverse contexts, such as scientific conferences, a museum, a primary school, or a hospital department. The mere analysis of these data represents a high stake for the study of human dynamics and raises fundamental issues such as the need of adequate tools and analysis techniques. This thesis presents the statistical characterization of physical proximity dynamics, put into relation with the context and other available metadata such as the age, the gender of participants or the structure of their virtual social networks. Although contact patterns considerably differ amongst the various contexts, the empirical distributions of interaction durations and of inter-contact times are very similar. An agent-based model, presented in this thesis, suggests simple microscopic interaction rules able to produce the complex macrostructure of interaction durations. In the last place, the characterization of contact dynamics constitutes a determining step for understanding spreading mechanisms of diseases such as the influenza. The human proximity data have allowed to analyze the level of information needed on contact dynamics for the elaboration of epidemiological models of contagion. Such models allow to better estimate the impact of public health strategies, e.g. the closure of school classes and targeted vaccinations.