# Deep dive into social network and economic data: a data driven approach for uncovering temporal ties, human mobility, and socioeconomic correlations

Yannick Léo


Numéro National de Thèse : **2016LYSEN066**

# THESE de DOCTORAT DE L'UNIVERSITE DE LYON

opérée par

**l'Ecole Normale Supérieure de Lyon**

## Ecole Doctoral 512

**Ecole Doctorale en Informatique et Mathématiques de Lyon**

**Spécialité de doctorat : Informatique**
**Discipline : Informatique**

Soutenue publiquement le 16/12/2016, par :

**Yannick LEO**

---

**Deep dive into social network and economic data: a data driven approach for uncovering temporal ties, human mobility, and socioeconomic correlations.**

Immersion dans les réseaux sociaux et les données économiques : une étude des interactions temporelles, de la mobilité humaine et des corrélations socio-économiques à travers le prisme du « big data ».

---

Devant le jury composé de :

| | | | |
|---|---|---|---|
| Magnien Clémence | Dir. Rech. CNRS | LIP6 | Rapporteure |
| Saramaki Jari | Professeur | Aalto University | Rapporteur |
| Cardon Dominique | Professeur | Science Po Paris, Medialab | Examinateur |
| Lambiotte Renaud | Professeur | University of Namur | Examinateur |
| Fleury Eric | Professeur | ENS de Lyon, INRIA, LIP | Directeur de thèse |
| Crespelle Christophe | Maître de Conf. | UCBL Lyon 1, LIP | Superviseur de thèse |
| Karsai Marton | Maître de Conf. | ENS de Lyon, INRIA, LIP | Superviseur de thèse |

# Deep dive into social network and economic data: a data driven approach for uncovering temporal ties, human mobility, and socioeconomic correlations,

A DISSERTATION PRESENTED

BY

YANNICK LEO

DIRECTED

BY

ERIC FLEURY

SUPERVISED

BY

CHRISTOPHE CRESPELLE & MARTON KARSAI

TO

THE LABORATOIRE INFORMATIQUE DU PARALLELISME

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF COMPUTER SCIENCE

ENS DE LYON

LYON, FRANCE

OCTOBER 2013-2016

# Contents

# Résumé en français

Ces dernières années, la quantité de données digitales personnelles enregistrées explose. La facilité de stockage et d'enregistrement de grosses données combiné au développement rapide d'outils qui permettent d'analyser des données utilisateurs prédire le comportement des utilisateurs ouvrent de nombreuses possibilités d'analyses. Les chercheurs peuvent utiliser une approche Big Data afin d'attaquer avec un nouvel angle des hypothèses de longues dates ou mme répondre à de nouvelles questions innovantes. Notamment, les liens entre utilisateurs, impossible à obtenir à grande échelle sans enregistrement numérique, jouent un rôle prépondérant dans l'organisation de notre société moderne. De surcroît, étant une approche novatrice, la méthodologie utilisée tout au long de l'étude Big Data se doit d'tre minutieuse pour garantir la justesse des résultats obtenus.

Cette thèse, composée de 8 chapitres, proposent des études diversifiées se basant toutes sur des grosses données dites Big Data. La pluralité des études menées tend à montrer la diversité des types de résultats qu'il est possible d'obtenir en se basant sur des données de qualité. Plus précisément, ces travaux sont centrés sur les réseaux d'utilisateurs pouvant évoluer au court du temps dans l'espace. Premièrement, des outils sont développés au chapitre 2 et 3 afin de comprendre la justesse des analyses Big Data. Ils permettent de quantifier la différence entre l'objet mesuré et l'objet analysé lors d'une l'étude. Deuxièmement, l'analyse pure et dure de données téléphoniques selon les aspects notamment temporel et spatial ont pour but de mieux comprendre le rôle de paramètres qui impactent le trafic. Troisièmement, les données réelles permettent de réaliser des expériences sur des cas concrets. Par exemple, au chapitre 4, à partir des enregistrements de SMS incluant le temps, le lieu, l'origine et la destination, il est possible de tester de nouveaux protocoles. En plus de comprendre la nature de l'activité, il est donc faisable d'expérimenter de nouvelles idées sur des scénarios existants et récents. Quatrièmement, la richesse des données traitées étant vraiment exceptionnelle, il nous a été possible de proposer une étude sociologique poussée. La combinaison de données téléphoniques et bancaires donne l'accès à de nombreux paramètres individuels comme l'âge, le sexe, l'adresse du logement, l'adresse de travail, l'activité téléphonique mais aussi le salaire et les dépenses.

J'introduis ma thèse par un chapitre qui présente les données téléphoniques et bancaires utilisées dans les travaux de recherche qui composent les parties 1 & 2. Ce chapitre 1 permet d'appréhender non seulement la qualité des données mais aussi son potentiel et ses limites. Les possibilités de questions sont très nombreuses compte tenu de la quantité des dimensions individuelles incluses dans les données partagées par Grandata

Labs à l'échelle nationale du Mexique. Il est montré notamment que la répartition spatiale est cohérente avec la distribution réelle et que la répartition d'âge est logiquement différente dès que l'on considère les clients bancaires. Par ailleurs, une sous-représentation de la proportion de femmes témoigne d'un déséquilibre homme-femme. Dans les études sociologiques, il faut garder ce biais en tte afin de proposer des résultats valides. Il est aussi important de noter les limites, nos résultats sont bien évidemment dépendant de l'espace et de la période temporelle considérée.

Dans la partie 1, uniquement les données téléphoniques sont analysées. Les appels et SMS sont enregistrés dans le temps à la seconde près et dans l'espace via l'antenne de télécommunication. Les contacts entre personnes constituent un réseau social évoluant en fonction du temps. Bien que les contacts téléphoniques ne constituent pas l'unique moyen de communiquer, il s'avère très représentatif du réseau social réel de chacun des utilisateurs. Beaucoup de réseaux dynamiques se représentent par une liste de lien (u,v,t) o u et v sont deux nuds ayant un lien à l'instant t. Un très grand nombre d'études commence par agréger les données en fentres temporelles de taille fixe afin d'obtenir une série de graphes plus facile à analyser. Dans le chapitre 2, nous étudions l'impact du choix de la taille de la fentre temporelle. Nous nous posons la question fondamentale suivante : combien une série de graphes obtenue par l'agrégation temporel est représentative de la série de contacts ponctuels originale ? Nous répondons empiriquement à cette question en montrant qu'il y a un seuil au delà duquel les transitions temporels (ou causalités) ne sont plus préservées. Si on agrège trop la série de contacts en choisissant une fentre trop large, le signal est dit saturé et les propriétés tel que la propagation sont altérées. Nous proposons une méthode automatique qui détermine cette échelle de saturation pour n'importe qu'elle flot de liens que nous validons sur plusieurs données réelles.

La méthodologie s'avère cruciale pour assurer des résultats valides et intéressants. S'assurer que l'objet étudié est proche de l'objet réel en est la clé. En parallèle de l'information temporelle, l'information spatiale fait aussi l'objet de beaucoup d'approximations. Dans la plupart des cas, les localités dans les données téléphoniques sont enregistrées lors de l'appel. La mobilité de l'utilisateur est souvent approximée par sa mobilité lors de son activité. Nous avons, dans le chapitre 3, montré la corrélation entre la mobilité et l'activité téléphonique d'un utilisateur. Ensuite, nous proposons de déterminer la distribution des mouvements réels à partir des localités des appels téléphoniques. Cette étude a une approche intéressante car elle se base sur la théorie de Palm qui permet de faire le lien entre plusieurs distributions.

En enregistrant des liens entre personnes, les données sont le témoin de scénarios réels. A ce titre, en supposant que peu de biais sont introduits dans la mesure du système dynamique, l'objet à étudier se transforme en terrain d'expérimentations. Au lieu de partir de modèles plus ou moins proches de la réalité, les scénarios réels permettent d'obtenir les exactes propriétés du tissu social dynamique comme le phénomène petit monde, les distributions temporelles des contacts, etc... Ainsi, au chapitre 4, nous expérimentons des protocoles de communication avec deux buts majeurs. D'une part, il est important de faire une analyse des données téléphoniques afin de mieux comprendre et prédire lactivité. D'autre part, différents protocoles de communication sont expérimentés à travers le réseau ad hoc constitué par les utilisateurs mobiles. Cette analyse ne permet pas seulement de tester de nouvelles approches mais avant tout de quantifier empiriquement la potentialité du réseau d'utilisateurs au vu d'une grande densité urbaine, d'une mobilité accentuée

pendant les rush hours et des communications le plus souvent très locales. Par exemple, à Mexico City, en s'appuyant sur ces atouts, il est possible de délivrer plus de 50% des SMS. Ce taux de réussite peut atteindre 78% pendant les pics d'activités au moment mme o l'infrastructure des opérateurs est mise à l'épreuve.

La première partie, se focalisant uniquement sur les contacts temporels et spatiaux, développe l'idée selon laquelle une attention très particulière doit tre faite afin d'éviter les fausses conclusions. En faisant un réel travail sur les biais, les données téléphoniques deviennent un terrain de jeu très intéressants. Elles permettent de reproduire des scénarios à grande échelle sur des périodes assez longues. Pour complémenter ces études sur les réseaux dynamiques, nous combinons des données démographiques et bancaires aux enregistrements téléphoniques. En plus, d'avoir le réseau de communications, les attributs individuels concernant des millions d'utilisateurs sont disponibles : l'âge, le sexe, les adresses du domicile et du travail, le salaire, les dépenses et bien entendu le réseau ego centré. La seconde partie présente une étude sociologique assez complète qui propose un nouveau point de vue pour valider ou infirmer d'anciennes hypothèses ou établir de nouveaux résultats.

L'entrée en matière de la seconde partie, au chapitre 5, explique le passage entre les données bancaires brutes et des paramètres basées sur les dépenses qui décrivent la richesse. Nous montrons non seulement l'inégalité de répartition de la richesse mais aussi des dettes, elles suivent toutes deux une loi de Pareto. Les classes sociales, bien que difficiles à définir, sont présentées et analysées. La pyramide des âges inclut aussi le paramètre d'appartenance à une des 9 classes sociales. Nous montrons que les classes plus élevées ont une moyenne d'âge plus élevée. De plus, l'accès aux classes les plus hautes reste très limitée pour les femmes. En effet, seulement un quart des femmes composent la classe la plus riche alors qu'elles représentent la moitié des utilisateurs.

L'inégale répartition de la richesse a une influence directe sur l'établissement des sociétés modernes et de la structure sociale induite. Cependant, l'étude des corrélations entre le réseau social et le statut économique au niveau de l'individu est difficile en raison d'un manque de grosses données disposant de ces dimensions très différentes. Dans notre étude sociologique, nous comblons ce manque avec l'analyse simultanée de données téléphoniques et bancaires pouvant tre relié grâce à un identifiant commun pour des millions d'utilisateurs à Mexico. Au chapitre 6, nous observons une structure très fortement stratifiée. Les personnes sont, en général,bien plus connectées et vivent plus proches avec les relations appartenant à la mme classe sociale qu"avec les autres plus pauvres ou plus riches. Ce phénomène fait notamment apparaître des clubs de riches connectés. Ces résultats sont obtenus par l'utilisation d'outils statistiques mais surtout en comparant des propriétés du graphe réel avec un graphe aléatoire qui garde la distribution de degré (appelé configuration model graph ). Cette comparaison permet de saffranchir des biais liés au degré et quantifie uniquement la tendance qu'ont les gens à se lier en fonction de leur statut social.

La consommation de biens et de services est un élément crucial du bien-tre de chacun. Les différences de pouvoir d'achat est un des facteurs des inégalités socio-économiques. La capacité économique d'un individu restreint sa façon de consommer et ira de pair avec ces habitudes d'achat et donc sa classe sociale. Dans les données bancaires, chaque achat possède un code parmi les centaines de codes de catégorie marchande. L'entreprise American Express propose de regrouper ces catégories en 28 groupes nous permettant d'étudier

les achats des clients à des niveaux différents. Nous montrons qu'au niveau des groupes, la consommation est très différente. Les pauvres se concentrent sur des groupes nécessaires comme la nourriture alors que les riches se permettent de dépenser dans les voitures, les bijoux, etc. De surcroît, nous démontrons que le réseau social a un impact sur la façon de consommer, spécialement sur les groupes de dépenses reliés à l'éducation, au transport et aux hôtels. Dès lors, en considérant simplement les clients qui achètent les produits, il est possible de relier un produit à un âge moyen, une tendance de genre et mme une classe sociale moyenne. En analysant les dépenses des utilisateurs sur les différentes catégories, certaines corrélations positives entre les catégories émergent. A partir des fortes corrélations, nous avons obtenu un graphe des catégories marchandes. L'idée est de mettre en avant l'organisation des dépenses de la société à partir d'étapes claires et justifiées.

Les raisons des inégalités socio-économiques se constatent sur de longues périodes de temps et se retransmettent de génération en génération. Nos données correspondent à une période de 2 ans et permettent de suivre la façon dont les utilisateurs communiquent. Ainsi, pour chaque utilisateur, nous étudions la quantité de relations qui persistent, qui apparaissent ou qui disparaissent. En comparant les étudiants entrant à l'université et un panel aléatoire d'utilisateurs, nous montrons que dans la période universitaire des relations se créent alors que plus généralement, la quantité d'apparitions et de disparitions de relations est constante dans le temps. Nous démontrons aussi que la façon de concentrer les contacts sur très peu de relations ou de disperser son effort de communication sur de nombreuses relations est propre à chaque individu mme si elle est influencée par différents paramètres extérieurs comme le réseau social. Une façon de continuer ce projet serait de comprendre à quel point la façon de créer des relations à l'université impacte le succès (ou le salaire) après l'université. Il semble très intéressant de réussir à faire un premier lien entre les relations naissantes, la façon de communiquer avec les autres et le succès.

Ces travaux montrent à quel point le big data est intrinsèquement multidisciplinaire. Ma thèse, majoritairement dans le domaine informatique, présente donc des résultats applicables aussi bien dans la théorie des graphes, dans les réseaux de télécommunication que dans la sociologie et l'économie. Ces exemples d'études rendent compte de l'importance des discussions entre les domaines et par dessus tout de la complémentarité des approches dans le but de confirmer empiriquement des hypothèses et d'obtenir de nouveaux résultats. Outre l'aspect interdisciplinaire, avoir le contrôle de la méthodologie, comprendre l'usage des utilisateurs, développer de nouveaux systèmes de communications et analyser les corrélations socio-économique à l'échelle du réseau est directement profitable aux entreprises. Les prédictions, les recommandations et plus généralement l'expérience utilisateur sont améliorés par la qualité et la diversité des contributions obtenues dans ce domaine.

# Introduction

In recent decades, three major computer revolutions have changed the world: the personal computer, the graphical browser, and the Internet. They all have had an incontestable impact on how people live, think and interact with each other. As big data provides room for great innovation, people, especially communities in science, industry, and even government, intend to exploit these novel capacities to develop tools to better understand human behavior.

Big data has already changed the way people live, and has done so on many levels [26]. The PageRank algorithm, used by Google, ranks websites for their search engine results, giving people unprecendented access to information. Further, recommendation methods influence the way we extend relationships through social networks, influence one's choice when buying a new phone, or influence who one contacts on dating services. Recorded data gives the opportunity to manage challenges in sport organisation, to listen to music, to have access to education, to travel, to be more efficient by applying email filters and to do online administrative tasks. In addition, big data has an impact on artificial intelligence, medicine, epidemic predictions and virtual brain projects. For these reasons, it helps to increase the longevity of human life. For all these reasons, big data may bring more productivity to work and personal tasks, more connectivity between people and greater human longevity.

Definitions of big data are manifold, as the term has become a label commonly used in science, industry and media. In 2009, Gartner first proposed a definition in [83] that encompasses three "Vs"of big data: Volume, Velocity and Variety. The computational power of computers and clusters can generate enormous amounts of information compared to traditional data. In 2014, 3500 petabytes of data were stored in North America. The number of data sets is huge and is increasing every year. As the McKinsey Global Institute estimates, data volume is growing 40% per year, and will grow 44-fold between 2009 and 2020. More data is transferred through the Internet every second than was stored over the entire Internet just 20 years ago. As it is becoming easy to record, store and transfer data, the velocity of the data flow is continuously increasing. The Internet is generating 2.9 million emails each second, 20 hours of YouTube videos each minute and 50 million tweets every day. The diversity of data sets is rising inexorably too, as companies and institutions collect data from users and citizens, recording actions (surfing, transactions), transactions (bank, online bills, contracts), post contents (tweets, web blog posts, Wikipedia pages), geolocalizations (Google Maps, mobile apps), personal attributes (health, diplomas, wealth), preferences (rates, likes, follows), pictures (Instagram, social profiles), transportation logs (subway, airports, buses), online interactions

(Call Detail Records, emails, social network friendship) and face-to-face contacts (sensors). Interestingly, for human scientists, big data can be very small because the complexity and the richness matter more than the size. Going further, all large datasets are not big. Big data tends to extend the capacity or capability of common methods and current systems. The veracity and value of the data are important parameters that allow researchers to answer relevant questions. All in all, volume, velocity, variety, value and veracity are the five "Vs"of rich big data.

Yet, as usual, many issues emerge from big data that we call "the big data problems ". People do not ask to have all their actions, movements and interactions logged. Data is registered and sold without any well-understood agreement between companies and users. The personal intrusion is becoming the major big data issue especially when it is used to make political decisions rather than improving services. As a consequence, governments have become more reliant on computers to control society, and criminals more cunning via digital means. Therefore, people worry more and more about what personal information is stored on external servers. Moreover, it also serves to remind us that the relationships between our values and big data are not always clear or distinct. Unfortunately, it is still very difficult to understand how much the big data problems affect our values, goals and quality of life. In this context, researchers, designers and computer scientists should have a predominant role in taking part in the discussion and helping to shape the future.

After thousands of years of experimental science, hundreds of years of theoretical science and decades of computational science, research is facing a new area called data-intensive scientific discovery. The new challenge is to collaborate in order to manage all the available data and propose solid validation processes of findings. For many researchers, getting access to big data is not common. As big data is usually collected by companies, agreements between research teams and companies with millions of users have to be reached to provide access to rich and modern large data. The challenge of creating new platforms to exchange and share data sets never has been so pressing. Dealing with data sets is a great opportunity for many research fields. Big data helps geographers to study human mobility, sociologists to validate deep rooted hypotheses on social structure, medicine to anticipate spreading of diseases, economists to understand consumption patterns and inequalities, historians to analyze data from archives more efficiently, biologists to explore genetics and the brain, physicists to model accurately spreading processes, mathematicians to apply new statistical tools and computer scientists to compare their model to reality. As evidence, large scale data is spurred by its inherent interdisciplinary aspects. Although the data-driven approach has, at some points, overtaken the classical approach, the two approaches remain complementary.

However, the methodology of a big data study differs in many ways from survey studies. Even if one can study specific data sets, it is not yet possible to get deep answers as in an interview process with a sociologist. Data scientists have to ask questions that fit with the studied data set. Many dimensions can be provided for each user like age, gender, wealth and purchase patterns. In many studies, well-known correlations can give the possibility of inferring certain parameters from others. Besides personal attributes, interactions between users are also recorded such as calls, short messages, emails and online messages. These interactions constitute social structure have a major role in the organization of society. In comparison

to interviews or personal surveys, big data provides the tremendous advantage of having access to huge human networks. Considering a network, which corresponds to the global structure induced by interactions, one can use graph theory to cluster a set of nodes into groups, to predict the spread of information or to point out central nodes. After the collection of interactions or contacts, one can build the network and directly apply the convenient graph theory parameters and objects such as degree, clustering coefficient, path, connectivity, flow and modularity parameters. One can compare to other models and detect specific properties such as the small-world effect and power law degree distribution. Usually, human contacts are approximated as an aggregated graph for which there exists a rich toolbox. Graph theory is a complete field of computer science that can be used in the analysis of structured data like mobile data.

In many cases, the edges are not continuously active. As an example, in networks of communication via e-mail, text messages, or phone calls, edges represent sequences of instantaneous or practically instantaneous contacts whereas in some other cases like face-to-face contacts, edges are active for non-negligible periods of time. Therefore, in many studies, the temporarily of interactions is considered, the object is called temporal network. The temporal dimension greatly improves the analysis as it contains the causal information that we completely miss on the aggregated network. Interactions not only vary in time but also in space. The understanding of human mobility and of the mobility of interactions is the key feature to understand epidemics or migrations. The temporal and spatial dimensions of social interactions combined with personal attributes open up many novel interesting questions.

Following millions of mobile users at an individual level is a great opportunity in order to understand human mobility. For example, can we use Call Detail Records in order to understand the spatial and temporal organization of a city? How much can we take advantage of the density and mobility of people in order to design and experiment efficient communication protocols?

Adding individual demographic and bank information to the communication network gives us the possibility of using a data-driven approach for answering long-lasting questions concerning the social stratification and more generally the role of the social network on the social inequalities. From bank datasets and CDRs, can we quantify the social stratification of the society? How much does the social status determine relations? What is the role of the university period on the social stratification? Are there specific purchase patterns for each social classes?

Temporal contacts contain the temporal information of the network. For convenience, many studies start by choosing a specific temporal window in which all the contacts are approximated to happen at the same time. In this context, it is important to quantify how much the studied object represents faithfully the real object. In other words, can we quantify the amount of temporal information lost according to the chosen time scale of the study?

# Outline

Research carried out for this PhD thesis has relied on large scale communication and bank data sets, and has lead to the formulation of novel and interesting questions. It has also provided the opportunity to explore the limit and capacity of the data-driven approach. As all chapters rely on detailed communication and bank data sets, the thesis begins, in Chapter 1, with the presentation of these data sets. For all chapters, except for Chapter 2, we consider the data set presented in Chapter 1.

Subsequently, the thesis is organized in two parts. The first part concerns data analysis at a country and dense-city level, as well as methodology improvements and performance evaluation. In Chapter 2, we analyze the impact of coarse-grained temporal analysis on the loss of information. Starting from fine-grained data sets, one can quantify the impact of degrading the time granularity in two ways: how much the temporal granularity of a contact series impacts the results, and given a time scale how we can estimate the loss of temporal information.

Regarding the temporal aspect of the following data sets, human mobility is a major component of my synthesis. Yet, in the communication data sets, we only have the locations during communication events. In Chapter 3, we investigate the gap between phone event locations and general mobility of users in order to understand the biases and correlations when one considers phone event locations as a faithful representation of user mobility. Can we infer precisely properties of human mobility from mobile events, and is there a correlation between human mobility and human phone usage? Apart from this mobility study, we conduct a large scale communication analysis of classical parameters such as the duration, the temporal activity and the inter-arrival time of communication events at the base station level.

A consequence of the temporal and spatial mobile user activity is an evolving DTN structure constituted by phone users and their emerging interactions. From the communication data sets, we evaluate the potentiality of density and mobility of users. In Chapter 4, using the communication data, we design a data-driven experiment suitable for big cities, in this case Mexico City. A spatial and temporal analysis provides insight about the activity patterns of a dense city. Then, innovative protocols are defined and tested under real conditions in order to estimate interesting properties like the density of mobile phones, the mobility of phone users and the locality of the communications.

In the second part of this thesis, a bank data set is combined with the communication network to study correlations between social stratification and the socioeconomic status of individuals. Discussions with economists and sociologists helped us realize that a data-driven approach may validate and bring new in-

sights to deeply rooted ideas. As there are more than 200 features for each user in the bank data set, we extract a few representative ones that appear interesting, to define social class and analyze inequalities over gender, age and social class in Chapter 5.

In Chapter 6, we study social structure and its organization. The hypothesis of social stratification, of the existence of rich clubs and spatial segregation are discussed. This data-driven approach validates certain hypotheses and quantifies accurately the correlations at the network level considering personal demographics, as well as geographic and socioeconomic parameters.

In Chapter 7, we take into account the consumption dimension of bank clients and extend the study of social class and social structure. Can one associate typical consumption patterns to people and to their peers belonging to the same or different socioeconomic classes, and if so how much do such patterns vary between individuals of different classes? Is there a causal link between position within the network and consumption? If so, does the impact of the network position depend on the type of good or service being consumed? Can one draw relations between commonly purchased goods or services in order to better understand individual consumption behavior? By answering these questions, Chapter 7 develops an overall picture of consumption patterns.

In Chapter 8, we expose the validations of recent findings and present the first findings of a more general study that makes the connection between the way a student socializes while entering university and his or her success (social status) after graduating. We propose, in the first part of the chapter, to comprehend how the ego network is evolving over a period of 21 months in terms of contact turnover, intensity of contacts and global communication effort. Finally, we discuss the benefits and the limits of a data-driven approach and conclude the thesis.

# 1

# Communication & Bank Datasets

In this thesis, as we are working with communication and bank data sets, several features are considered to analyze human behavior. We only focus on human attributes and interactions, and do not explore animal, machine or organization behavior. Personal demographic attributes such as age and gender, attached to our data sets, are part of the demographic features. User demographic parameters enable studies on age differences, gender inequalities or user profiling. The role of the age and gender in our society can be discussed within a data-driven approach. Furthermore, the spatial coordinates of the user are registered in communication logs from the coordinates of the attached antenna when a user is making a call and in bank trace from the postal code linked to each user. In mobile data, locations are registered during specific events such as calls. From these information, we have access to the favorite places of the user and identify the home and work address. Our understanding of how individual mobility patterns shape and impact the social network is limited, but is essential for a deeper understanding of network dynamics and evolution. This question is largely unexplored, partly due to the difficulty in obtaining large-scale society-wide data that simultaneously capture the dynamical information on individual movements and social interactions.

Whereas there is no doubt about the richness of demographic and geographic parameters, analysis should be embellished with other given or inferred personal parameters. Here, we consider fine-grained economic parameters that usually are inferred from aggregated governmental data. Economic parameters are one of the key features to assign to each individuals a social status and start building a sociological analysis. In this thesis, we have access to a bank dataset at the client purchase level. In some studies, from aggregate data at a region level, it may be possible to infer economic parameters like in [52] but the impact of the approximation is hard to quantify. On the contrary, the bank trace opens up possibilities to answer to difficult economical

and sociological questions at the individual level. We propose to develop some of them in part II.

From our data sets, many aspects are provided with some approximations that we have to take into account. With huge data sets and fine-grained measurements, statisticians and computer scientists note that there is an increased risk of false discoveries [55]. When one gets some locations, one has to ask whether the location samples in the trace are representing faithfully the mobility of egos. Furthermore, the boundary of the considered set of users we explore is also a limit. A company may only have clients from a limited region of the world like a set of countries, a country, a city or even a more located area. Whereas studies at each level are building the global knowledge, the generalization is not straight. In this thesis, biases are presented and all steps are explained clearly.

Demographic, geographic and economic features all combined allow us to answer to a large amount of questions concerning human behavior. As a consequence of city organizations, small-world effect and homophily, the network induced by communications between people is well-known to be structured. However, to have access to the underlying social structure, interactions between people have to be part of the trace. Interactions can be modeled as graphs of vertices coupled by edges. Some real interactions such as face-to-face contacts can be captured by RFID sensors and virtual contacts directly measured from web like emails, social network friendships, instant messaging or voice calls. In our case, we consider communication interactions such as calls and SMS. Even if we miss or have some false relations [160], the communication contacts have the advantage to temporally be a good approximation of the real network compared to social network ties that are often kept even if there are no more interactions as Facebook friendships. A recent study [54] demonstrated that real social ties can be effectively mapped from mobile call interactions with precision up to $95\%$. Recently, large mobile data are the purpose of many studies as explained in this recent survey [13]. But again, results obtained on the measured network may not be extended to real system. In one hand, we face the impossibility of having access to all the human interactions and in the other hand, it is difficult to quantify the gap between the measured social network and the real social network. Therefore, as all the results are part of a specific context of study, all the pieces of the context of our experiments are explained.

The communication and bank data sets are the base of our studies. In this chapter, we introduce and build clean data sets in order to prepare clean inputs for our studies. The access to the communication and bank data sets at the country level of Mexico is provided by Grandata Labs (a company that integrates first-party and telco partner data to understand key market trends and predict customer behavior). Out of this thesis, some works have already been recently published from these data sets [20, 97, 114, 136]. The quality of the data sets we have access to is quite rare. As data sets are coming from national companies, the number of users in the Call Detail Records (CDRs) and the bank trace are very large (several millions). In addition, there is the possibility of joining communication and bank clients with the insurance of the anonymization (no direct identification by the phone number or the name). It provides a unique access to a large network of nodes for which we have individual demographic, geographic and economic parameters.

## 1.1 CALL DETAIL RECORDS - DS1

Communication data DS1 used in our study records the temporal sequence of 7,945,240,548 call and SMS interactions of 12,317,219 clients of a single mobile phone operator for 22 months in Mexico: from January 2014 to October 2015. As we not only have the client ID but also the caller-callee IDs, we reach 111,719,360 users that correspond to most of the active phone users in Mexico. It is important to notice that IDs are anonymized. In order to make individual identification of any person impossible from the data, phone numbers of mobile clients are hashed and salted by the mobile company and names are removed from the trace. Each call detailed record (CDR) contains the time, unique caller and callee IDs, the direction and duration of the interaction, and the cell tower location of the client(s) involved in the interaction. Other mobile phone users, who are not clients of the actual provider also appear in the dataset with unique IDs. Using this dataset we constructed a large social network where nodes are users (whether clients or not of the actual provider), while links are drawn between them if they interacted (via call or SMS) at least once during the observation period.

In order to filter out call services and other non-human actors from the social network, after construction we recursively removed all nodes (and connected links) who appeared with either in-degree $k_{in} = 0$ or out-degree $k_{out} = 0$. We repeated this procedure recursively until we received a network where each user had $k_{in}, k_{out} > 0$, i.e. made at least one outgoing and received at least one incoming communication events during the nearly two years of observation. After construction and filtering the network remained with 82,453,814 users connected by 1,002,833,289 links, which are considered to be undirected after this point.

As the communication network is centered on the telco company clients for whom all the calls and SMS are recorded, a part of the complete social network is missing. The trace encompasses both the links between two clients and the links between non-clients and clients. Yet, it does not contain connections between non-clients. Consequently, if we need the list of neighbors, we focus only on telco clients for which we have access to the exhaustive list of calls and SMS. Moreover, it is certain that a part of the users have more than one mobile phone or change their phone number during the experiment. Some personal, business and mobile phone numbers are merged according to the bank information to reduce this bias but it is still unclear how much it impacts the results. In the coming discussion, we neglect this bias and treat every unique hashed phone ID as an individual.

## 1.2 CREDIT DATA - DS2

Besides the Call Detail Records (CDR), an access to a bank data sets at the individual level was provided. Information such as the income, the salary and the total debt are given each month whereas purchases are daily recorded. Beyond this information, individual features like the gender and the age are also given.

In some of the studies, we use individual socioeconomic indicators which are estimated from a datum provided by a single bank. These data records financial details of 6,002,192 of people assigned with unique anonymized identifiers over 8 months from November 2014 to June 2015. The data provides time varying

customer variables as the amount and type of their daily debit/credit card purchases, their monthly loan measures, and user personal attributes as their billing postal code (zip code), their age, and gender. In addition for a subset of clients the records of monthly salary (38.9% of users) and income (62.5% of users) defined as the sum of their salaries and any incoming bank transactions are given.

All the users are clients of a single bank. Considering this large amount of bank clients, it is important to note that these clients are quite representative of the society we consider. It is fair to note that people in the bank trace have a bank account and mostly have money. Yet, this bank proposes a free access to bank accounts. Every Mexican citizen can open a bank account to this bank without any fee. Therefore, not only rich people have an easy access to bank accounts.

In addition, many parameters related to the social status of a person are missing such as the family characteristics. More generally, there is no direct access to the type of interactions between people. By choice, we do not try to predict these connections as it should, again, introduce a strong unquantifiable bias for the sociological study. The wealth of an ego is computed according to the spendings. Unfortunately, the traces do not encompass any information about the taxes the user is paying or properties he or she owns.

### 1.3 PURCHASE DATASET - DS3

To study consumption behavior we used purchase sequences recording the time, amount, merchant category code of each purchase event of each individual during the observation period of $8$ months. Purchase events are linked to one of the 281 merchant category codes (mcc) indicating the type of the actual purchase, like fast food restaurants, airlines, gas stations, etc. Due to the large number of categories in this case we decided to group mccs by their types into $28$ purchase category groups (PCGs) using the categorization proposed in [1]. After analyzing each purchase groups $11$ of them appeared with extremely low activity representing less than $0.3\%$ (combined) of the total amount of purchases, thus we decided to remove them from our analysis and use only the remaining $K_{17}$ set of $17$ groups. Note that the group named *Service Providers* ($k_1$ with mcc $24$) plays a particular role as it corresponds to cash retrievals and money transfers and it represents around $70\%$ of the total amount of purchases. As this group dominates over other ones, and since we have no further information how the withdrawn cash was spent, we analyze this group $k_1$ separately from the other $K_{2\text{-}17} = K_{17} \backslash \{k_1\}$ set of groups.

This dataset collects data about the age and gender of individuals together with their purchase sequence. To receive a set of active users we extract a corpus of 4,784,745 people that were active at least two months during the observation period. Then for each ego $u$, purchase distribution vector $PV(u) = [r(u, c_1), ..., r(u, c_{281})]$ is assigned and defined as :

$$r(u, c_i) = \frac{m_u^{c_i}}{\sum_{c_i} m_u^{c_i}} \tag{1.1}$$

From the amounts of money spent $m_u^{c_i}$ on each merchant category $c_i$. This vector assigns $r(u, c_i)$ the fraction of money spent by user $u$ on a merchant category $c_i$ during the observation period. We excluded purchases corresponding to cash retrievals and money transfers, which would dominate our measures otherwise. A

minor fraction of purchases are not linked to valid mccs, thus we excluded them from our calculations as well. This way, DS3 collects 3,680,652 individuals all assigned with a purchasing vector $PV(u)$ and demographic details.

## 1.4 COMBINED DATASET - DS4

As the phone number IDs linked to CDR and bank data are hashed and salted in the same way by mobile and bank companies, it is possible to unify the bank and mobile data sets and to create a combined dataset DS4. A subset of IDs of the anonymized bank and mobile phone customers were matched. The matching, data hashing, and anonymization procedure was carried out through direct communication between the two providers (bank and mobile provider) without the involvement of the scientific partner. After this procedure only anonymized hashed IDs were shared disallowing the direct identification of individuals in any of the datasets.

This combined dataset allowed us to simultaneously observe the social structure and estimated economic status of the connected individuals. The combined dataset contained 999,456 IDs, which appeared in both corpora. However, for the purpose of the study only the largest connected component of this graph containing IDs valid in both data corpora is considered. This way, a connected social graph of 992,538 people connected by 1,960,242 links is obtained. For all of the nodes, communication events and detailed bank records are available.

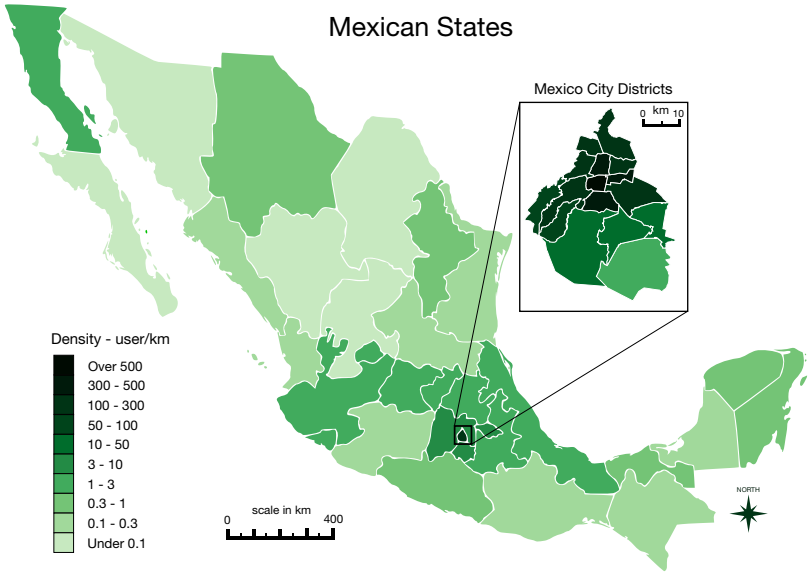



Figure 1.4.1: State level population distribution of egos in DS4 based on their zip locations. Inset depicts a zoom on Mexico D.C. district.

In Fig. 1.4.1, the density population in the combined data sets DS4 is showed. Interestingly, at the state level, the density distribution is correlated with a positive correlation value of 0.81 with the real density distribution [3]. Even if the context is only limited to the country of Mexico, the nodes are well distributed all over the country. Therefore, our results are valid at the population level of Mexico without any spatial distribution biases.

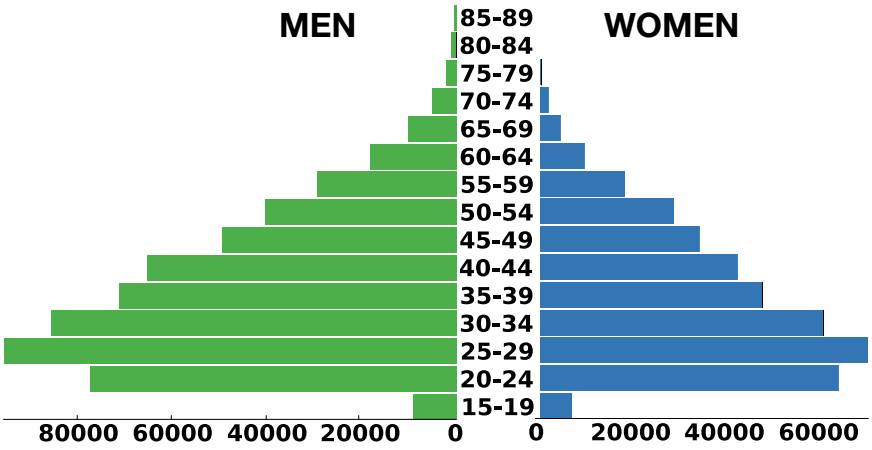



Figure 1.4.2: Age pyramids for men and women in DS4.

To demonstrate some information about the demographic structure of the observed population, the age pyramids of DS4 users are represented in Fig. 1.4.2. The blue bars correspond to women and the green bars to men. Whereas, in reality, there are more females (51.3%) than males (48.7%) in Mexico, there are 54% of males in DS4 suggesting that males have a greater access to bank and mobile accounts. It is important to take into account the gap between the sample of DS4 and real population. Yet, the gender difference reveals inequalities between males and females suggesting the unbalanced organization of the society. Without any surprise, the age distribution is very sparse for children. Managing money and being a client is reserved for adults. A deeper analysis in chapter 5 will point out the same age pyramids with more parameters such as socioeconomic status. One needs to keep in mind these inequalities to better understand the context of the findings. It is reasonable to only consider social status of adults and children, yet, the gender inequity has to be considered when conclusions are made. For example, when the percentage of women in each social class is quantified (like in chapter 5), it should be normalized according to real gender ratios in order to obtain the real numbers. Even more, there is maybe a slight difference between the women that have a bank account and the women in general, this part of the bias is very difficult to precisely quantify.

## 1.5 LOCATION DATASET - DS5

In order to infer the locations of customers we use two types of location data from DS4. The first location parameter is the zip code of billing address of bank customers (also called zip location) like in the study [82]. The second type of parameters is coming from the communication dataset. Using geo-localized mobile communication events, it is possible to estimate the work and home locations for a set of users . Home (resp. work) locations are determined by looking at the most frequented locations during nights and weekends (resp. during daylight at working days) [35]. From the total 992,538 individuals, 990,173 have correct zip codes, and 94,355 have detectable home and work locations (with at least 10 appearances at each location). Each method has some advantages and disadvantages. While frequency dependent locations are more precise, they strongly depend on the activity and regularity of users in terms of mobility. On the other hand, zip codes provide coarse-grained information about the location of individuals but they are assumed to be more reliable due to reporting constraints to the bank and because they do not depend on the call activity of individuals.

It is important to note that work locations can refer for unemployed people to the favorite location of the day. Moreover, if an ego has several home addresses and several work addresses, our algorithm will take the most frequented one. As the home and work locations are computed at the antenna level, they are more precise in large city such as Mexico City where antennas are planted denser rather than in countryside. Yet, regions that have no coverage are very rare in Mexico, mistakes mainly come from people that do not follow the weekly pattern of sleeping at home during the night and working during weekdays.

## 1.6 STUDENT DATASET - DS6

For this last data set, we are looking forward to get temporal communications of freshmen students entering the university. DS1 contains 12,317,219 clients of a single mobile phone operator for 22 months in Mexico: from January 2014 to October 2015. In parallel, in DS2, a student flag is provided by the bank revealing rather if the account is a student account or not. Then, by neglecting the students that repeat or ship grades, we assume that all the freshmen are 18 years old and only keep 18-year-old students. Finally, we manage to obtain 1675 freshmen students with the whole ego network. In Fig. 1.6.1a, we see an example of types of nodes and links that the dataset contains. For instance, we do not have links between two nodes that are not clients. Yet, for this 22-month period, we know all the contacts of clients including nodes that are not clients of the telecommunication company.

In parallel, we build a reference sample of nodes that have the same global activity distribution. To do so, we picked 1675 clients that have the same degree distribution. The important properties that we study in this chapter are the evolution and persistence of the communication effort. In Fig. 1.6.1b and c, the temporal degree $< D >$ defined as the average number of relations per month and the temporal activity $< A >$ defined as the averaged total number of contacts per month are shown. Even if the total distribution of degree and activity of the students and the random sample are equal for the whole period, these distributions

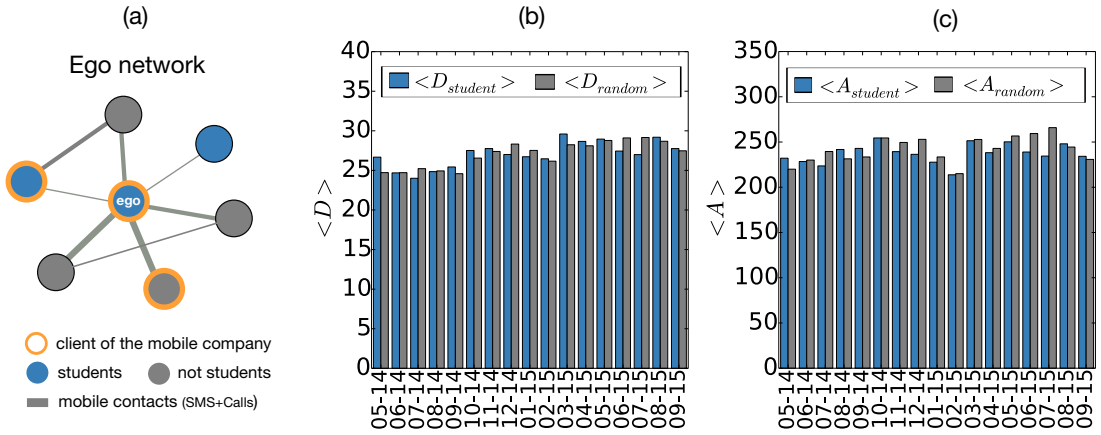



Figure 1.6.1: **Data presentation and temporal activity (a)** An ego network of a single client student from the 1675 that the dataset contains. The student is linked to a set of nodes by phone communications (SMS or calls). The neighborhood is exhaustive as it does not only contain students and clients but also non clients and non students that have at least one contact with the ego. The averaged **(b)** temporal degree (number of different contacts) and **(c)** the temporal activity (total amount of contacts) per month of the student sample (in blue) and a random sample (in grey).

may differ for a single month.

## 1.7 Discussion

As depicted in Fig. 1.7.1, we can point out several properties of our datasets that show their potential to investigate several scientific problems. First, the number and the quality of personal attributes such as gender, age and socioeconomic parameters are substantial for a sociological study as for many questions, the answer will change according to these two dimensions. Second, the evolution in time and space of each individual is the key to study urban areas and to design models for the mobility of individuals and crowds, primarily important in the design of smart cities. For example, the spatio-temporal evolution have an impact on transportation and communication services. Third, individual actions such as purchases make the link between individuals and short-term activity (shopping, practice sport, ...) and long-term attributes (social status or preferences). Fourth, the network position is fundamental. The social structure coming from social interactions is the key to understand human behavior. I pointed out many correlations at the network level revealing that personal traits depend greatly on network position. Consequently, as the homophily of ties is present, the personal traits of an ego can be inferred by the analysis of his or her network position. Finally, in practice, the context of the data is important in order to get true findings. For example, having a large representative sample of users and making an analysis over a long period improve a lot the quality of the study.

Therefore, the quality of the data sets open up many possibilities in order to push interesting questions in four disciplines such as graph theory, communication network, sociology and economy. The richness of

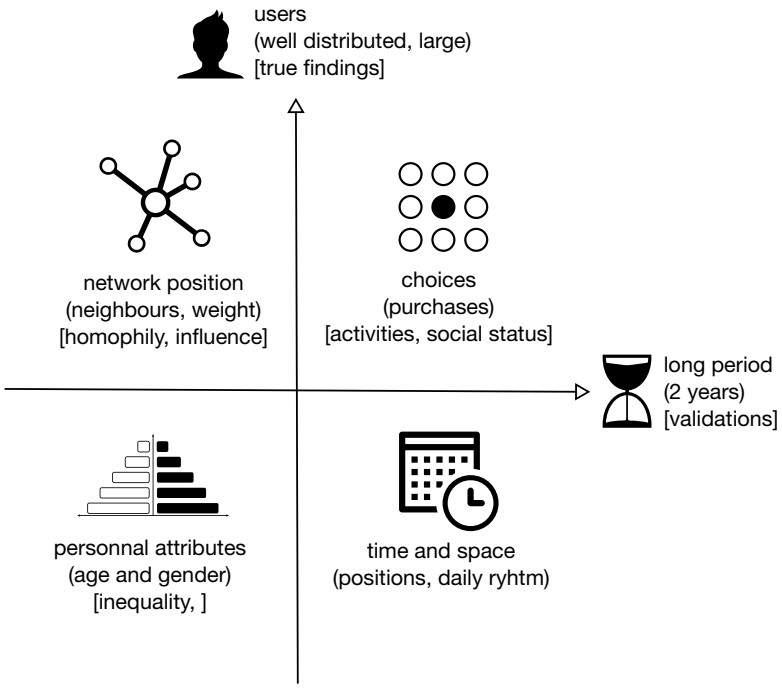



Figure 1.7.1: Schema of big data of this manuscript

these data sets is coming from their diversity. Having the combination of geo-locations, personal attributes, economic parameters and the communication network at the individual level is very rare especially for millions of users and for several months or years. In this chapter, biases have been discussed in order to understand better the ground truth of our findings. When a conclusion is made, it is important to take care about the male gender overestimation in DS4 compare to real ratio.

In the chapters that follow, we organize our didactic around well formed scientific questions, and explain the data sets build from DS1-6 what that are used to investigate the answers. For example, in chapter 3 only the nationwide communication trace is considered and in chapter 4 the communication dataset is filtered out and centered on a very dense area : Mexico City. In the part II, we mainly use DS4 like in chapter 5 and 6. Finally, in chapter 7, as we focus on purchase types, we do not consider bank clients that only use cash and in chapter 8, we only target the 18-year-old students entering the university. One can notice that we always map a research question to a very specific type of dataset. Yet, because of the complexity of these data sets, the heterogeneity of the questions we can ask is significant.

# Part I

# Interactions and mobility patterns in dynamic social networks: methodology and modelling

# Part I: Introduction

Communication is the process of sending and receiving information among people. Humans communicate with others not only by face-to-face, but also by giving information via the Internet, phones and printed products such as books and newspapers. Many people believe that the significance of communication is like the importance of breathing. It is no doubt that communication plays a vital role in human life. As the foundation of all human relationship, communication is the base of organizational structure and change [127].

The need of communicating, especially in dense areas, is increasing. Every day, millions of SMS are sent in a large city like Mexico City. Traditional SMS is challenged by alternative messaging services such as Facebook Messenger, WhatsApp [32], Tango, Skype and Viber by using data connection and/or wireless hot spots [151]. One can also use P2P applications to enable the connection of smartphones via Bluetooth or Wi-Fi without an Internet connection. In this case, people have to constitute a dense network to reach the connectivity. Though it was not designed in purpose, FireChat was used as a communication tool in some civil protests. Nevertheless, SMS is still a growing market and remains a very popular service over cellular networks since 82% of mobile users are sending SMS in DS1. The SMS is well-known and well-used in both developed and developing countries.

The constant evolution of mobile technologies and usage of cellular networks tends to change deeply. The analysis of phone calls from real logs is thus fundamental in order to understand them and adapt the protocols and infrastructures especially when the traffic is challenging. The load is varying in time and space, high activities are concentrated during rush hours and in urban areas such as dense cities. Urban performance currently depends not only on a city's endowment of hard infrastructure, but also, and increasingly so, on the availability and quality of knowledge communication and social infrastructure. The latter form of capital is decisive for urban competitiveness. Against this background, the concept of the smart city has recently been introduced as a strategic device to encompass modern urban production factors in a common framework and, in particular, to highlight the importance of Information and Communication Technologies (ICTs) in the last 20 years for enhancing the competitive profile of a city. People are constantly moving from an area to another, there are interacting with each other. The human daily pulse, especially in dense areas, reveals a dense mobile network that smart cities need to understand. Smart cities may propose better services and organization if they take advantage of users mobility, activity and density.

All in all, the possibility to obtain insights from CDRs has never be so important. As the amount of applications is tremendous, the methodology is fundamental. In this part, we make a longitudinal study of CDRs that represent a 1-year nationwide data set to better understand the mobile activity resolved in time and space.

In chapter 2, we are concerned with the impact of the time scale chosen to measure and study a link stream. We address the fundamental question of knowing whether a series of graphs formed using a given time scale faithfully describes the original link stream. We define an automatic tool based on graph theory and path analysis for quantifying the loss of temporal information when one choose a specific studied time scale.

In chapter 3, we analyze a very large mobile data sets over 12 months in Mexico. It contains 8 millions users and 5 billions of call events. Our first contribution is the study call duration and inter-arrival time parameters. Then, we assess user movements between consecutive calls (switching from a station to another one). Based on Palm Calculus theory, our study suggests that user mobility is dependent on user activity. Furthermore, we show properties of the inter-call mobility by making a analysis of the call distribution.

In chapter 4, part of a smart city project, a global performance evaluation of the human DTN network defined by mobile users for delivering SMS is presented. Four different protocols are experimented in the dense area of Mexico City. Results are shown according to several dimensions such as time, space, density, mobility, activity, storage cost in order to deeply catch the mobility and time-reacting potentiality of the human dynamics of the system. Such key characterizations allow us to answer the question: is it possible to transmit SMS using phones as relay in a large city such as Mexico City?

# Part I: Related Work

Whereas it is a great advantage to understand better the mobile activity, the studies of nationwide cellular networks are quite rare [165]. First, it may help to limit the congestions and anticipate high peak activities. The overload during rush hours or specific events that induces high traffic is a great problem, several solutions have been proposed [70, 115]. The amount of mobile phone traffic has an overriding impact on the quality of services. The understanding of time evolution and spatial arrangement of the activity, studied in [62, 67, 90], helps to enhance the network infrastructure and its capacity.

As an example, the traffic analysis brings around a set of tools to detect specific local events and anomalies [25, 46, 47] that commonly induce overload [120]. Predict and adapt protocols to respond to high activity periods is a substantial benefit [70]. Other parameters such as the inter-arrival time also has its importance to understand the daily activity [28, 168]. The traffic is the result of a causal chain where the way users communicate to each others is the starting point. From CDRs, it is possible to understand better the human behavior and to predict the traffic.

Yet, some events are unpredictable, such as natural disasters or riots. Both global and local events can be uncertain. Many studies propose general or specific methods to detect such events [57, 152, 164]. During these overloaded periods while the network is saturated, some messages are not delivered unless they use alternative ways of communicating [32], or more decentralized applications such as the BitTorrent's P2P Encrypted Messaging App where messages are directly sent from one phone to another without being routed through or stored on any servers.

However, mobile data are even richer and applications are broad. It is possible to characterize users from their spatio-temporal activity. For instance, detecting personal events leads to the identification of his or her religion, soccer team or music bands [118]. The paper [109] defines categories of mobile call profiles and [38] makes the links between user phone usage and personal behavior.

In addition, the activity is resolved in time and space. Recent studies investigate spatial individual mobility [25, 37, 113, 123, 139, 168]. In [67], the human mobility is shown not to be random like modeled with random walk or Levy flights. This idea, comforted in [141], is intuitive as soon as one understands that each user usually spends the majority of his or her time in few favorite places [9, 35].

At a city level, the mobility of users in contacts to each other induce a DTN network with certain capacities [45]. The studies [28, 74] try to understand better the potentiality of the network in some small connected areas like scientific conferences. For example, in [68], they show that mobility increases the

capacity of the DTN network.

The understanding of the DTN network induced by phone users have many applications especially for building connected smart cities [71, 76]. More generally, the knowledge coming from mobile structure and activity implies great improvements of recommendation [166] and predictions of demographic attributes [20], locations [163] or even personality of the user [22, 31].

However, working on measured link streams induces a gap between the studied object and the real object especially when you miss a part of the network [85]. Furthermore, the time in seconds contains the causal information that we may miss if one make a study at the day level. In this context, it seems to be difficult to catch the true of the findings when one chooses an incoherent time scale regarding the real system. It is paradoxical to note that, while the question of the influence of time scale choice on the properties of the formed graph series is largely ignored in most of the studies on dynamic networks [11, 19, 25, 27, 30, 43, 53, 64, 72, 73, 80, 84, 86, 91, 93, 98, 108, 110, 116, 121, 122, 128, 133, 133, 135, 142, 144, 148, 149, 153, 155], this question actually already received a lot of specific attention [23, 33, 60, 79, 129, 133, 142, 147]. Furthermore, in the paper [28], they show how the inter-arrival distribution impacts on the algorithms and suggests to have take care of the distribution of the arrival times of interactions in order to obtain these results.

In this part, we propose a method that helps to control the loss of information while processing link streams like a call and SMS events. Besides, we propose a nationwide study that push the understanding of phone usage and show properties of the DTN network induced by the users.

# 2

# Non-Altering Time Scales for Aggregation of Dynamic Networks into Series of Graphs

Many dynamic networks coming from real-world contexts are *link streams*, i.e. a finite collection of triplets $(u, v, t)$ where $u$ and $v$ are two nodes having a link between them at time $t$. A very large number of studies on these objects start by aggregating the data in disjoint time windows of length $\Delta$ in order to obtain a series of graphs on which are made all subsequent analyses. Here we are concerned with the impact of the chosen $\Delta$ on the obtained graph series. We address the fundamental question of knowing whether a series of graphs formed using a given $\Delta$ faithfully describes the original link stream. We answer the question by showing that such dynamic networks exhibit a threshold for $\Delta$, which we call the *saturation scale*, beyond which the properties of propagation of the link stream are altered, while they are mostly preserved before. We design an automatic method to determine the saturation scale of any link stream, which we apply and validate on several real-world datasets.

## 2.1 Introduction

Many real world dynamic networks are naturally given in the form of a finite collection $\mathcal{L}$ of triplets $(u, v, t)$, which we call a *link stream*, where $u, v \in V$ are two nodes of the network and $t$ is a timestamp[1], with the meaning that nodes $u$ and $v$ have a link between them at time $t$. Depending on the context, these

[1]Time can be continuous or discrete. The method we design works in both frameworks. Though, the sample datasets on which we illustrate it all use discrete timestamps.

links can represent physical contacts between individuals, exchanges of emails between people, commercial interactions between companies, etc. When one wants to study such dynamic networks, a very common approach [19, 25, 30, 43, 53, 64, 73, 80, 91, 93, 98, 110, 121, 122, 135, 148, 149, 153, 155] is to transform them into series of graphs. The process used to do so is called *aggregation*. It consists of choosing a time window $[a, b] \subseteq [0, T]$ in the initial series, where $T$ is the length of the period of study, and forming the graph $G_{[a,b]}$ with all edges $u, v$ such that there exists a triplet $(u, v, t) \in \mathcal{L}$ with $t \in [a, b]$. Doing so for a collection of windows that covers the entire period of study, one obtains a representation of the dynamic network as a graph series: the graphs formed for each window, called *snapshots*. Very often, as in this chapter, the windows are disjoint and all have the same length, but in some studies, they may also overlap [11, 27, 84, 108, 133, 144] or have different lengths [128, 142] or all start at the beginning of the period of study [72, 86, 116, 133]. In all cases, once the series is obtained, all analyses are conducted on it instead of the original link stream.

There are two main reasons to use aggregation for studying dynamic networks. First, in many cases, it does not make sense to study the network at the scale of the time resolution of the timestamps of the given link stream. For example, in an email dataset, the timestamps of the events (sending of email) are often given with a 1-second resolution. However, studying the dynamic network at this time scale does not give a general and comprehensive view of its organization, like someone watching a painting with a microscope. Hence, aggregation allows to study the network at a scale which is relevant compared to its activity. The second reason for using aggregation is that it produces graphs. They give an instantaneous view of the network (snapshot) which is practical in itself to get a view of what the object under study looks like and one can use the rich set of notions developed in graph theory to analyze the considered dynamic network.

If the benefits of aggregation are clear, on the other hand, it also raises some important concerns. Indeed, the length chosen for the aggregation window usually has a strong impact on the properties of the aggregated graph series [79, 129, 133]. This raises the question of which time scale should be chosen to study a given dynamic network and how much the properties studied, based on which conclusions are derived, are sensitive to the length of the aggregation period used [110, 122, 154]. It points out that in any case, this period should not be chosen without well established evidence, as it is currently done in most of the studies cited above. Pushing further, it is not even clear whether an aggregated series faithfully describes the original link stream. Indeed, the aggregation process goes along with a loss of information: in each aggregation window, the information on the exact times at which links occur in this window is lost. In particular, in a given time window,the causality information is lost, thus it is impossible to know whether a given link $(a, b)$ has occurred before or after another one $(b, c)$. This question, which determines whether it is possible to go from node $a$ to node $c$, via $b$, within this time window (only if $ab$ has occurred before $bc$), is crucial for many phenomena taking place on the dynamic network, such as epidemic spreadings, possibilities of communications and cascade driven by influence for example. The wider the aggregation period, the greater the amount of information lost. At the limit, aggregating a link stream over the whole period of study yields one single static network which misses all the information on the order of occurrences of links and which therefore very poorly captures the structure of the original dynamic network, see e.g. [106]. Then,

more generally, for a given aggregation period, one can ask whether the obtained graph series is a faithful representation of the original link stream. This is precisely the question we address here.

We show that for many dynamic networks, the length $\Delta$ of the window chosen for aggregating the network into a graph series exhibits a threshold, which is proper to each network. Beyond this threshold, the propagation properties of the graph series obtained from aggregation show evidence of alteration, while they are mostly preserved below it. We design a method, called the *occupancy method*, in order to determine this threshold, which we call the *saturation scale* and denote $\gamma$. We apply and validate the occupancy method on various real-world datasets, as well as on synthetic dynamic networks.

This answers the fundamental question of deciding whether a given aggregation period gives rise to a graph series that faithfully describes the original dynamic network. The aggregation periods beyond the saturation scale alters the properties of propagation of the dynamics. This range of scale must then be avoided or used only for analyzing properties of the series that do not suffer this alteration.

Moreover, the saturation scale, which is the larger non-altering aggregation period, is a characteristic time scale of the network. It can then be used to compare the properties of different dynamic networks at a same level of aggregation, which is very interesting in itself. Finally, let us emphasize that our method is fully automatic and does not require any parameter as input. Therefore, it can easily be incorporated into any automatic tool for analyzing dynamic networks.

## 2.2 Related work

It is paradoxical to note that, while the question of the influence of aggregation on the properties of the formed graph series is largely ignored in most of the studies on dynamic networks, this question actually already received a lot of specific attention [23, 33, 60, 79, 129, 133, 142, 147].

In [79], the authors lead a systematic analysis of what is visible from the structure of a dynamic phone-call network when it is aggregated at different time scales. They show that significant characteristics of the dynamics of the network appear at different scales of analysis, which implies that one should use the broad spectrum of possible scales in order to reveal these different properties of the dynamics. Though we are also concerned by the impact of the aggregation period on the properties of the formed graph series, our motivation and goal are clearly different from those of [79]. Here, we do not intend to find aggregation scales that reveal the key properties of the dynamic network. Instead, we aim at determining the range of aggregation scales such that the formed graph series faithfully describe the original network. Making statistics on the network out of this range of scale may still reveal interesting facts that are invisible at other scales. Nevertheless, for such aggregation scales greater than $\gamma$, one should consider only statistics that are not sensitive to the loss of information induced by aggregation (like those used in [79] for example), excluding all statistics based on propagation properties of the dynamics.

This is also the point of view developed in [129], where the authors study the impact of aggregation over the properties of random walks in a dynamic network. They show that the probability of occupation of nodes

of the network by such random walks is deeply impacted by aggregation, implying that it should be used with great caution when dealing with phenomenon that depends on propagation properties of the dynamics. The key contribution of the work of [129] is to emphasize and analytically explain the impact of aggregation on random walks, but it does not provide any way of determining a maximum aggregation period that can be used safely, which is precisely our goal here.

In [133], the authors study the impact of the length of the aggregation window, as well as the impact of the type of windows used (disjoint or overlapping or starting at the beginning of the period of study), on the output of a dynamic community tracking algorithm taking as input a series of graphs. Their results show that both the length and the type of the windows used have a strong impact on the dynamic communities outputted by the algorithm. As [129], the purpose of their work is to provide a deeper understanding of the effect of aggregation, but it is not intended to choose a suitable aggregation period.

Contrastingly, the goal of [147] is precisely to determine an ideal aggregation period. In their method, this period is obtained as a trade-off between two metrics that vary monotonically and oppositely with regard to aggregation: one describing the loss of information (increasing with aggregation) and one describing the noise contained in the series of snapshots (decreasing with aggregation). Compared to them, here, we are concerned only with the loss of information. This allows us to avoid some drawbacks and limitations inherent to the approaches based on a trade-off: i) the value selected for the aggregation period strongly depends on the importance given to each metrics and ii) the selected value does not reveal any particular behavior of the properties of the network used in the trade-off, as each of them varies smoothly and monotonically from one extremal value to another one. On the contrary, our method does not depend on any arbitrary choice of ponderation and reveals a natural change in the way the network responds to aggregation at a certain aggregation scale that we determine.

[33] also aims at determining an appropriate time scale for aggregating a link stream into a graph series. Their method does not take into account the loss of information but is instead based on the modes of periodicity and on the self-similarity of the time series of some properties of the snapshots. They observe that the offset time for which the self similarity of these time series is zero is close to half of the period of the highest frequency visible in their spectra, which is the aggregation period suggested as a result of their method. Though this provides a very relevant time scale for analyzing dynamic networks, its meaning is different from the meaning of the saturation scale we are looking for in this chapter. Indeed, an important part of the activity of dynamic networks takes place at time scales much smaller than their modes of periodicity. Therefore, using such periods for aggregation usually induces an important loss of information, which we aim at avoiding here. Let us mention that a similar approach based on modes of periodicity of some time series associated to the network was previously used in [51, 53].

The approach of [142] is noticeable in that they develop a method to aggregate a link stream on variable length windows. To this purpose, they fix the beginning time of the current aggregation window and they observe the evolution of some statistics of the aggregated graph as the ending time of the window increases. When the observed statistics has converged, they end of the current window and start a new one. The idea is to form a series of so-called ”mature” graphs, meaning that these graphs have been aggregated on a time

window long enough so that the properties of the formed graph would not change much if it was aggregated on a longer period of time. This motivation is clearly different from ours and the loss of information due to aggregation may occur before the convergence of the properties of the formed graph.

[60] and [23] consider the aggregation of a particular class of link streams: those that are obtained as the result of the oversampling of a dynamic network where links do not occur punctually but instead last over a time interval. Such dynamic networks are often measured by sampling processes that repeatedly check (often periodically) what are the links existing in the network at different times along the period of study, e.g. using sensor devices to measure contacts between individuals [27, 53]. These sampling processes introduce some noise in the data, for example due to failure to measure some links that do exist. The aim of [23, 60] is to find an aggregation period that removes the noise introduced by the sampling process and allows to retrieve the original signal. Here, both our purpose and the kind of dynamic networks we consider are different. We deal only with link streams where links are punctual and do not last over time. The approach of [23, 60] is not intended and not directly applicable to this kind of link streams. Conversely, it must be clear that applying our method to lasting links would require some adaptation and is one key perspective of our work.

For sake of completeness, let us mention two other works that address in different ways the problem of aggregation of link streams into graph series. [11] design a tool for visualizing a dynamic network as a series of snapshots, which takes as parameter the length of the aggregation window. One of the interest of this tool is that it helps to visually choose an aggregation period that gives a comprehensive view of the evolution of the network. Finally, [107] provides alternatives to aggregation by designing two representations of a dynamic network that encode both time and links in the form of a static graph structure.

## 2.3 Preliminaries

We describe our methodology in discrete time and with non-directed links, but actually, it applies the same if the time $t$ is continuous and if the links are directed (as in the real-world datasets we consider in Section 2.6). The only restriction which is meaningful here is that links are punctual events and therefore have no duration. The case where links exist during one interval of time instead of one instant requires some adaptation, both for continuous and discrete time. We now formally define some of the concepts we use in the chapter, starting with the process of aggregation of a link stream into a graph series, see example in Figure 2.3.1.

**Definition 1** (Aggregation)**.** *The aggregation, on disjoint time windows of equal length, of a link stream $\mathcal{L}$ on the period of study $[0, T]$ consists in choosing a constant time period $\Delta$ such that $\Delta = T/K$ for some integer $K \geq 1$ and forming the graph series $\mathcal{G}_\Delta = (G_k)_{1 \leq k \leq K}$ defined by $G_k = (V, E_k)$ with*

$$E_k = \{uv \mid \exists (u, v, t) \in \mathcal{L}, (k-1)\Delta \leq t < k\Delta\}$$

*and $V$ is the same for all graphs of the aggregated series: it is the set of nodes involved in the link stream $\mathcal{L}$.*

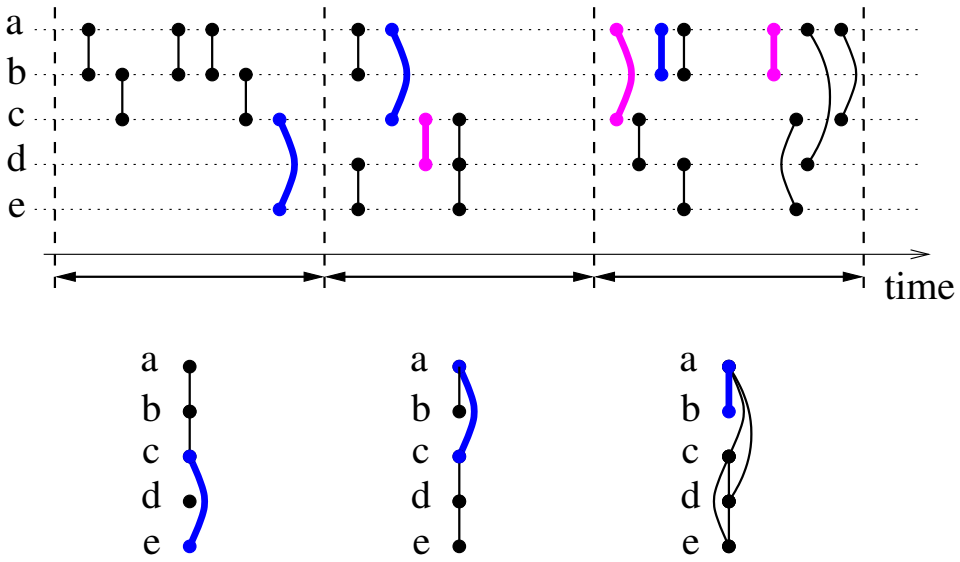



Figure 2.3.1: A link stream $\mathcal{L}$ and the graph series $\mathcal{G} = (G_1, G_2, G_3)$ obtained by aggregating $\mathcal{L}$ using a period $\Delta$. The bold dark-blue links depict a temporal path, from $e$ to $b$, in the link stream and its corresponding temporal path in the graph series. The bold light-pink links also form a temporal path in the link stream, from $d$ to $b$, but there is no temporal path from $d$ to $b$ in the graph series, because it would require to use two links of graph $G_3$, which is not allowed (see Remark 1).

A temporal path, in a link stream or a graph series, is a sequence of edges defining a path and occurring at strictly increasing time along the path (see examples given in Figure 2.3.1).

**Definition 2** (Temporal path in link stream)**.** *In a link stream $\mathcal{L}$, a temporal path $P$ is a sequence $(u_i, v_i, t_i)$ of triplets, with $1 \leq i \leq l$ and $l > 0$, such that $\forall i, (u_i, v_i, t_i) \in \mathcal{L}$ and $\forall i > 1, u_i = v_{i-1}$ and $\forall i, j$, if $i < j$ then $t_i < t_j$.*

**Definition 3** (Temporal path in a series of graphs)**.** *In a series of graphs $\mathcal{G} = (G_k)_{1 \leq k \leq K}$, a temporal path $P$ is a sequence $(u_i, v_i, k_i)$ of triplets, with $1 \leq i \leq l$ and $l > 0$, such that $\forall i, u_i v_i \in E(G_{k_i})$ and $\forall i > 1, u_i = v_{i-1}$ and $\forall i, j$, if $i < j$ then $k_i < k_j$.*

**Remark 1.** *Note that, in Definition 2 and 3, the inequalities are strict. This implies that a temporal path cannot use two links belonging to the same graph of the series or occurring at the same time in the link stream.*

Temporal paths are an essential notion as they capture the propagation properties of the dynamic network. Indeed, all diffusion phenomena in the network, such as communication of information, spreading of epidemics and cascades of influence for example, respect time causality: a node needs to be reached by the diffusion before it can propagate it further. Therefore, all these phenomena occur on and follow temporal paths of the dynamic network.

There are two notions of length associated to a temporal path: the topological length, which is the classical one for static graphs, and the duration of the path.

**Definition 4** (hops(P) and time(P))**.** *In a link stream or a graph series, the* topological length *of a temporal path $P = ((u_i, v_i, t_i))_{1 \leq i \leq l}$ is the number $l$ of edges in the path. In the rest of the chapter, we call it the* number of hops *of $P$ and denote it $hops(P)$.*

*The* duration *of path $P$, denoted $time(P)$, is $t_l - t_1$ in a link stream and $t_l - t_1 + 1$ in a graph series (because each $t_i$ is not an instant as in a link stream but an interval of time which has a duration).*

**Remark 2.** *By definition, in a graph series, we always have $hops(P) \leq time(P)$ for any temporal path $P$.*

In the rest of the chapter, we also use three notions of distance at time $t$ between two nodes $u, v$ of a link stream or a graph series. These notions are based on the minimal arrival time $t_{arr}$, if any (otherwise $t_{arr}$ is undefined), among all paths from $u$ to $v$ whose departure time is not before $t$. The *distance in time*, denoted $d_{time}(u, v, t)$, is simply defined as $t_{arr} - t$ in a link stream and $t_{arr} - t + 1$ in a graph series, with the convention $d_{time}(u, v, t) = +\infty$ if $t_{arr}$ is undefined. The *distance in hops*, denoted $d_{hops}(u, v, t)$, is the minimum number of hops among all paths realizing the distance in time $d_{time}(u, v, t)$. By convention, $d_{hops}(u, v, t) = +\infty$ when $d_{time}(u, v, t) = +\infty$. Finally, the distance in absolute time, which is dedicated to aggregated graph series only, is denoted $d^{abs}_{time}(u, v, t)$ and defined by $d^{abs}_{time}(u, v, t) = \Delta . d_{time}(u, v, t)$. It is the absolute time needed to go from node $u$ to node $v$ in the aggregated graph series, with a departure time not before $t$, taking into account the fact that each graph of the series represent a time interval of length $\Delta$. Interestingly, this last notion contains in itself the imprecision of the timestamps of the aggregated series.

## 2.4 TEMPORAL SCALE : DIFFICULTY OF THE PROBLEM

As previously explained, the larger the length of the aggregation window, the greater the loss of temporal information due to aggregation, as the exact times of occurrence of the links within one given window are lost. Then, we can reformulate the problem as follows: what is the maximum aggregation period that induces no significant loss of information in the graph series compared to the original link stream? The natural way to proceed to answer this question is to make the aggregation period vary from its minimal value to its maximal value and to observe the variations of the properties of the obtained series of graphs in the meanwhile. Then, one would hope to find a time scale beyond which the variation of these properties exhibit a qualitative change. Unfortunately, this does not happen for the classical properties of interest of the graph series. On the opposite, when the aggregation period varies, these properties varies smoothly from one extremal value to another one. Figure 2.4.1 shows the results for several properties of the Irvine network, which we use as an example to describe our method in the first sections of this article (see Section 2.6 for a description of the Irvine dataset).

Figure 2.4.1 top-left shows the variations of the mean density of the snapshots of the series, which is also equivalent, up to a multiplicative constant, to the mean degree of the nodes in all the snapshots. The plot shows that when the aggregation goes from the minimal temporal resolution of the timestamps (1s) to the total length of the period of study ($\sim$1175h), these two properties linearly varies from a very small value ($5.7 \times 10^{-7}$) to their maximal value ($7.2 \times 10^{-3}$), which is the one obtained by aggregating the whole dynamic network into one single graph.

The plot in Figure 2.4.1 top-right shows that in the meanwhile the mean size of the largest connected component in each snapshot as well as the mean number of non isolated vertices per snapshot (which are very close) exhibit the same behavior: their values go increasingly from a minimal one (2.3 nodes) to a maximal one (1509 nodes), which is the total number of nodes in the network, without exhibiting any non-smooth behavior at any time scale.

Let us now examine the variations of distance properties according to the aggregation period. Figure 2.4.1 bottom-left gives the variation of the mean distance in time $d_{time}(u, v, t)$ (see Section 2.3) for all couples $(u, v)$ of nodes and all time $t$ (such that $d_{time}(u, v, t)$ is finite), in logarithmic scale. As one can see, the curve is almost a straight line, indicating that the mean distance in time depends on $\Delta$ following a power law. This comes from the fact that the number of graphs formed in the aggregated series varies as $1/\Delta$: the plot shows that the mean distance in time varies accordingly. This does not help much to detect a time scale at which the properties of the aggregated series significantly change their behavior.

Pushing further, in Figure 2.4.1 bottom-right, we also plotted the mean distance in hops (empty squares) and the mean distance in absolute time (filled squares), both in linear scale. The rational for using the distance in absolute time, defined as $d^{abs}_{time}(u, v, t) = \Delta.d_{time}(u, v, t)$, is that it does not suffer from the dependence on $1/\Delta$ previously highlighted for the distance in time, since it is canceled by the multiplication by $\Delta$. Then, it gives a clearer insight into the variations of the distance in time with the aggregation period.

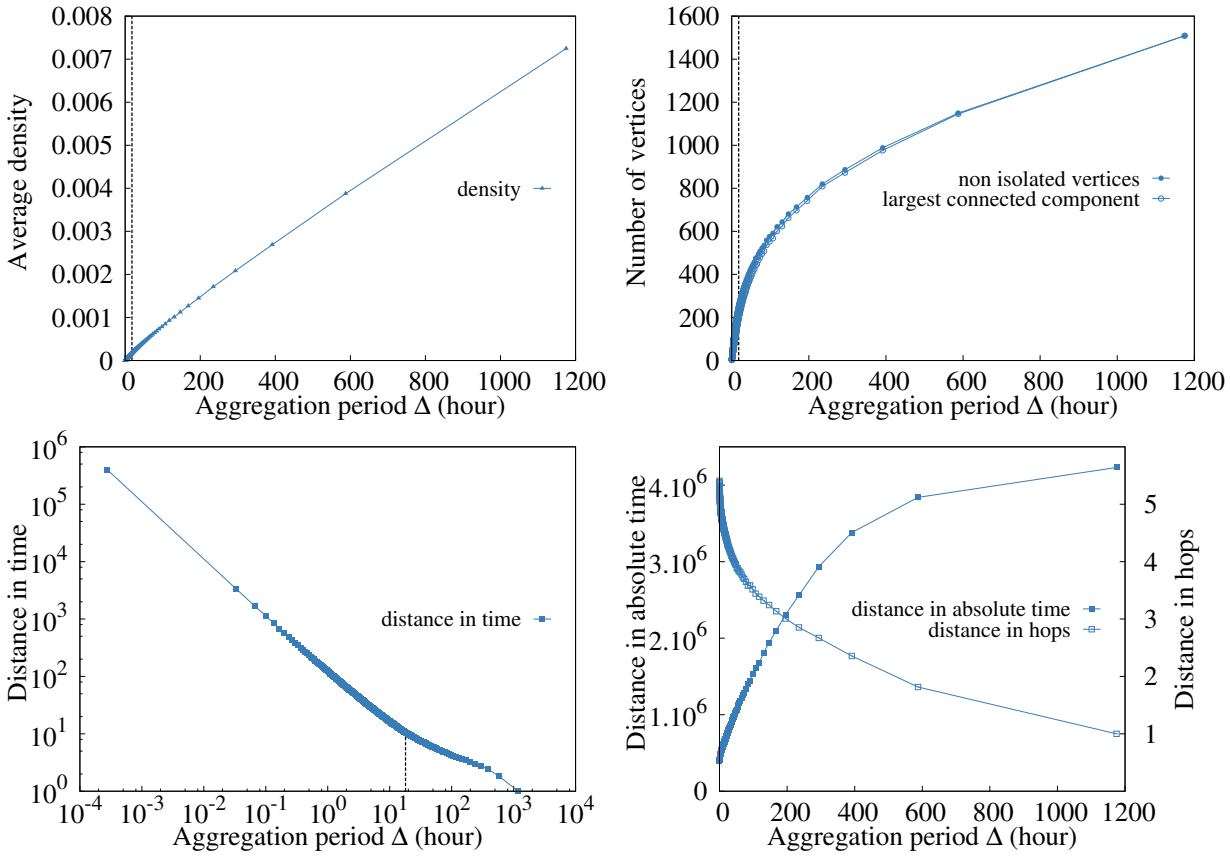



Figure 2.4.1: Variation of some classical parameters of the aggregated series of graphs (y-axis) according to the aggregation period $\Delta$ (x-axis), for the Irvine network (see Section 2.6). Top-left: density. Top-right: connectedness properties. Bottom-left: distance in time (log-log scale). Bottom-right: other distance properties (linear scale). The dotted line shows the aggregation period returned by the occupancy method: 18h.

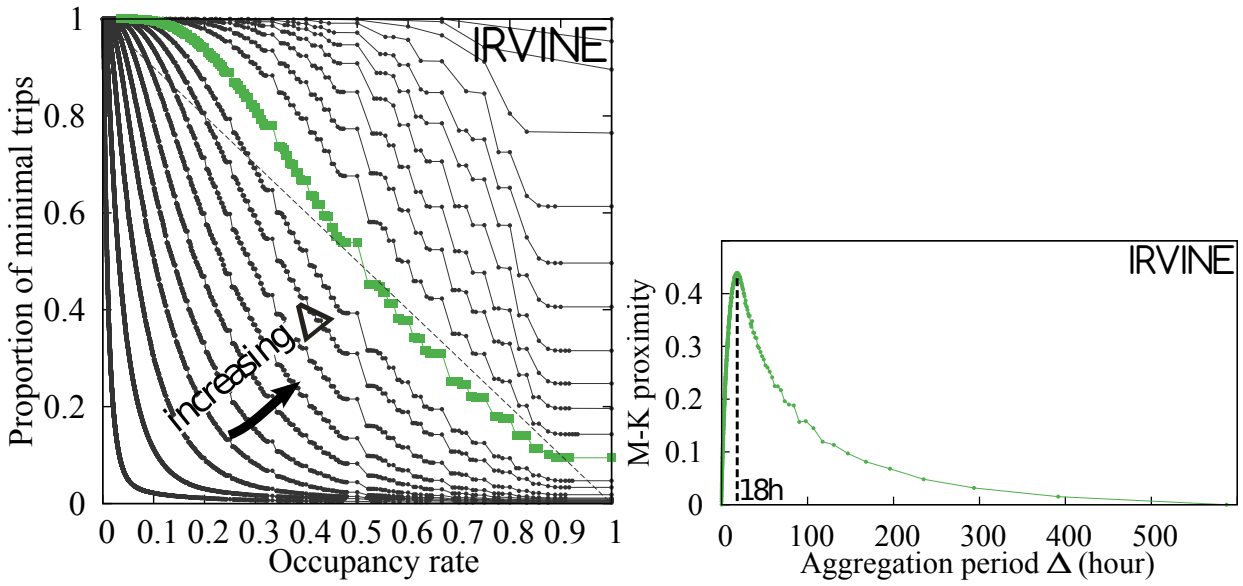



Figure 2.4.2: Left: Inverse Cumulative Distributions (ICD) of the occupancy rates (x-axis) of the minimal trips of the aggregated series $\mathcal{G}_\Delta$ for several values of the aggregation period $\Delta$ in the range $[1, T]$, for the Irvine network. Right: M-K proximity (y-axis) of these distributions with the uniform density distribution according to $\Delta$ (x-axis).

Unfortunately, as one can see, the situation is the same as for the other parameters previously studied. When the aggregation period $\Delta$ increases from the minimal temporal resolution of the timestamps (1s) to the total length of the period of study ($\sim$1175h), the mean distance in absolute time increases as well, going monotonically from its minimal value ($\sim$110h) until its maximal value ($\sim$1175h), which is by definition equal to the total length of the period of study, as there is only one graph in the series formed using the maximum value of $\Delta$. In the meanwhile the mean distance in hops (empty squares in Figure 2.4.1 bottom-right) decreases, from $5.4$ to $1$, without exhibiting any remarkable change at any value of the aggregation period.

Thus, the observation of the variations of the classical properties of the graph series with the aggregation period does not point out scale at which some qualitative changes occur in the way the dynamic network responds to aggregation. Instead, one finds a regular drift from one extreme value to another one[2]. This constitutes the main difficulty of the problem we consider. An important contribution of our work is to exhibit a finer property of the graph series that is able to reveal a time scale where such qualitative changes occur.

[2]Note that we present results for only one dataset but they hold similarly for all the four datasets we consider in this chapter, cf. Section 2.6.

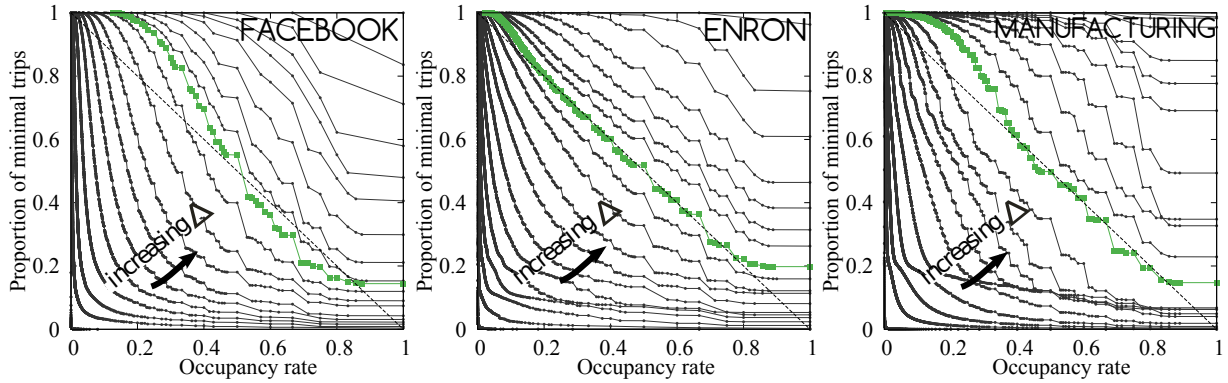



Figure 2.5.1: Inverse Cumulative Distributions (ICD) of the occupancy rates (x-axis) of the minimal trips of the aggregated series $\mathcal{G}_\Delta$ for several values of the aggregation period $\Delta$ in the range $[1, T]$, for the Facebook, Enron and Manufacturing networks.

## 2.5 THE OCCUPANCY METHOD

We now give the definitions necessary to describe our method and we illustrate it on a sample real-world network, the Irvine network (cf. Section 2.6).

**Definition 5** (Trip and minimal trip)**.** *A* trip *is a quadruplet* $(u, v, t_{dep}, t_{arr})$ *such that there exists a temporal path from* $u$ *to* $v$ *whose starting time from* $u$ *and arriving time at* $v$ *are both in the interval* $[t_{dep}, t_{arr}]$*. A trip* $(u, v, t_{dep}, t_{arr})$ *is* minimal *if there exists no trip from* $u$ *to* $v$ *in an interval* $[t'_{dep}, t'_{arr}]$ *strictly included in* $[t_{dep}, t_{arr}]$ *(i.e.* $[t'_{dep}, t'_{arr}] \subsetneq [t_{dep}, t_{arr}]$*).*

**Definition 6** (Transition and shortest transition)**.** $P = ((a, b, t_1), (b, c, t_2))$*, a temporal path* $P$ *on two hops, is called a* transition*, and* $P$ *is a* shortest transition *if* $(a, c, t_1, t_2)$ *is a minimal trip.*

**Definition 7** (Occupancy rate)**.** *For a graph series* $\mathcal{G}$ *and a temporal path* $P$ *in* $\mathcal{G}$*, the* occupancy rate *of path* $P$*, denoted* $occ(P)$*, is defined as* $occ(P) = hops(P)/time(P)$*. The occupancy rate of a minimal trip* $(u, v, t_{dep}, t_{arr})$ *is the occupancy rate of a temporal path starting from* $u$ *at* $t_{dep}$ *and arriving at* $v$ *at* $t_{arr}$ *and having the minimum number of hops among such paths.*

The rational behind the occupancy rate $occ(P)$ is to count the proportion of time steps between $t_{dep}$ and $t_{arr}$ that are effectively used by the path $P$ to move from one node of the dynamic network to another one. In particular, note that since $0 < hops(P) \leq time(P)$ (cf. Remark 2) then we always have $0 < occ(P) \leq 1$.

In order to determine the saturation scale $\gamma$, we make the aggregation period $\Delta$ vary from its minimal value, the resolution of the timestamps, until the whole length $T$ of study of the network. For each value of $\Delta$ we form the aggregated graph series $\mathcal{G}_\Delta$ for which we compute the set of minimal trips and their occupancy rates. Then, for each $\Delta$, we plot the distribution of occupancy rates of all the minimal trips in $\mathcal{G}_\Delta$ (considering all pairs of nodes and all time intervals), see Figure 2.4.2 left.

Necessarily, when $\Delta$ is close to its minimal value, provided that the resolution of the timestamps is fine enough, the distribution of occupancy rates must be concentrated on values close to $0$. The reason is that the aggregation windows contain only few data and the shortest paths therefore need to wait several slot of times before finding one opportunity to perform the next hop. On the opposite, when the aggregation period reaches its maximum value, by definition, all the minimal trips are made of one single link (because there is only one graph in the aggregated series) and their occupation rate is $1$. Then, the distribution is again concentrated, this time on the value $1$. What is remarkable here (Figure 2.4.2 left) is that the distribution changes from values concentrated near $0$ to values concentrated on $1$ in a very specific manner: it first progressively stretches toward $1$ until it almost equally occupies all the values on the range from $0$ to $1$ and then it contracts again, leaving the low values to progressively concentrate on the values close to $1$.

The saturation scale $\gamma$ is precisely the value of $\Delta$ for which the distribution is maximally stretched on the interval $[0, 1]$ (curve marked with green squares on Figure 2.4.2 left). In order to detect it, we compute for each value of $\Delta$ in the total range of variation, the M-K distance $d(\Delta)$ (see Section 2.8 for a definition) between the distribution obtained for $\Delta$ and the uniform density distribution on $[0, 1]$, i.e. the distribution whose inverse cumulative is the straight line $y = 1 - x$. We then plot the M-K proximity, defined as $1/2 - d(\Delta)$, in Figure 2.4.2 right. This confirms the observation made above on the way the distribution first stretches and then concentrates again: accordingly, the M-K proximity first increases and then decreases. As a consequence, the value $\gamma$ returned by our method is the value of $\Delta$ that realizes the maximum of the M-K proximity. Of course, one may think of many other ways to determine which $\Delta$ gives the maximum stretch of the distribution. We actually tried several of them (see Section 2.8) and we chose to use the M-K distance with the uniform density distribution because it gives results that are visually satisfying and it is conceptually simple.

Let us now explain the meaning of $\gamma$. A very low occupancy rate for most minimal trips of the series denotes that the data in each aggregation window is sparse, which implies that the information contained in the link stream is mainly preserved in the graph series. On the opposite, a very high occupancy rate for most of the minimal trips reveals a loss of information. Indeed, this indicates that, at each time in the graph series, there is a high probability to find a next hop to perform on any given shortest path, meaning that, in each snapshot, a high proportion of nodes are involved in a high number of edges. Then, at the same time, the information on the existence or the non existence of a transition, in the original link stream, using a couple of these edges incident to one same node is lost, which constitutes the essential loss resulting from the aggregation process.

In the first phase of variation of the aggregation period, below $\gamma$, only the low values of the distribution increase, while the proportion of high occupancy rates almost does not change. This means that during this phase, the effect of increasing the aggregation period is mainly to fill the lack of data in the aggregation windows without inducing a significant loss of information. On the opposite, in the second phase, beyond $\gamma$, there is a strong increase of the proportion of minimal trips having a very high occupancy rate, $1$ or close to $1$, indicating that the loss of information due to aggregation becomes non-negligible. Therefore, the saturation scale $\gamma$ appears as a separation between the range of values, below $\gamma$, where the aggregated graph

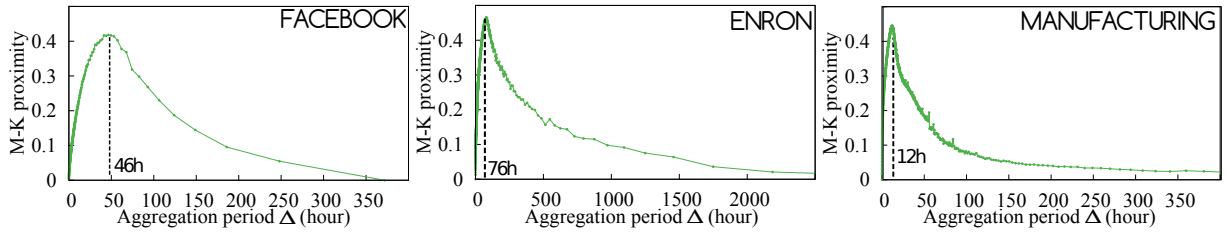



Figure 2.5.2: M-K proximity (y-axis) of the distribution of occupancy rates of minimal trips of the aggregated series $\mathcal{G}_\Delta$ according to the aggregation period $\Delta$ (x-axis), for the Facebook, Enron and Manufacturing networks.

series still faithfully describes the original link stream and the range of values, beyond $\gamma$, where aggregation alters the properties of propagation of the original link stream.

## 2.6 Results on real-world datasets

In this section we apply our methodology and discuss the results obtained on four link streams, whose timestamps have a resolution of 1s. The *UC Irvine messages* network [117], which is the one used for presentation of the method in the previous section, is made of 48 000 messages sent between 1 509 users of an online community of students from the University of California, Irvine, over a period of 48 days. The *Facebook wall posts* network [155] is made of 11 991 wall posts between a group of 3 387 Facebook users over a period of 1 month. The *Enron emails* network [77] contains 15 951 individual emails sent between a group of 150 employees of the Enron company during year 2001. Finally, the *Manufacturing emails* network [100] contains 82 894 internal emails between 153 employees of a mid-sized manufacturing company over a period of 8 months.

We applied the occupancy method on each of these four datasets. The distributions of occupancy rates of the minimal trips in the aggregated graph series are given on Figure 2.4.2 left for Irvine and Figure 2.5.1 for the three other networks, their M-K proximity with the uniform density distribution is given on Figure 2.4.2 right and Figure 2.5.2. One can see that the observations made on the Irvine network in Section 2.5, hold for all the four datasets. When the aggregation period $\Delta$ increases, the distribution of occupancy rates, initially concentrated near $0$, stretches until it occupies almost equitably all the range of values between $0$ and $1$, and then concentrates again on the values close to $1$. Consequently, the proximity with the uniform density distribution first increases, until it reaches a maximum for $\Delta = \gamma$, which is the saturation scale returned by the occupancy method, and then decreases until the aggregation period reaches its maximum value $T$. This shows that the way the distribution of occupancy rates evolves with the aggregation period is a fundamental phenomenon common to many dynamic networks, therefore guaranteeing that our method is sound and that it can be used for a wide range of dynamic networks.

The values returned for $\gamma$ in each of the four cases are: 18 hours for the Irvine message network, 46

hours for the Facebook wall-post network, 78 hours for the Enron email network and 12 hours for the Manufacturing email network. These values, between half a day and three days, are in accordance with the fact that both emails and on-line social network messages are generally not dedicated to live discussions. In the case of email networks for example, most of people only send some emails a day and frequently wait for some hours or some days before getting a reply. Therefore, this range of values seems appropriate for the largest aggregation scales providing accurate views of the original link streams.

The aggregation periods returned by our method also appear to be in accordance with the level of activity of these 4 networks. The two greater values, 46h for Facebook and 78h for Enron, are obtained for the two networks that have the lower activity, $0.12$ and $0.29$ messages sent in average per person per day for Facebook and Enron respectively. The two other networks have higher activities, $0.66$ messages per person per day in the Irvine network and $2.22$ in the Manufacturing network, and have smaller saturation scales, 18 hours and 12 hours respectively. As one can see, the average activity has a strong influence on the saturation scale, even though this is not the only parameter affecting it. We further investigate the relationship between the level of activity and the saturation scale in the next section.

Finally, let us emphasize that the aggregation period $\gamma$ returned by our method should not been interpreted as the best possible one but instead as an upper-bound on the aggregation periods that are suitable for studying the network. For many practical studies, one may prefer to choose an aggregation period slightly lower than $\gamma$, which will preserve more carefully the properties of the network. For example, in the case of the four networks we study here, one can note that the proportion of minimal trips having occupancy rate $1$ started to increase just before the distribution of occupancy rates reaches its maximal stretched position (the one selected by our method). Then, one could prefer to use an aggregation period smaller than $\gamma$ in order to get a finer grain representation of the dynamic network. In Section 2.9, we give some ways to directly estimate the loss of information in the aggregated graph series that can be used to choose more accurately the aggregation period in the range of scales immediately preceding $\gamma$.

## 2.7 Results on synthetic networks

We now investigate how the aggregation period returned by our method depends on the level of activity of the link streams considered, i.e. the number of links per node and per unit of time, and on the temporal heterogeneity of this activity. To this purpose, we use two kinds of synthetic dynamic networks, where the activity is uniformly distributed between all pairs of nodes. The first kind, called *time uniform networks*, is generated by assigning $N$ links ($N << T$) to each pair of the $n = 100$ nodes of the network and uniformly randomly choosing each of their timestamps between $0$ and $T = 100\,000$s. We make the value of $N$ vary from $10$ to $100$ and for each of these values, we compute the aggregation period $\gamma$ returned by the occupancy method. Results are given in Figure 2.7.1 left, which shows $\gamma$ as a function of the average inter-contact time of one node, that is $T/(N(n-1))$. For these time uniform networks, the aggregation period returned by the occupancy method is perfectly proportional to the average inter-contact time, showing that our method

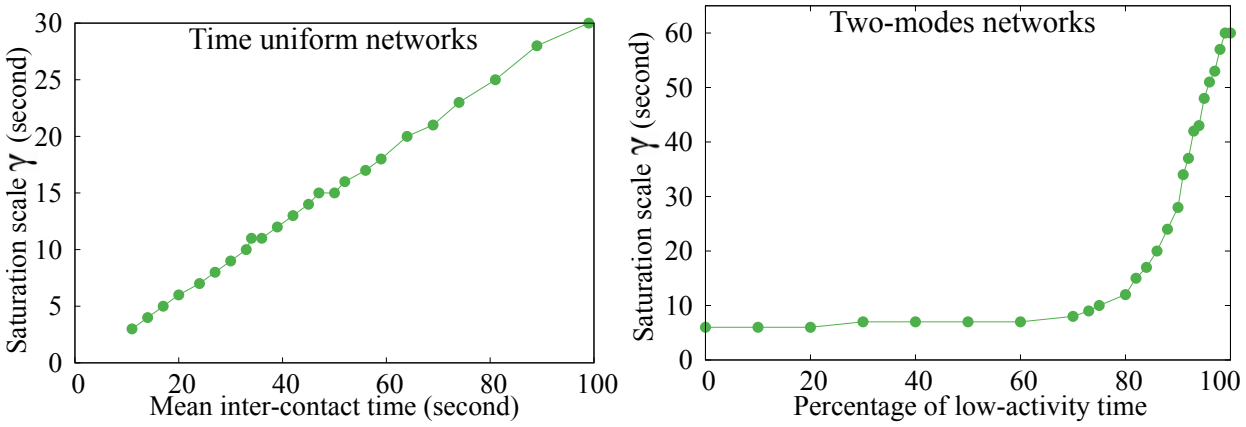



Figure 2.7.1: Left: for time uniform networks, saturation scale (y-axis) in function of the mean inter-contact time of nodes (x-axis). Right: for two-mode networks, saturation scale (y-axis) in function of the percentage of low-activity time (x-axis).

correctly takes into account the level of activity of the link stream.

However, most of the dynamic networks encountered in practice are far from being uniformly active over time. Many of them instead alternate periods of intense activity with periods of lower activity. In particular, this is the case for networks coming from human activities, such as the ones considered in Section 2.6, which often exhibit circadian rhythms. Then the question naturally arises to know how the saturation scale behaves according to this temporal heterogeneity. Does it simply make the average between the different levels of activity? Or does it favor one of them? To answer these questions we generate *two-mode networks* that are built by $10$ alternations of one period of high activity and one period of low activity, which are time uniform networks with parameters $N_1, T_1$ and $N_2, T_2$ respectively. $N_1$, $N_2$ and the whole length $T = 10(T_1 + T_2)$ of study are fixed and we vary the ratio between $T_1$ and $T_2$.

Figure 2.7.1 right gives the saturation scale $\gamma$ as a function of the percentage $\rho = T_2/(T_1 + T_2)$ of low-activity time in the network. The curve goes from the value of $\gamma$ for the high-activity mode (for $\rho = 0\%$) to the one, much larger, for the low-activity mode (for $\rho = 100\%$). The plot shows that when the proportion of low activity varies from 0% to 70-80%, the saturation scale almost does not increase: it remains very close to the smaller value of the high-activity network, which preserves better the information contained in the original link stream. This is surprising as one would rather expect the saturation scale to be a compromise between its value for the low-activity periods and its value for the high-activity periods. This shows that in presence of heterogeneity of the activity along time, even with high-activity periods occupying only 30% to 20% of the time, the saturation scale returned by the occupancy method is respectful of this important part of the dynamics. Moreover, and importantly, the fact that the saturation scale does not linearly vary with respect to the percentage of low-activity time in the network shows that, for networks that are not time uniform (which is in particular the case of real-world networks), the saturation scale returned by the occupancy method does not only depend on the mean inter-contact time of nodes in the network (or equivalently on the

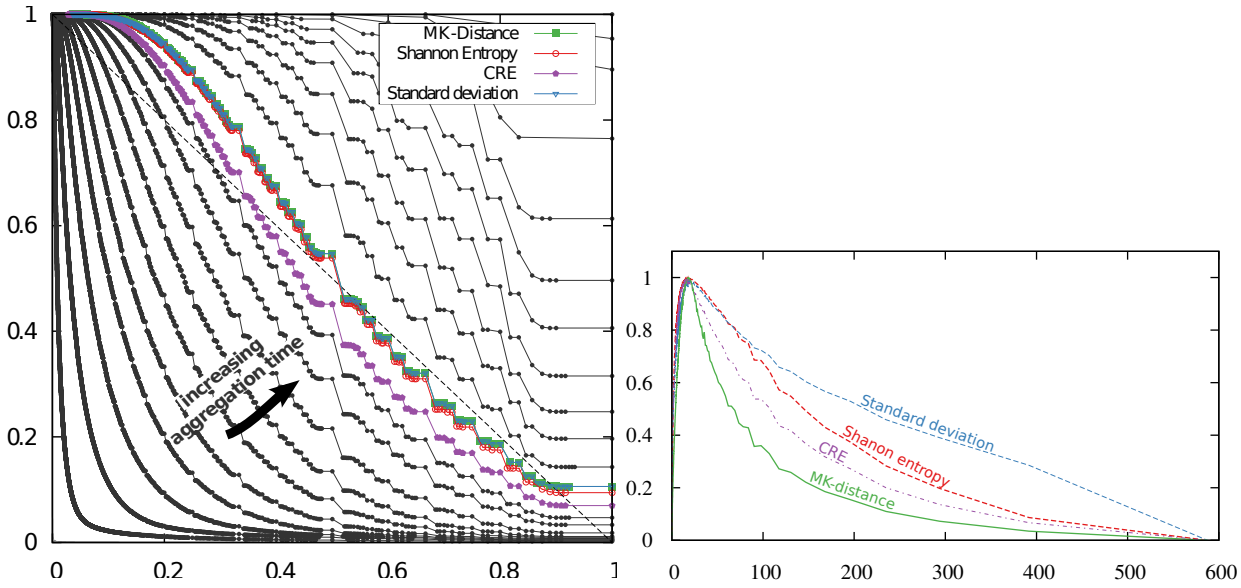



Figure 2.8.1: Results of four methods for selection of the more uniformly spread distribution: M-K proximity, standard deviation, Shannon entropy with 10 slots and cumulative residual entropy (CRE). The left plot shows the distributions selected by the maximum of each of the metrics and the right plot shows the variations (normalized to have maximum 1) of each metric (y-axis) depending on the aggregation period $\Delta$ (x-axis).

frequency of links in the network).

When the proportion of low-activity time goes beyond 80%, the aggregation period returned starts to increase until it reaches the value for the low-activity network when its proportion in time is 100%. This seems natural as when the low-activity part takes most of the time of the dynamic network, it does not make sense to continue to study it with a scale which is suitable only for a marginal part of the time. Nevertheless, we note that the increase of $\gamma$ is progressive. For example for 90% of low activity, the returned value is close to the arithmetic mean between the values for the two modes of the network. This shows that the returned aggregation does not forget too quickly the high activity part of the dynamics, which once again is a desirable feature of such a method.

## 2.8 Detection of the most uniform distribution

In this section we consider several methods for selecting the aggregation period $\gamma$ that gives the distribution of occupancy rates that is the most uniformly spread on $[0, 1]$, and we study the dependence of $\gamma$ on the chosen selection method. Until now, all the results we gave were obtained by selecting the distribution which minimizes the M-K distance with the uniform density distribution on $[0, 1]$. Of course, one may think of many other ways to select $\gamma$. This includes plotting the sets of distributions obtained when the aggregation period spans its entire range of variation, like on Figure 2.4.2, and selecting one distribution

by visual mean. This empirical method is likely to give the most satisfactory results in practice and will therefore be preferred in many studies. However, here, we are interested only in quantitative methods of selection, for two reasons. Firstly, we want to provide a uniquely defined value which can be used as a reference for comparing the saturation scales of different dynamic networks. Secondly, we want our method to be fully automatic in order to be easily incorporable to any tool for analysis of dynamic networks.

In addition to the method based on the M-K distance, which we used until now, we now consider four other selection methods and compare their results. These four additional methods are based respectively on: standard deviation, variation coefficient, Shannon entropy and cumulative residual entropy. Figure 2.8.1 gives the results obtained when applying these methods on the Irvine data set. We now describe and analyze them one by one each before giving a global comparison of their respective results.

M-K DISTANCE WITH THE UNIFORM DENSITY DISTRIBUTION. The Monge-Kantorovich distance is a way to measure the distance between two distributions of probability on the same support, here $[0, 1]$. It is defined as the area comprised between the two inverse cumulative distributions of the probability distributions to be compared. Here, as we are looking for the distribution which is maximally spread over $[0, 1]$, we compare each distribution with the uniform density distribution, which gives $dist_{M-K}(X) = \int_{[0,1]} |P(X > \lambda) - (1 - \lambda)| d\lambda$, where $X$ is the random variable defined by the occupancy rate. Then, the aggregation period we select is the one for which the distribution of occupancy rates $X$ realizes the minimum of $dist_{M-K}(X)$. In order to get the desired distribution for the maximum of the measure, instead of the minimum, as for all the other measures we consider, we rather use the corresponding proximity measure defined as $1/2 - dist_{M-K}(X)$, as $dist_{M-K}(X)$ is always less than $1/2$. This metric is the one we use throughout the article. It gives visually very satisfying results for all the data sets.

STANDARD DEVIATION. This method selects the distribution having the maximum standard deviation $\sigma = \sqrt{E[(X - \mu)^2]}$, where $X$ is the random variable defined by occupancy rate and $\mu$ is its mean value. This is one of the most direct measure one can think of in order to compare the spread of distributions on support $[0, 1]$. It gives very satisfactory results, comparable to the one obtained with the M-K distance. Nevertheless, it tends to select slightly higher aggregation period than the M-K distance, as the standard deviation is less penalized by the increasing of occupancy rates 1, which is the maximal value in the distribution. Then, usually, the aggregation period selected by the M-K distance is visually a bit more satisfying. This is the reason why we prefer to present our methodology using the M-K distance, but the two metrics actually give comparable results.

VARIATION COEFFICIENT. Another very natural method is the one that selects the distribution having the maximum variation coefficient $c_v = \sigma/\mu$, where $\mu$ is the mean and $\sigma$ the standard deviation of the distribution. Moreover, it could possibly correct the slight drawback of the standard deviation pointed above, which tends to select a little bit higher value than would desire. Unfortunately, the method based on variation coefficient suffers from a much more severe limitation: it favors too much distributions having

a small mean and therefore proposes only very short aggregation periods, or even not to aggregate at all. Among all the methods we tried, this is the only one which gives clearly unsatisfactory results to select the more spread distribution.

SHANNON ENTROPY. In information theory, the Shannon entropy $H(X) = E[-ln(P(X))]$ of a random variable $X$ (here the occupancy rate) allows to measure the spread of distributions on a given fixed finite support. This means that in order to compare different distributions, the set of possible values taken by the distributions (the support) must be the same and must be finite. As for the measure based on the M-K distance, the distribution which maximizes the Shannon entropy is the one with uniform density on the considered support. The difficulty we face here to use this measure is that the supports of the distributions we want to compare are not the same: the set of possible values of the occupancy rate is different for each aggregation period. There are different ways to deal with this issue in order to compare all the distributions we obtain when varying the aggregation period. The first one is to artificially take one common support for all the distributions, the minimal such support being the union of the supports of all obtained distributions. Unfortunately, when applied with this support, the method always selects the distributions that effectively use the larger part of the support, that is those obtained for very short aggregation periods. As noted for the variation coefficient, this does not give satisfactory results.

Another possibility to solve this issue is to discretize the segment $[0, 1]$ into $k$ slots of equal length and to compute for each distribution what is the probability that the value of the occupancy rate belongs to each slot. For example, when applied with $k = 10$, this measures gives very satisfactory results. On the other hand, the returned aggregation period depends on the number of slots chosen. The results are sensibly different using $k = 5$ or $k = 20$. With smaller number of slots, the method tends to select higher aggregation periods, while with greater number of slots, like previously, it favors the distributions having greater original supports, which are those obtained for short aggregation periods. For $k = 100$ this trend is already clearly marked: the value returned for $\gamma$ is less than half of the one returned with $k = 10$ and visually, the distribution selected does not appear to be spread over $[0, 1]$ in the best possible way among all the distributions obtained. Despite of this, we note that in the range of variation we consider for $k$, namely $[2, 100]$, and for the data sets we use, the selection method based on the Shannon entropy properly determines the order of magnitude of the saturation scale $\gamma$ and even gives visually very satisfying results for values of $k$ between 5 and 20. Nevertheless, because of its sensitivity to the chosen $k$, we decided not to use this selection method. The next metric is another attempt to correct the difficulties arising from the use of the Shannon entropy.

CUMULATIVE RESIDUAL ENTROPY (CRE). This is a variation of the Shannon entropy that is able to compare distributions with the same *infinite* support. It is suitable for our purpose as all the distributions we consider have support $[0, 1]$. In this case, the cumulative residual entropy is defined as $\varepsilon(X) = -\int_{[0,1]} P(X > \lambda) \log(P(X > \lambda))d\lambda$. As for the Shannon entropy, the maximum value of the CRE is reached for the uniform density on the considered support. It turns out that this selection method performs well on all the data

sets we used. It gives aggregation periods close to the one obtained by the M-K distance, usually shorter. Visually the results are satisfying, even if on some example this method appears to favor a bit too much distributions with large supports. But this is only a slight effect and this method seems quite suitable for our needs, and theoretically well funded. The reason why we preferred the M-K distance with the uniform density distribution is that it is conceptually much simpler and gives as good results.

Let us now compare the aggregation periods selected by the 5 methods above for the Irvine data set. First of all, let us note that all the selected aggregation periods are very close between $14.5$h and $18.7$h, except one of them, the one based on the variation coefficient. This method proposes an aggregation period of 1 second, which is the resolution of the timestamps, and the distribution it selects is very far from being uniformly spread on $[0, 1]$. The variation coefficient method therefore appears not to be suitable for our purpose. On this data set, the distributions selected by the M-K distance and the standard deviation methods are exactly the same. They are the distribution obtained with an aggregation period of $18.7$h. The distribution selected by the method based on Shannon entropy with 10 slots of width 0.1 is almost indistinguishable from the previous one, it is the distribution obtained for an aggregation period of $18.1$h. These two distributions are indeed visually quite well spread on $[0, 1]$ and therefore, these three methods give here very satisfying results, and they also did on other data-sets (not presented here). The method based on the cumulative residual entropy selects the distribution obtained with an aggregation period of $14.5$h, which is slightly lower than the three previous method. In this case, this distribution is also visually very well spread on $[0, 1]$ and then quite good for our purpose.

As a conclusion, except the method based on the variation coefficient, all the four other methods we considered appear to give satisfactory results on all the data sets we use. We chose the method based on the M-K distance because it is conceptually simple and it gives very satisfactory results. But more importantly, beyond the slight differences between these methods, the fact that they all give very close values of $\gamma$ shows that each of them is sound and is appropriate to detect the aggregation period that maximally stretches the distribution of the occupancy rates in the interval $[0, 1]$.

## 2.9 VALIDATION

In this section we quantify the amount of information which is lost when one aggregates the network using a given period $\Delta$. This allows us to validate our approach by evaluating the loss obtained for $\Delta = \gamma$. Moreover, this provides tools to select more accurately an aggregation period, in the range preceding $\gamma$, that is suitable for representing a given link stream as a graph series.

The first measure of loss we use is the proportion of shortest transitions (minimal trips with two hops, cf. Definition 6) that lay entirely in one aggregation window. These are exactly the shortest transitions of the original link stream that do not exist anymore in the aggregated series of graphs: all the other minimal trips having their two hops, say $(a, b, t_1), (b, c, t_2)$, in two different aggregation windows, say indexed $t'_1$ and $t'_2$, still exist in the form $(a, b, t'_1), (b, c, t'_2)$ in the aggregated series. We chose this way of measuring the loss as the shortest transitions are the key units that capture the possibilities of propagation in the link streams.

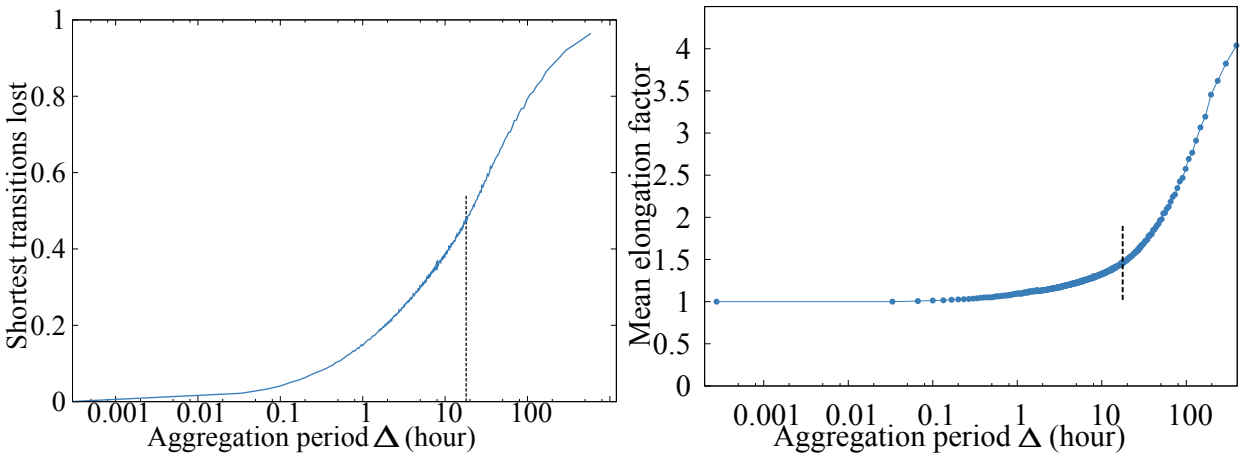



Figure 2.9.1: Left: proportion of shortest transitions lost (y-axis) in the aggregated series $\mathcal{G}_\Delta$ according to the aggregation period $\Delta$ (x-axis). Right: mean elongation factor of minimal trips of $\mathcal{G}_\Delta$ (y-axis) according to $\Delta$ (x-axis).

In other words, note that if all the shortest transitions of the link stream are conserved in the graph series (in the sense above), so are all the minimal trips, and therefore, the possibilities of propagation in the dynamic network are unchanged.

Figure 2.9.1 left depicts the proportion of lost transitions as a function of the aggregation period $\Delta$, for the Irvine network. One can see that when the aggregation increases, starting from 1 second, the number of lost transitions first remains very low during several orders of magnitude, until an aggregation period of 0.5h where only 10% have been lost. The main part of the loss ($80\%$) is concentrated on the range between 0.5h and 235h, i.e. a bit more than 2 orders of magnitude. The saturation scale $\gamma = 18h$ returned by the occupancy method is in the beginning of this range, and in the middle in terms of order of magnitude. This shows that the occupancy method successfully detects the order of magnitude of the time scale from which the loss of information starts to be visible. For $\Delta = \gamma$, 48% of the shortest transitions are lost. Therefore, one may prefer to limit further the range of aggregation periods used, for example one order of magnitude below $\gamma$.

On the other hand, it must be clear that the measure of the loss used above is rather pessimistic. Indeed, some of the shortest transitions of the original link stream that are lost can be replaced by some others slightly longer or occurring a bit later. This limits the actual impact of this loss on the possibilities of propagation in the aggregated series. As lost transitions can be replaced, the duration of a minimal trip that was using some of these transitions may be only slightly altered by their loss (or even not at all). For this reason, we also use a measure of loss which is based on the elongation of minimal trips in the aggregated series $\mathcal{G}_\Delta$ compared to the original link stream $\mathcal{L}$.

**Definition 8** (Elongation factor)**.** *The* elongation factor *of a minimal trip* $P = (u, v, t_u, t_v)$ *of* $\mathcal{G}_\Delta$*, with*

$t_u \neq t_v$, *is defined as the ratio* $(t_v - t_u + 1).\Delta / time_{\mathcal{L}}(P)$, *where*

$$time_{\mathcal{L}}(P) = min\{t'_v - t'_u \mid (u, v, t'_u, t'_v) \textit{ is a minimal trip of } \mathcal{L} \textit{ and } t'_u, t'_v \in [(t_u - 1).\Delta, t_v.\Delta]\}$$

Note that when $t_u \neq t_v$, we necessarily have $time_{\mathcal{L}}(P) \neq 0$. Therefore, the elongation factor is properly defined. Figure 2.9.1 right gives the mean elongation factor (y-axis) of all minimal trips of the series aggregated with period $\Delta$ (x-axis), for the Irvine network. When $\Delta$ increases, the elongation factor of minimal trips first stays very close to 1 during several orders of magnitude, before it suddenly raises when the aggregation period reaches values around the saturation scale $\gamma$. This shows that our method properly determines the scale at which the properties of propagation of the link streams start to be altered by aggregation. For $\Delta = \gamma$, the mean elongation ratio of minimal trips is less than 1.5, showing that despite the 48% of shortest transitions lost, the propagation properties of the original link stream are not yet too drastically altered.

## 2.10 Discussion

Our contributions are the following:

- we showed that there exists a threshold, called the saturation scale $\gamma$, for the aggregation period of a link stream at which a qualitative change occurs in the way the network responds to aggregation,
- we empirically demonstrated that this change of behavior reveals an alteration of the properties of propagation of the dynamics, implying that dynamic networks should not be aggregated with a period larger than $\gamma$ to perform analyses that depend on these properties,
- we designed a fully automatic and parameter-free method to determine the value of $\gamma$ for an arbitrary link stream.

Our work open several perspectives to improve the method and broaden its field of application. The first of these perspectives is to extend the occupancy method to the case where links have a duration. The method presented in this article applies to both discrete and continuous time, to both undirected links and directed links, but it is able to deal only with links that are punctual events. However, in some contexts, the links of the dynamic network last during an interval of time (e.g. phone calls and physical contacts between individuals). Adapting the occupancy method to this case would be highly desirable. One particularly interesting way to do so would be to develop a notion of minimal trip that is specifically adapted to links that have a duration.

In Section 2.7, we pointed out a nice behavior of the occupancy method in presence of temporal heterogeneity in the activity of the link stream processed: the aggregation scale $\gamma$ returned in this case gives more importance to the parts of the dynamics that have a high level of activity, even if they do not occupy the majority of the time. Nevertheless, if these periods are really too short, they will have only a limited impact on the value of $\gamma$. As a consequence, these highly active parts of the link stream, which are likely to contain

a valuable information for the whole dynamics, may be smoothed out by the aggregation process. Avoiding this phenomenon by better taking into account the temporal heterogeneity of the activity of the link stream would constitute a key improvement. To this end, one could enhance the method so that it is able to separate the high activity periods from the lower activity periods and to determine an appropriate aggregation scale for each of these parts independently. Then one could decide either to aggregate the whole link stream at the shortest aggregation scale detected, which is the one that better preserves the information contained in it, or to partition the period of study and aggregate each part with a different length of window.

# 3

# Call Detail Records to Characterize Usage and Mobility Events of Phone Users

Cellular technologies are evolving quickly to constantly adapt to new usage and tolerate the load induced by the increasing number of phone applications. Understanding the mobile traffic is thus crucial to refine models. In this context, one has to understand the temporal activity and movements of users. At the user scale, the usage is not only defined by the amount of calls but also by his or her mobility. At a higher level, the base stations have a key role on the quality of service. In this chapter, we analyze a very large Call Detail Records over 12 months in Mexico (DS1). It contains 8 millions users and 5 billions of call events. Our first contribution is the study call duration and inter-arrival time parameters. Then, we assess user movements between consecutive calls (switching from a station to another one). Our study suggests that user mobility is pretty dependent on user activity. Furthermore, we show properties of the inter-call mobility by making an analysis of the call distribution.

## 3.1 Introduction

With the constant evolution of mobile technologies and digital networks, such as new generation of smartphones, and new applications, usage of cellular networks tends to change deeply. The analysis of phone calls from real logs is thus fundamental, both from phone operators and from other stakeholders' points of view. For the operators, it gives insights on the network usage and load, and consequently on possible dimensioning issues. It also allows to adapt or propose services according to the user trends. More generally,

mobile phone datasets allow to derive an analysis and statistics of human activities at a fine level of details. This unprecedented flow of continuous information on human activity represents a tremendous opportunity for research and real-world applications. Indeed, models or simulations that are used in order to study and dimension cellular networks, as queuing theory for instance, need to take into account the recent evolution of networks load and may progress by considering our new observations that concern the call duration and the inter-arrivals (time between two successive calls), users mobility, etc.

As we noticed in related work of the part I, human dynamics is an important dimension for comprehending the social behavior. In mobile data, studies that gives the possibility to follow each users temporally are very rare [37, 54]. In many cases, many studies do not have information about locations but use only locations of the calls in order to determine the density of population [44], the urban mobility [25, 123], distribution of favorite places [35]. However, it is not clearly how the phone event locations are correlated to the user mobility. In the paper [59], these authors study the user movement between two calls. This fine study on a small amount of users ($n = 56$) gives us insights about the inter-call mobility (ICM) of users. The ICM model represents a spatio-temporal probability distribution of users position in space and time between two consecutive communication records at distinct places. Here, instead of considering the whole trajectories, using DS1 presented in Section 1.1, we are estimating the user movement distribution from the call distribution based on the activity of around 7.7 millions of users during one year.

In this chapter, we focus on the analysis of this trace from the network/operator point of view. Contributions can be summarized through three items :

- First, a macroscopic analysis of communication dataset is performed. We show that activity, computed here as the number of calls per hour, varies at different scales. When the activity is seen as a signal, an empirical mode decomposition (EMD) allows us to derive its different cyclo-stationary components.

- Second, we assess phone usage and traffic properties through three different quantities: load, inter-arrival time between two calls and duration of a call. They are studied through two point of views: globally i.e. considering phone calls in the whole Mexico city, and per base station. For the load, we establish a landscape of the usage of the Base Stations (BS). For the inter-arrival and duration distributions, we confirm that the statistical traffic properties is the same from a Base Station to another, and also at the network scale. We compare these distributions to the classical distribution that is systematically considered in the models, the exponential law, and discuss its pertinence. It appears these very recent logs (2014) still leads to the classical exponential distribution at both scale (globally and on particular BS). For call duration, the distribution tail (the part that impacts performances in queuing system) fits by an exponential law.

- Third, the last part of this chapter is an analysis of user movements. This contribution is twofold. We model calls and users movement through two point processes. Whereas the first one is perfectly described by our dataset, the second one is unknown, except that a node movement is detected at the time of a call. Indeed, when a user changes of base stations, it does not appear explicitly in the logs,

but is detected only when a call occurs on the new BS. We show that for this kind of problems, application of Palm calculus theory [8, 146] offers relevant estimators for the second process (describing nodes movement). The use of this mathematical tool to the analysis of dataset, to our knowledge, is original. It applies to data that can be described with stationary point processes. In our context, it allows to derive: (i) an estimator of the number of calls per time unit, (ii) a simple test on the independence between the two processes (calls and movements), and (iii) an estimation of the movement distribution. It highlights the benefits of Palm calculus for data analysis to offer a formal framework to derive interesting and practical estimators even in presence of partially observable data. Numerical results based on this framework show that users mobility is correlated to the calls.

The chapter is organized as follows. In Section 3.2, we analyze our dataset (extracted from DS1 presented in Section 1.1). We also present the different results on the calls in time and space. Section 3.3 proposes a method to infer the statistical properties of the user movements, and presents the corresponding results. We conclude in Section 3.4.

## 3.2 Data analysis : temporal activity, inter-arrival time and call duration

For this study, we used DS1 presented in section 1.1 and extracted only geo-localized phone calls from the January 1, 2014 and ending on the December 31, 2014. For this period, we have more than 4.75 billions of calls. For 77% of call records, there is one location which determines the location of the phone user belonging to the telco company (either the callee or the caller). When both caller and callee are clients of the telco company, two locations are provided in our trace, one that notifies that the caller is calling the callee and the other indicating that the callee is receiving a call from the caller. These fully geo-localized calls represent around 6% of global internal calls in the country of Mexico. As in our study we focus on the user movements, we will mostly consider the geo-localized calls. We can notice in Fig. 3.2.1, the missing locations have the same activity as the ratio of geo-localized calls is quite constant over time. Therefore, it should not have any impact on our results. This subset of calls is representing the activity of 7,700,208 telco users during one year. In Fig. 3.2.1, we note that there are the same number of incoming and outgoing calls.

### 3.2.1 Temporal activity

The activity varies through time at several scales. During the day (from midday to 8pm), the activity is greater than during the night. The number of calls as function of the hours of the day (Figure 3.2.2) points out the typical period of lower activity during the night and greater activity during the day. Although the number of calls varies during the day, it varies between different days too. For instance, the activity during weekdays is greater than during the week-end. The peak is reached on Friday at 6 pm just after the end of the work.

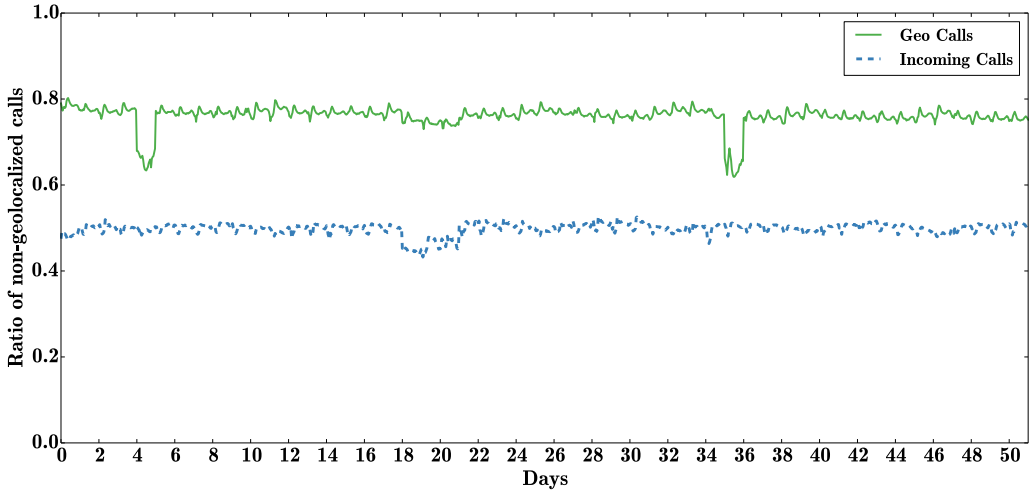



Figure 3.2.1: Ratio of the number of geo-localized calls (green line) and ratio of the number of incoming calls (dashed blue line) over 51 days. One can observe that there are the same number of incoming and outgoing calls. Geo-localized calls constantly represent around 76% of the calls. We will use these data for the experiments that follow.

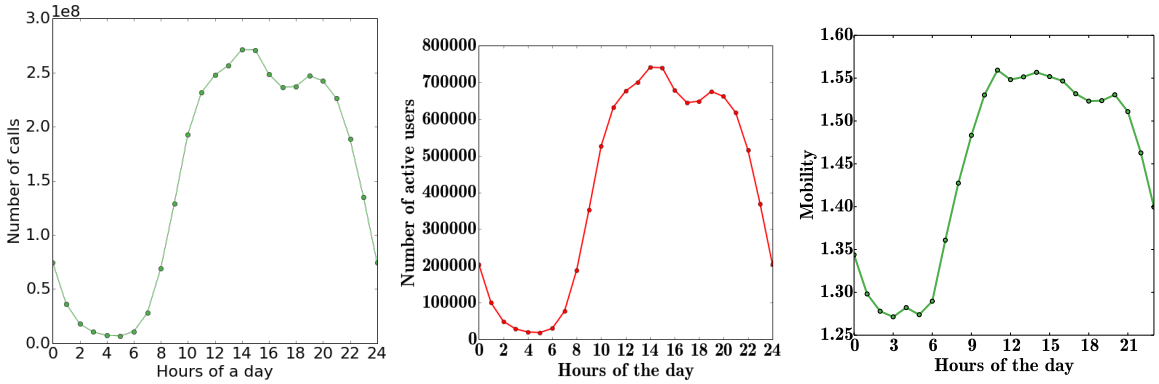



Figure 3.2.2: The number of calls (left) and the number of active users (middle) during the 1-year period as function of the hours of the day. We note a period of lower activity during the night and higher activity during the day. The peak is reached at 2pm. (right) The mobility defined as the average number of base stations reached within less than 30 minutes according to the hour of the day.

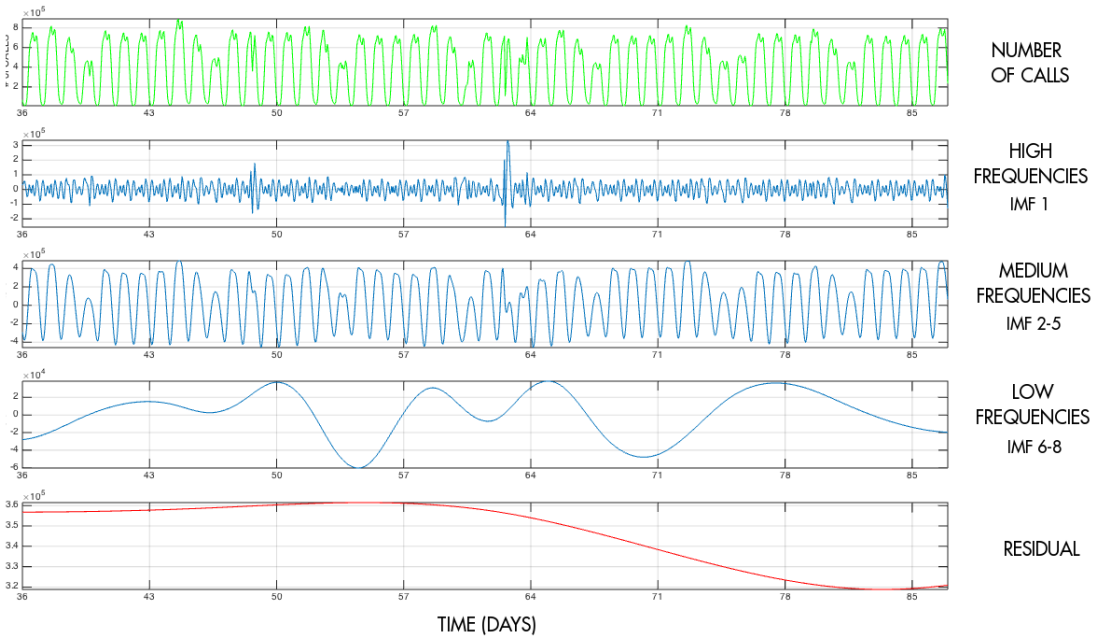



Figure 3.2.3: EMD of the signal linked to the number of calls per hour. From top to bottom, there is the original signal, high to low frequencies. One can clearly identify a day oscillation in the IMF 2-5. IMF 1 is high frequency variation and other IMF (6 to 8) are low frequencies.

As one can notice in Figure 3.2.2 (right), a mobile user tends to call more times in average than a static one. The mobility, corresponding to the average number of BS explored within half an hour, and the activity (left) are well correlated. This quick observation will be developed in section 3.3.

If we consider the activity as a signal, we can observe daily circadian patterns. People are organized on a daily base of 24 hours such that the activity signal will have statistical properties that vary cyclically with time and can be viewed as multiple interleaved stationary processes. To show this intuitive point, an Empirical Mode Decomposition (EMD) [130] is performed on the activity signal, the number of calls per hour during 51 days. The EMD allows to represent the non-stationary signal as sum of zero-means Intrinsic Mode Function (IMF) and one residue. Figure 3.2.3 gives the decomposition of the global call activity in high and low frequencies. The IMF 2 to 5 clearly gives a daily periodic signal (a spectral analysis also gives an harmonic decomposition in days of the signal) which validate the cyclo-stationarity of the activity signal and the fact that globally, people are used to call or not at the same moment of the day. The high frequency IMF 1 is also plotted on Figure 3.2.3. We plot in red the mean of the residual. The signal is clearly oscillating around the mean in a compact envelope with few extra peaks of activity. The low frequency signal is useful when one tends to detect special events and anomalies on the activity.

In a mobile data trace, a lot of measures are quite heterogeneous like the number of contacts, the number

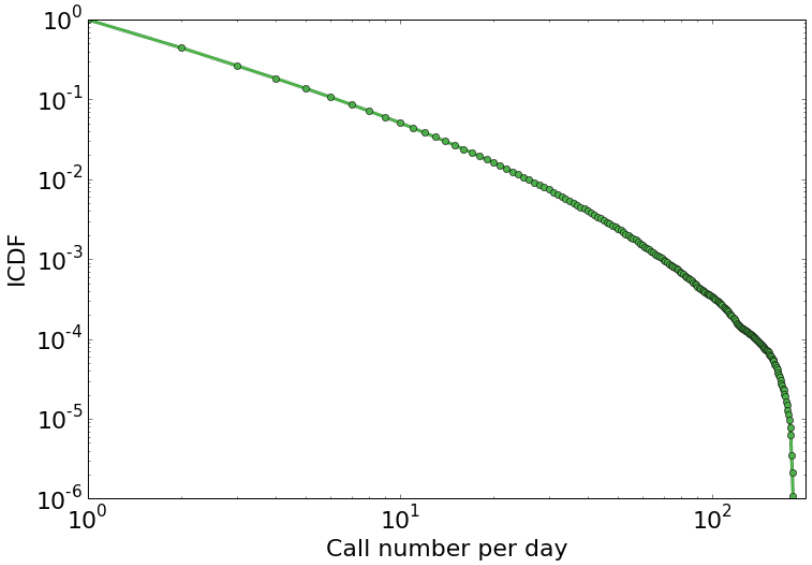



Figure 3.2.4: ICDF of the number of calls per user. The activity of users is heterogeneous, many people have few calls and some others have an active usage of the voice channel.

of calls and the time between two calls. We show the distribution of the number of calls per day (Figure 3.2.4). We note that around 25% of the users have more than two calls per day whereas 25% of users have less than 10 calls per week. Running a user movement study on a very long period of time will not make sense in such conditions of strong heterogeneity. Indeed, it is impossible to determine rather if the user changes precisely his location during a long inactive period. We do need a weak hypothesis on the stationarity of the signal. As we want to catch the movements of people during the day, we decided to cut all the signal (one year long) by slots of 2 hours. During each 2-hour period, we consider that the signal is stationary.

### 3.2.2 Call duration and inter-arrival time analysis

In the two next sections, we analyze two parameters : inter-arrival time between calls and call duration. These two quantities are the main input of queuing theory. The inter-arrivals describe the traffic nature, *i.e.*, the distribution of the clients arriving in the queue. The duration of a call is related to the service time of a client once it accesses to a resource. In our context, a resource is a couple slot-frequency or a set of resource blocks depending on the generation of cellular network we consider. In most of the queuing models, both inter-arrivals and call durations are supposed to be independently and exponentially distributed, leading to the famous $M/M/.$ queues. The reader can refer to [125], for a deeper presentation of queuing models applied to cellular networks. This assumption on the exponential distribution is common when considering phones traffic and call durations [168]. For the call duration, it is the distribution tail that is supposed to be exponential. Indeed, the first interval of the distribution is known as non exponential, because call durations

are usually lasting more than very few seconds. But, the exponential assumption still offers a good approach as it is the tail distribution, "the big clients", that impacts the performance of the system.

**INTER-ARRIVALS ON A BASE STATION.** When a user is calling someone, the origin and the destination are linked to a single BS. The attached BS are the first and last steps of the routing. In the trace, we only have the coordinates of the attached BS of the origin or the destination. As we miss many non-geolocalized calls, the measured activity of a BS is underestimated by a factor 10. As a consequence, the inter-arrival between two calls is overestimated by the same factor. Yet, the shape of the distribution may be the same.

In Figure 3.2.5a, we plot the inverse cumulative distribution function of the inter-arrivals. It corresponds to the time between two successive calls to a same BS. The distribution at the network scale, that gathers all geolocalized calls, is plotted in Figure 3.2.5a. It shows that the inter-arrivals range from $0$ to several hours. These very high values of inter-arrivals could correspond to periods where a BS is switched off (for maintenance or other reasons). Also, the figure shows that $99\%$ of the samples are less than $180$ seconds, and $80\%$ less than $21$ seconds. By considering all samples, we get very large range of values for which many have a small inter-arrivals. It corresponds to peaks of traffic during the day. On the opposite, great inter-arrivals are due to the night traffic. Nevertheless, these statistics usually help to dimension the network, which is usually performed with regard to the peak of traffic. We are thus interested to the traffic nature when the network is loaded. For these reasons, we perform the same statistic evaluation for specific BS and time ranges. We considered three particular BS at the peak of traffic. We have first ordered all the BS as function of their load and choose three BS (BS numbered $1175$, $157$ and $100$) that are respectively at $60\%$, $70\%$, and $90\%$ in this classification. The distributions are shown in Figure 3.2.5b. For these distributions, at least $80\%$ of the samples are less than $15$ seconds ($12$ times less than the case with all samples). The three distributions have been fitted with an exponential law, represented by the dotted lines in Figure 3.2.5c. Even if it does not match exactly, the exponential is very close to these distributions. The parameters of the exponential are $0.14$, $0.19$, and $0.21$ and correspond to the mean number of calls per second. The standard deviation errors of the fit is respectively $0.0005$, $0.0009$ and $0.0007$. The assumption on Poisson traffic is thus verified in our case.

**CALL DURATION.** Here, we propose a study on the duration of a call. For each call for which the destination replied, there is a duration in second. The duration of a call is one of the parameter that has a major impact on the load.

We plot this distribution from our trace, by extracting a single duration of a random call per user. Each user counts only for one in the distribution 3.2.7. In our trace, a long call is cut in several 10-minute calls. So, the distribution is ending at 10 minutes because the end of the tail is unknown. Apart from that, the ratio of long calls is quite small and so 10-minute sessions have a very small impact on the average and quartile results. We also noticed that there are more values when the number of seconds corresponds to a

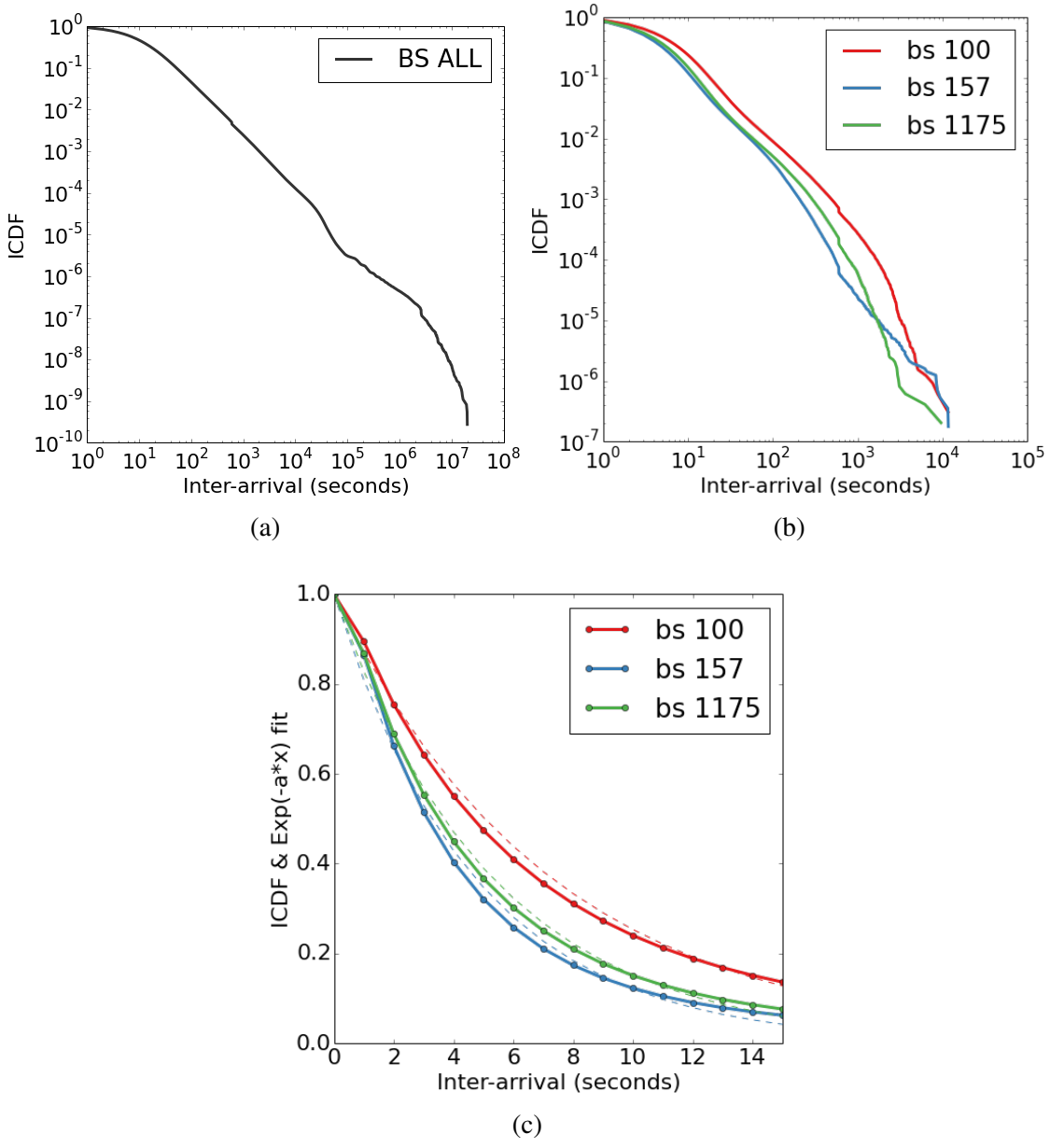



Figure 3.2.5: (a) For each BS, inter-arrival times between two consecutive calls are computed. The plot is obtained by merging all the distributions. (b) For 3 specific BS, that corresponds to the 40%, 30% and 10% more active BS (60%, 70%, and 90% in terms of load) the distribution of the inter-arrival time in second between two consecutive calls is plotted in log-log scale. (c) For the same 3 specific BS, the ICDF from 0 to 15 seconds is fitted by an exponential function (dashed lines). For practical reasons, x-axis is shifted by 1 second, we can so take the log as all values are strictly positive.

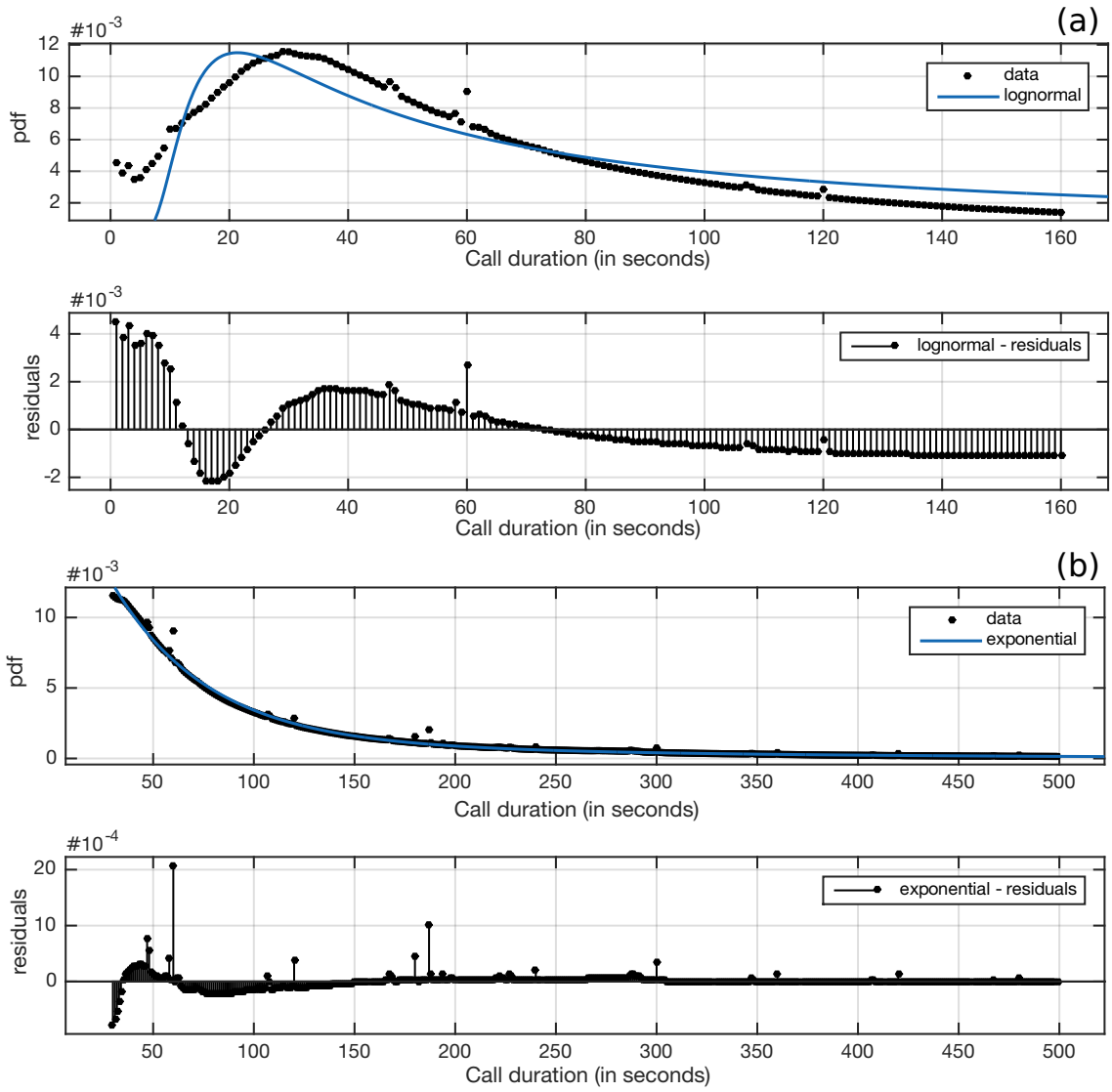



Figure 3.2.6: (a, top) Distribution of call duration in seconds fitted by a log-normal distribution in blue. (a, bottom) Residuals of the log-normal fit (b, top) Tail distribution of call duration in seconds fitted by an exponential distribution in blue. (b, bottom) Residuals of the exponential fit.

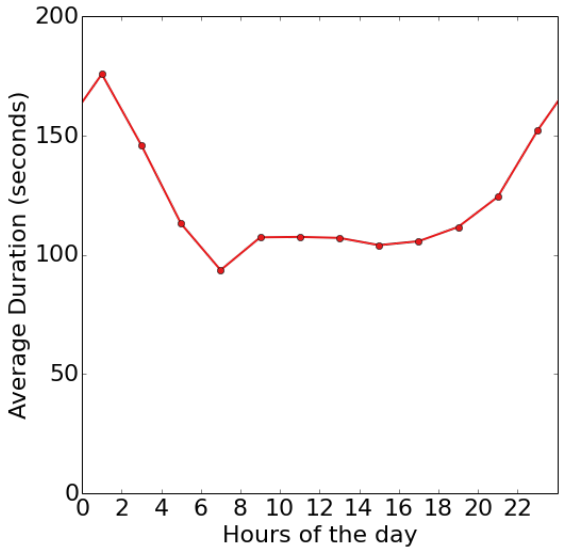



Figure 3.2.7: Average for the whole year of duration calls for each 2 hours slot

minute like 60s, 120s,... It is probably due to external artifacts like per-minute billing. The peak is reached for 34 seconds. The average duration of a call is 121 seconds and 25% of calls last more than 30 seconds whereas 75% last less than 2 minutes. All in all, 50% of calls take between 30 seconds and 2 minutes. In Figure 3.2.7a, the fit of the distribution with a log-normal function is quite good, the goodness of fit is $R^2 = 0.82$. The log-normal presents a gap with the empirical distribution for small values and for values in the tail. In Figure 3.2.7b, the fit of the distribution tail by an exponential distribution is very good as $R^2 = 0.99$. The residual between the exponential law and the empirical law is very small and confirms that the exponential distribution models perfectly the distribution tail as many studies already noticed it. This long tail induces an heterogeneity for the duration parameter, many durations are around 30 seconds and 2 minutes but some calls are still quite long.

In Figure 3.2.7, the time is divided in 12 slots of 2 hours each and the average duration is computed. From 6am-8am to 0am-2am, the average of the call duration is increasing. As the day is going, people tends to exchange more during a voice communication. Then during the night (2am to 6am) people who answer do not take the time for long conversations. The shortest durations are recorded between 6am and 8am. According to parts of the day, the duration is changing and the average can double from a slot time to another. This preliminary study on duration points out the fact that duration is not stationary and homogeneous but contains a lot of information that is useful to refine models or adapt performance of telecom companies. These starting observations may help to refine models and improve performance.

## 3.3 USER MOVEMENT ANALYSIS

Data collected describes sent and received calls of users. For each call, the localization of the BS associated to the user is known. It allows us to know the BS location at the time the calls are made. Based on this knowledge, we can study the statistical properties of the BS changes, *i.e.* the different times at which a user is associated to a new BS. It reflects users mobility between two calls and should be interesting for the telecom operator as it corresponds to user movements that it has to manage. Like in many CDRs, these times are only partially observable: we are able to detect that between two successive calls the user is not bound to the same BS but we do not know when it does happen exactly between these two calls.

In this Section, we propose two estimators. The first one describes the mean number of user movements per time unit, and the second one is related to the cumulative distribution function (CDF) of the time between user movements. Also, we propose a simple test that allows us to check if the two processes, calls and user movements, are dependent. The different computations and proofs rely on Palm calculus. This mathematical framework offers a set of tools on stationary point processes. The reader can refer to [8] for the definition and tools of Palm Calculus in $\mathbb{R}$, or [146] for a more pedagogic introduction and its application in $\mathbb{R}^2$. As it will be shown, Palm calculus is particularly adapted to this study.

A stochastic point process is a random variable. It can be seen as an ordered set of points distributed in $\mathbb{R}$. The observation of a set of events occurring at different times can thus be modeled through a stochastic point process. Therefore, calls and user movements can be modeled through two-point processes. They are represented in Figure 3.3.1. The first point process is denoted $N_{call}$. A sample represents the time of the calls for a user. At the time of a call, we know the BS which the user is bound. Formally, it can be seen as a mark associated to the point process $N_{call}$. In Figure 3.3.1, we used different patterns to represent the points of $N_{call}$ and its associated marks: a given pattern corresponding to a given BS/mark. For instance, when the user 1 is bound to BS $x$, the points/calls are depicted through black discs. When the user is bound to BS $y$, it is black ring, etc. A user movement is thus detected when the mark of $N_{call}$ changes. This marked point process is an exact representation/model of the data set in our possession.

The second process is $N_{BS}$ and is depicted through the vertical arrows in the figure. It represents a movement of the user, a change of the BS, between two calls. Our data set does not describe $N_{BS}$, but the marked process $N_{call}$ allows us to determine between which calls there was a BS change, or equivalently between which points of $N_{call}$ there is a point of $N_{BS}$. For instance, for user 1 in Figure 3.3.1, we observe a change of BS, from BS x to BS y, between the points/calls $T_i^{call}$ and $T_{i+1}^{call}$. Consequently, we infer the presence of a point of $N_{BS}$ between the points $T_i^{call}$ and $T_{i+1}^{call}$.

Formally, $N_{BS}$ and $N_{call}$ are random variables taking their values in the counting measures set on $(\mathbb{R}, \mathcal{B})$ (where $\mathcal{B}$ denotes the Borel $\sigma$-field of $\mathbb{R}$). We will use this definition in the different formulas, but as previously mentioned, it is more convenient to see a sample as a set of points (the support of the counting measure). A sample of $N_{call}$ and $N_{BS}$ can thus be seen as a set of points in $\mathbb{R}$, and correspond to the different time calls ($N_{call}$) and BS changes ($N_{BS}$) for a given user (a sample = a user).

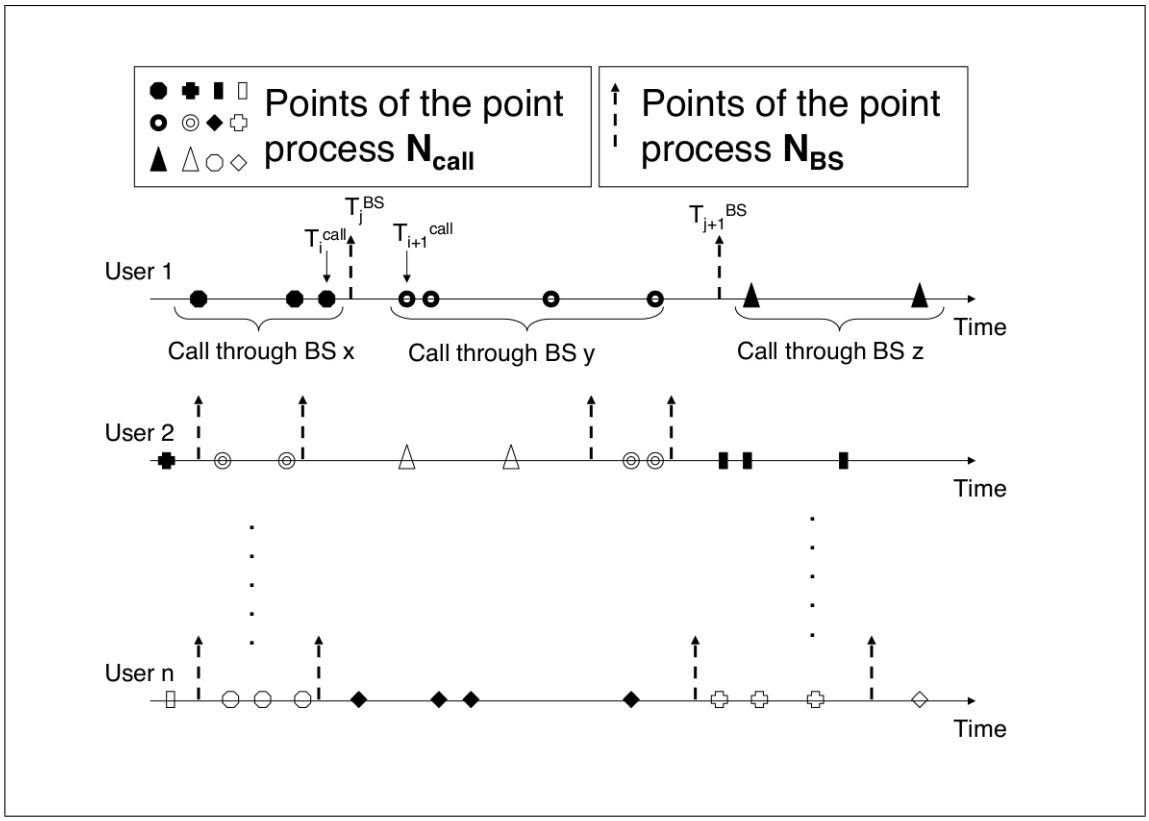



Figure 3.3.1: Description of the two-point processes $N_{call}$ and $N_{BS}$. The point process $N_{BS}$ is unobservable but the change of marks/BSs at the time calls (at points of $N_{call}$) allows us to know the intervals of $N_{call}$ where they are located and to derive statistical properties of $N_{BS}$.

A rapid analysis of the data showed that the process $N_{call}$ is not ergodic, *i.e.*, statistics made on a given sample do not allow to obtain convergent estimators. For instance, the mean number of calls per time unit are very different from a user to another. The different statistical estimators that are derived in this section are then systematically based on all samples/users. In other words, we do not make statistics as the average of the observable quantities on large period of times, but instead we consider an event for each user/sample, the time between two calls for instance, and we compute the average of this event over all users/samples. We assume that the two-point processes are stationary. From the statistical point of view, we assume that the process is stationary on the interval of times where the statistics are computed. In the numerical results, the statistics are then given for different periods in the day. We also assume that there is at most one point of $N_{BS}$ in an interval of $N_{call}$. It is thus seen as the movement of the user even if it may be composed in practice of several BS changes. The impact of this assumption on the results are discussed in the end of the section.

### 3.3.1 INTENSITY

The first quantity that is studied is the intensity of $N_{BS}$, denoted $\lambda_{BS}$, *i.e.* the mean number of BS changes per unit time. We propose an estimator $\widehat{\lambda_{BS}}$ of this quantity. Let $\Omega$ be the set of samples (our data set). The samples in $\Omega$ are assumed to be independent.

Our estimator is obtained through the application of Palm calculus. The points of $N_{call}$ (respectively $N_{BS}$) are denoted $(T_i^{call})_{i\in\mathbb{Z}}$ (respectively $(T_i^{BS})_{i\in\mathbb{Z}}$), in ascending order, and where $[T_0^{call}, T_1^{call}]$ (respectively $[T_0^{BS}, T_1^{BS}]$) is the interval that contains the origin. We apply the Neveu's exchange formula ([8] page 21) to the two-point processes $N_{BS}$ and $N_{call}$ for a function $f = 1$. We obtain:

$$\lambda_{call} = \lambda_{BS}\mathbb{E}^0_{N_{BS}}\left[\int_0^{T_1^{BS}} N_{call}(dx)\right] \tag{3.1}$$

$\mathbb{E}^0_{N_{BS}}[.]$ is the Palm expectation with regard to the process $N_{BS}$. Palm expectation, or Palm measure, may be seen as the probability measure under the condition that there is a point of the point process at the origin. The point process indexed under the expectation notation $\mathbb{E}^0_{N_{BS}}$ ($N_{BS}$ here) indicates which point process is supposed to have a point at the origin. It is worth noting that quantities under the classical and Palm expectation lead to different values. For instance, $\mathbb{E}\left[T_1^{BS}\right]$ and $\mathbb{E}^0_{N_{BS}}\left[T_1^{BS}\right]$ differs. $\mathbb{E}\left[T_1^{BS}\right]$ is the time from the origin (an arbitrary time) to the next BS change. It is thus the residual time to the next BS change. $\mathbb{E}^0_{N_{BS}}\left[T_1^{BS}\right]$ is the time between two BS changes. Indeed, under Palm expectation we know that there is a point of $N_{BS}$ at 0 and we evaluate the time to the next BS change $T_1^{BS}$.

In Equation (3.1), remind that $N_{call}(.)$ is a counting measure, and $\int_0^{T_1^{BS}} N_{call}(dx)$ is thus equal to $N_{call}([0, T_1^{BS}])$. Under Palm expectation, $\int_0^{T_1^{BS}} N_{call}(dx)$ is thus the mean number of points of $N_{call}$ between two successive points of $N_{BS}$. Consequently, this quantity can be fully determined/estimated based on our samples as we do not need the exact location of the points of $N_{BS}$. For instance, in Figure 3.3.1, a sample of this quantity (for user 1) is the number of points of $N_{call}$ between points $T_j^{BS}$ and $T_{j+1}^{BS}$ (equals to 4). The estimator is then:

$$\widehat{\lambda_{BS}} = \frac{\widehat{\lambda_{call}} card(\Omega)}{\sum_{N_{call}\in\Omega} N_{call}([T_i^{BS}, T_{i+1}^{BS}])} \tag{3.2}$$

where $\widehat{\lambda_{call}}$ is an estimator of $\lambda_{call}$.

Equation (3.2) corresponds exactly to Equation (3.1), but writen in a simpler form. Here, we pick one interval of $N_{BS}$ for each sample. The value of $i$ does not matter and may be different from one sample to another. Due to the stationarity constraint, we divide times of the day to slots of 2 hours. The estimation of $\lambda_{BS}$ is then performed independently for each slot. We take one sample for each user and each day. Results are shown in Table 3.3.1. The number of considered samples is shown in the last column of the table. With the constraint of 2 hours, all samples are not taken into account. Indeed, we consider only samples with at least two movements, otherwise it is obviously impossible to apply the method. The results show that the number of movements stays more or less constant during all days. With the filter that we apply on the data

set, the results tend to show that, in average, a user moves rarely more than two times on these slots. Clearly, our method leads to an over estimation of the real intensity $\lambda_{BS}$. But, a classical method consisting in evaluating the time between two movements with the same constraint on the stationarity should lead exactly to the same problem. Moreover, in our study, we do not have the exact time between two movements, such an approach would consequently be impossible.

| Hours | $\lambda_{BS}$ (second) | $\lambda_{BS}$ (hour) | $card(\Omega)$ |
|---|---|---|---|
| 0-2AM | 0.00058 | 2.09 | 1 044 566 |
| 2-4AM | 0.00061 | 2.19 | 265 693 |
| 4-6AM | 0.00061 | 2.18 | 129 196 |
| 6-8AM | 0.00060 | 2.17 | 215 682 |
| 8-10AM | 0.00058 | 2.07 | 1 676 401 |
| 10-12AM | 0.00059 | 2.11 | 6 899 027 |
| 12-14PM | 0.00059 | 2.14 | 9 543 073 |
| 14-16PM | 0.00058 | 2.10 | 10 545 188 |
| 16-18PM | 0.00057 | 2.07 | 8 899 166 |
| 18-20PM | 0.00058 | 2.07 | 8 409 011 |
| 20-22PM | 0.00056 | 2.01 | 7 574 970 |
| 22-24PM | 0.00056 | 2.01 | 4 058 205 |

Table 3.3.1: Results for the estimation of $\lambda_{BS}$.

### 3.3.2 Dependency test

An important assumption to estimate the distribution of the time between two successive points of $N_{BS}$ is the dependency between the two processes $N_{BS}$ and $N_{call}$. A formal hypothesis test is impossible to perform as $N_{BS}$ is not fully observable. Therefore, we propose a simple test, based on the length of the intervals $[T_i^{call}, T_{i+1}^{call}]$ where the points of $N_{BS}$ are located, to infer the dependency between the two processes.

According to Palm Calculus, if we pick a point $X$ in $\mathbb{R}$ independently of a stationary point process, e.g. $N_{call}$, this point will be likely located in a "big interval". More precisely, the mean of the interval length $[T_i^{call}, T_{i+1}^{call}]$ where $X$ is located will be greater than the mean size of the interval of the point process ($\frac{1}{\lambda_{call}}$ here). Intuitively, as "big intervals" occupy more space, $X$ is likely located in one of them. For instance, with a Poisson point process the mean interval length where $X$ is located is two times greater than the other intervals (in average). It is the famous Feller paradox ( [8] pages 33 and 295). As the process is stationary, pick a random point $X$ or a fix point leads to the same results. By convenience we consider the origin. The interval where the origin is located is $[T_0^{call}, T_1^{call}]$, and its mean length is equal to $\mathbb{E}[T_1^{call} - T_0^{call}] = 2 \cdot \mathbb{E}[T_1^{call}]$ (consequence of the stationarity of the point process). The mean length of this interval depends on the distribution of the process, but it can be easily calculated with the Palm inversion formula ([8] page 20). We give below the computation details but it is a classical result of Palm calculus (see [89] for instance where it is applied to a mobility study). In the first equation below, $\theta_t$ is the shift

operator. Here, it shifts the points of $N_{call}$ of a time $t$ (meaning that $N \circ \theta_x(C) = N(C - t)$ for an interval $C$ in $\mathbb{R}$, or more formally $C$ in $\mathcal{B}$). We get:

$$\mathbb{E}\left[T_1^{call}\right] = \lambda_{call}\mathbb{E}^0_{call}\left[\int_0^{T_1^{call}} T_1^{call} \circ \theta_t dt\right] \tag{3.3}$$

$$= \lambda_{call}\mathbb{E}^0_{N_{call}}\left[\int_0^{T_1^{call}} (T_1^{call} - t)dt\right] \tag{3.4}$$

$$= \frac{\lambda_{call}}{2}\mathbb{E}^0_{N_{call}}\left[\left(T_1^{call}\right)^2\right] \tag{3.5}$$

If $N_{BS}$ is independent of $N_{call}$, a point of $N_{BS}$ behaves as the random point $X$ presented earlier or the origin. Therefore, if the two processes are independent the mean interval lengths (of $N_{call}$) where the points of $N_{BS}$ are located must equal to $2 \cdot \mathbb{E}[T_1^{call}]$. Formally, in case of independence, we get:

$$\mathbb{E}\left[T_1^{call} - T_0^{call} | N_{BS}([T_0^{call}, T_1^{call}]) > 0\right] = \mathbb{E}\left[T_1^{call} - T_0^{call}\right] \tag{3.6}$$

$$= \lambda_{call}\mathbb{E}^0_{N_{call}}\left[\left(T_1^{call}\right)^2\right] \tag{3.7}$$

Let denote $\mu_1 = \mathbb{E}\left[T_1^{call} - T_0^{call} | N_{BS}([T_0^{call}, T_1^{call}]) > 0\right]$ and $\mu_2 = \mathbb{E}^0_{N_{call}}\left[\left(T_1^{call}\right)^2\right]$. We can use the confidence interval of these two expectations (both sides of Equation (3.6)) to accept the assumption on dependency with a certain probability ($(1-\alpha)^2$ in the proposed method). With our assumptions (i.i.d. samples), a confidence interval of $\mu_1$ at $1 - \alpha$ is given by:

$$\left[\bar{X_1} - z(\alpha)\frac{S_1}{\sqrt{n_1}}, \bar{X_1} + z(\alpha)\frac{S_1}{\sqrt{n_1}}\right] \tag{3.8}$$

where $\bar{X_1}$ is the expectation evaluated from the samples (intervals of $N_{call}$ where there was a BS change), $S_1$ is its standard deviation [1], $n_1$ is the number of samples, and $z(\alpha)$ depends on the parameter $\alpha$ (such that $\mathbb{P}(N \in [-z(\alpha), z(\alpha)]) = 1 - \alpha$ where $N$ follows a normal distribution $\mathcal{N}(0, 1)$). Obviously, the same holds for $\mu_2$, but as we have to consider $\lambda_{call} \cdot \mu_2$ instead, we get:

$$\left[\lambda_{call}(\bar{X_2} - z(\alpha)\frac{S_2}{\sqrt{n_2}}), \lambda_{call}(\bar{X_2} + z(\alpha)\frac{S_2}{\sqrt{n_2}})\right] \tag{3.9}$$

If the two intervals do not overlap then the probability that the two quantities $\mu_1$ and $\lambda_{call} \cdot \mu_2$ are different is greater than $(1 - \alpha)^2$. In other words, the probability that the two processes are dependent is greater than $(1 - \alpha)^2$. Otherwise, we cannot conclude to dependence or independence. It is worth noting that these two quantities do not depend on the exact locations of the points of $N_{BS}$ but only on the interval lengths of $N_{call}$ available from our data set.

[1] As the standard deviation is unknown, it is given by $\sqrt{\frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X_1})^2}$ where $X_i$ are the samples.

The results are shown in Table 3.3.2. Before describing the results, we give some elements on the method we followed. We considered intervals of two hours during the day to obtain intervals where the two-point processes are assumed stationary. Each temporal window was processed independently. For each user, we draw randomly one of the movements and we measured the interval $[T_i^{call}, T_{i+1}^{call}]$ where it lied. The result is the column "Movement interval" in the table. Beside, we selected an interval $[T_i^{call}, T_{i+1}^{call}]$ randomly chosen for each user and estimate these two first moments (compute as the average over all users). It leads to estimators of $E^0_{N_{call}}[T_1^{call}]$ and $E^0_{N_{call}}[(T_1^{call})^2]$, from which we deduce $E[T_1^{call} - T_0^{call}]$ (equal to two times Equation (3.5)). The selection of users making calls can have an impact only in case of correlation between the two point processes. It does not impact the result as the test is only able to valid correlation.

In Table 3.3.2, we can observe, as expected, that user movement happens in interval with a greater length in average with regard to $E^0_{N_{call}}[T_1^{call}]$. But, their mean lengths should equal to the $4^{th}$ column. We can observe a difference of approximately $30\%$ between these two quantities. With the number of samples used in the different computations, that are given in the last columns, the confidence intervals are close to $0$ for all these estimators, and so does not explain the gap. According to the test presented above, the two point processes are dependent with a probability of $0.9$ as $\alpha = 0.05$. This dependency confirms the temporal correlations we have between mobility and the activity in Figure 3.2.2. A possible interpretation of this phenomena, is that mobile users may call before a departure, at their arrival, or during the path, and consequently are likely to call when they are in movement or just after/before a movement. This result may present a bias as we do not know the number of BS changes between two calls. Indeed, several BS changes may happen between two successive calls. Therefore, our choice of the intervals with a BS change would be different if the number of movements is very different from an interval to another. Intuitively, in this case, we should more likely choose an interval with a great number of movements than an interval with only a small one.

| Hours | $E^0_{N_{call}}[T_1^{call}]$ | Movement interval | $E[T_1^{call} - T_0^{call}]$ | number of samples |
|---|---|---|---|---|
| 0-2AM | 591.79 | 942.25 | 1324.77 | 35058 |
| 2-4AM | 578.03 | 923.05 | 1377.29 | 9280 |
| 4-6AM | 564.49 | 1007.60 | 1453.25 | 4531 |
| 6-8AM | 565.61 | 1073.24 | 1515.29 | 8445 |
| 8-10AM | 634.54 | 1152.25 | 1632.98 | 62660 |
| 10-12AM | 719.78 | 1252.82 | 1797.30 | 193054 |
| 12-14PM | 727.06 | 1277.26 | 1825.02 | 237929 |
| 14-16PM | 718.87 | 1275.12 | 1816.50 | 260823 |
| 16-18PM | 716.41 | 1298.34 | 1827.37 | 218267 |
| 18-20PM | 715.20 | 1264.40 | 1810.68 | 218915 |
| 20-22PM | 687.99 | 1242.35 | 1741.00 | 200104 |
| 22-24PM | 655.38 | 1128.19 | 1628.88 | 118539 |

Table 3.3.2: Results on the dependency test.

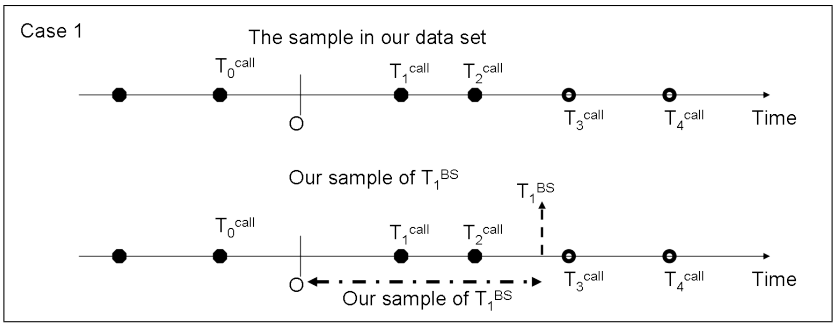



Figure 3.3.2: Case 1: the first point of $T_1^{BS}$ is in the interval $[T_2^{call}, T_3^{call}]$. The sample of $T_1^{BS}$ is then uniformly distributed in this interval.

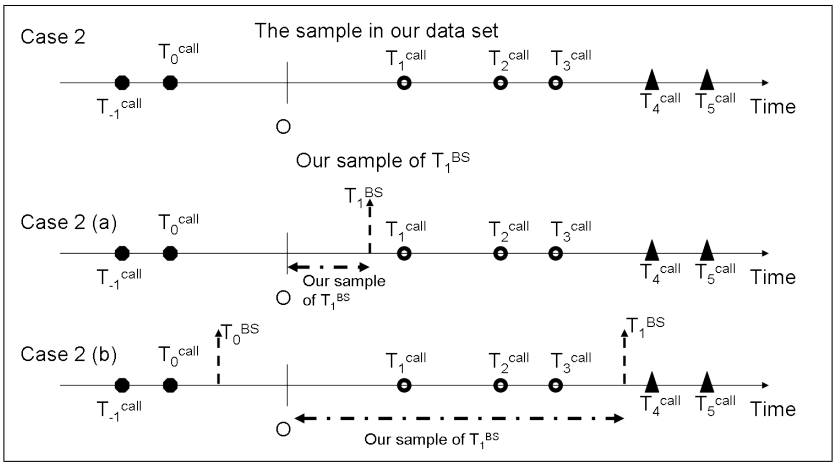



Figure 3.3.3: Case 2: there is a point of $N^{BS}$ in the interval $[T_0^{call}, T_1^{call}]$. We draw a point of $N_{BS}$ uniformly in this interval. It leads to two sub-cases: (Case 2(a)) if the point is in $[O, T_1^{call}]$, then we consider it as our sample of $T_1^{BS}$, (Case 2(b)) if the point is in $[T_0^{call}, O]$ then it corresponds to $T_0^{BS}$, so we look for the next interval that hosts a point of $N_{BS}$ ($[T_3^{call}, T_4^{call}]$ in the figure) and we distribute uniformly our sample of $T_1^{BS}$ in this interval.

### 3.3.3 DISTRIBUTION

In this section, we describe a method to obtain estimations of the distribution of $N_{BS}$. We assess the cumulative distribution function (CDF) of $T_1^{BS}$ under the classical probability measure $\mathbb{P}\left(T_1^{BS} \leq x\right)$ and Palm measure $\mathbb{P}^0_{N_{BS}}\left(T_1^{BS} \leq x\right)$. Under the Palm measure, it describes the distribution of the time between two successive movements. Under the classical measure, it is the time to the next movement: given a user at an instant $t$, it is the time to the next movement.

We do know the intervals where the points of $N_{BS}$ are distributed. In each of these intervals, we draw the point $N_{BS}$ uniformly. It would correspond to the real distribution in case of independence of the two processes. But, as we have seen in the previous Section, independence does not hold here and our method

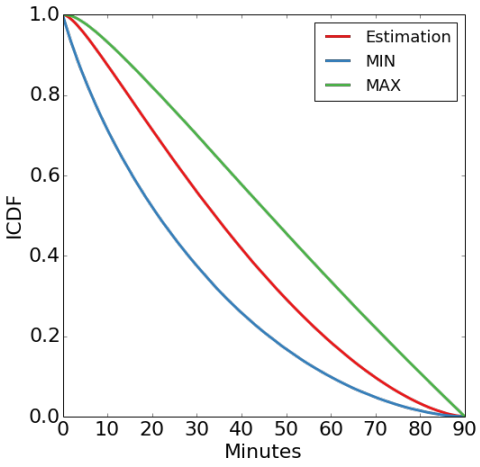



Figure 3.3.4: ICDF of $T_1^{BS}$.

is thus not exact.

From these samples we compute the empirical estimator of $\mathbb{P}\left(T_1^{BS} < u\right)$.

We detail below the method. A set of examples is given in Figures 3.3.2 and 3.3.3. The method:

- We set a common time $t$ as the origin for all our samples. It is chosen arbitrarily and independently of the two processes. It is denoted $O$ in the figures.
- To collect samples of points $T_1^{BS}$, we proceed as follows for each sample/user:
  - If the interval of $N_{call}$ that contains the first point of $N_{BS}$ is $[T_i^{call}, T_{i+1}^{call}]$ with $i > 0$, then we draw our sample uniformly in this interval. This case is illustrated in Figure 3.3.2 (Case 1).
  - If the interval of $N_{call}$ that contains the origin, $[T_0^{call}, T_1^{call}]$, hosts a point of $N_{BS}$, then we draw a point uniformly in $[T_0^{call}, T_1^{call}]$. If it belongs to $[0, T_1^{call}]$ then we select this point as our sample (Figure 3.3.3 - Case 2(a)). Otherwise, the point that is obtained is $T_0^{BS}$ and not $T_1^{BS}$. Consequently, we consider the next interval of $N_{call}$ ($[T_i^{call}, T_{i+1}^{call}]$ with $i > 0$) that contains a point of $N_{BS}$. Our sample is then the point uniformly distributed in this interval (Figure 3.3.3 - Case 2(b)).
- From the collected samples, we calculate the empirical distribution of $T_1^{BS}$ *i.e.* $\mathbb{P}\left(T_1^{BS} \leq x\right)$.

We also consider a lower and upper bound on the values of the samples that allows to bound the real distribution of $T_1^{BS}$. For the lower bound, we consider for each sample the beginning of the interval $[T_i^{call}, T_{i+1}^{call}]$, thus $T_i^{call}$. For the upper bound we consider $T_{i+1}^{call}$.

The inverse cumulative distribution function (ICDF) is shown in Figure 3.3.4. The ICDF under the classical expectation, shows that user movements occurs between $0$ and approximately $5400$ seconds. The proposed estimation thus offer an interesting trade-off between the two bounds. The two bounds do not

present negligible differences with the approximated distribution. It can reach up a difference of $0.2$ for the lower bound, and $0.15$ for the upper bound.

## 3.4 DISCUSSION

This chapter presented an analysis of calls in a cellular network from a CDR trace. In the first part:

- we assess the statistical properties of these calls. We exhibit a cyclo-stationarity of the number of calls per hour, with a lightweight different behavior in the week-end
- the distribution obtained for call durations and inter-arrivals have shown that the classical exponential distribution still fit to the empirical one. It confirms the classical assumptions on phone traffic. Moreover, our study gives example of current loads observed in cellular networks that can be considered as input in queuing models.

In the second part, we have proposed:

- a method to study user movements using Palm calculus. This theory offers a formal mathematical framework to obtain estimator on user movements. we have proposed methods to estimate the intensity of user movements,
- a dependency test that allowed us to check if calls and movements are correlated, and a method to generate samples of user movements. A required property to apply this theory is that the considered processes must be stationary. As it is clearly not the case for our data set, we had to consider range of $2$ hours. It led to a proportion of samples with no movements that could not be taken into account, and thus an overestimation of user movements. Also, for moving users, the proposed dependency test seems to show that their movements are correlated to their calls.

Results of our nationwide data analysis may be used in different ways. It can help to consider practical parameters in simulations and models. Results on the dependency between calls and movements still need to be improved. A more detailed characterization of this dependency could help to propose models able to generate joint calls and movements distribution. Also, we point out that our results on the call durations contain a lot of information by its variability through time and users. This quantity can help to improve models that describe social relationship between users. Taking advantage of this parameter can lead to identify people, define contacts between phone users, detect communities or predict links. A more fundamental work could consist in extending this study to non-stationary point process. The question is: may we rely on the cyclo-stationarity of the processes to derive equivalent estimators from Palm Calculus but applied to the full periods (complete weeks, months, or year).

# 4

# Performance evaluation of Delay-tolerant networking (DTN) protocols to deliver SMS in dense mobile network : empirical proofs

In dense areas, the density of mobile users is great. People are moving from home to work, from work to active places. The mobility of users, especially during the day, creates a DTN mobile network where the nodes are the smartphones held by mobile clients. In this chapter, we perform a temporal and spatial analysis of the Mexico City cellular network considering geo-localized SMS to characterize the traffic. Such key characterization allows us to answer the question: is it possible to transmit SMS using phones as relay in a large city such as Mexico City? Thus, we propose several DTN (Delay Tolerant Network) like basic network protocols for delivering SMS. We define four network protocols to transmit SMS from a source to a destination. We study a mobile dataset including 8 millions users living in Mexico city. This gives us a precise estimation of the average transmission time and the global performance of our approach. Our analysis shows that after 30 minutes, half of the SMS are delivered successfully to destination. In contrast to the cellular networks, we explain how much the potentiality of the mobile users network can take benefit from complementary properties such as the locality of SMS, the density of phones in Mexico City and the mobility of phone users. Moreover, we show that in a realistic scenario, our approach induces reasonable storage cost.

## 4.1 INTRODUCTION

As we noticed, the need of communicating in a dense city is always increasing. Every day, millions of SMS are sent in a large city like Mexico City. Even if SMS is challenged, SMS communication is now also used for many services and the list of uses keep increasing. It is now possible to use SMS to make health campaigns [42, 61, 81], schedule calls, conduct electronic surveys, provide e-voting services, send calendar notifications, search the Internet, and exchange status updates with servers on the Internet. Therefore, during rush hours, SMS traffic may consume a non negligible part of the backbone network capacity, and sometimes saturates it. This saturation may come from the SMS architecture itself which is totally centralized. Every SMS is delivered to a unique SMS CENTER (SMSC) which acts as a centralized, store-and-forward server that is responsible for accepting, storing, retrieving subscriber information, and forwarding SMS to the intended destination of the SMS.

It is becoming a great challenge to increase the amount of traffic delivered to the users while keeping the infrastructure stable (i.e., same number of relays and backbone capacity). Apart from the typical rush hours, the mobile network can be globally challenged during special events such as natural disasters or locally saturated during concerts, conferences, riots or sport matches. Mobile networks are dimensioned to sustain the load 99% of the time, but for those specific events, the activity can be incredibly higher than the expected traffic during rush hours. In [36], authors revealed that voice calls and SMS are still in use in these large scale events, despite frequent reports from users about the data network unavailability. It can thus be interesting to propose and test new protocols, less dependent of the cellular infrastructure and/or the backbone, that could carry a part of the traffic load.

In this study, we evaluate the benefits and the feasibility of a delay tolerant network (DTN) approach to transmit SMS and more generally data from a source to a destination. Instead of using classical routing, we use relays close to the source and phone users that are connected to those local relays to reach the destination. The advantage of our approach is that we do not perform a routing algorithm, we do not need global knowledge, neither its associated mechanisms such as neighbor discovery, exchange of control messages, etc. and do not need to know where the destination is. Moreover, as we only use local relays that are close to the source, the bandwidth cost of a SMS is smaller and the backbone infrastructure of the operator is not used. On top of that, our protocol works better when the capacity of the network is challenged during rush hours as the density of phones and the mobility of users are even higher.

It is important to notice that we clearly do not target an implementation of our approach in existing 3GPP standards or existing cellular networks protocols. Our goal is to demonstrate that the DTN approach could be feasible and helpful. The replay of millions of real SMS traffic shows the reliability and the gain of our proposals. Note that such approach could be used in future cellular standards, and also be used as a key enabler for P2P applications.

The contributions of this chapter are two-fold :

- First, we perform an analysis in time and space of SMS traffic of DS1. The temporal analysis includes the evolution of the SMS activity at different time scales. The spatial analysis is based on the distance

between source and destination in kilometers and number of hops. The number of hops is computed according to the Voronoï diagram of the base stations.

- Thanks to the traffic characterization and more precisely its regularity and locality, the second contribution is the proposal of four protocols that aim to carry SMS traffic, and to relieve the infrastructure network in terms of load. We use the same real SMS traces to evaluate the efficiency of these less centralized protocols. Through the replay of these SMS, we show that with a very simple and not optimized algorithm, it is possible to have a delivery ratio higher than 50%. It is obviously not perfect, but we consider that it is enough to prove the benefit of our approach. It is also worth noting that this rate is an underestimation as our dataset is partial.

The chapter is organized as follows. In Sections 4.2, we present our large trace that contains SMS and localized calls and we perform a spatial and temporal analysis to point out some pertinent characteristics of the cellular networks. Two basic DTN protocols are presented in Section 4.3. In Section 4.4, we show the performance evaluation of these two protocols in terms of transmission success and delay, and point out how much higher activity and mobility can increase the success rate of our protocols. We deeper analyze the complementarity of our protocols by mixing and extending them in Section 4.5. In Section 4.6, we discuss the experimental choices and estimate the storage cost of protocols. We conclude in Section 4.7.

## 4.2 Data Analysis

### 4.2.1 Data extraction and validation

**Data extraction.** For this study, we extract from DS1 a dataset containing only SMS (no geo-located) and geo-localized calls where origin and destination are in Mexico City from March to April 2014. Mexico City is covered by 775 base stations that are part of the telecommunication network of a cellular operator. From the geo-localized mobile calls, it has been possible to localize in Mexico City 1.5 millions of SMS for the study. For each SMS, we set the localization of the closest call in time. We assume that if the source and the destination received or made a localized call 30 minutes before or after the SMS, we can effectively set a source and destination location for the SMS. We also noticed that more than 92% of geo-localized SMS sent from Mexico City have the destination in Mexico City.

For our study, we define a graph with the base stations as nodes, for which there is a link/edge between two nodes if they are neighbors in the Voronoï tessellation: there is a link between two base stations $(bs_1, bs_2)$ if and only if $bs_1$ and $bs_2$ have a common border in the Voronoï tessellation built according to the base station locations. Consequently, we can define a distance between two base station not only in terms of kilometers but also in terms of hops through this graph. Two neighbors are at distance 1, the neighbors of the neighbors are at a distance 2, and so on. As we do not have any hierarchical information on base stations, we consider that every base station has the same role.

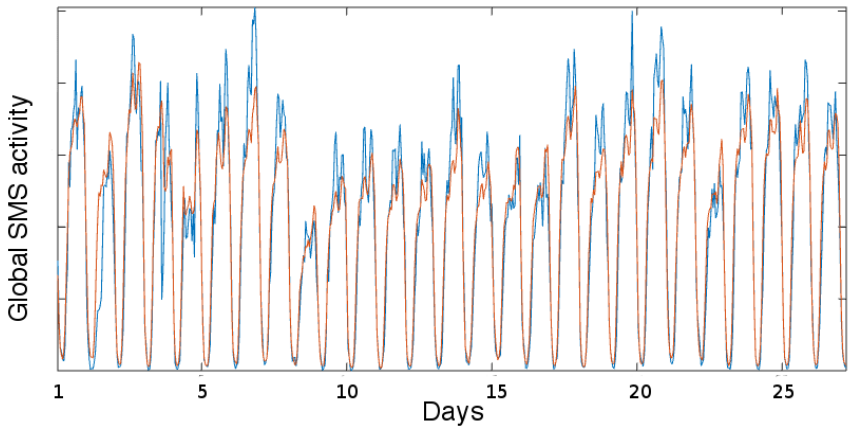



Figure 4.2.1: Global SMS activity per hour in Mexico City for 28 days (March 1 to March 28 in 2014) (in blue) and the geo-localized sampling for which the locations of the source and the destination are known for the same 28-day period (in red). These curves are normalized by the mean activity per day.

**VALIDATION.** As we take a sampling of SMS for which we have the locations of the source and the destination, it is important to test that our sample is representative of the global trace. In Figure 4.2.1, we show the global activity and the sampled activity. Both are normalized by their mean. It seems at a first glance that both signals of the activity per hour are similar. The Pearson correlation coefficient of the two time series is equal to 0.95, a linear equation describes the relationship between both signals perfectly. As the $p$-value associated is null, the computed correlation is significant. We also confirm this validation by a cross-correlation of two discrete-time signals. It shows a strong similarity between both signals.

### 4.2.2 TEMPORAL AND SPATIAL CHARACTERISTICS

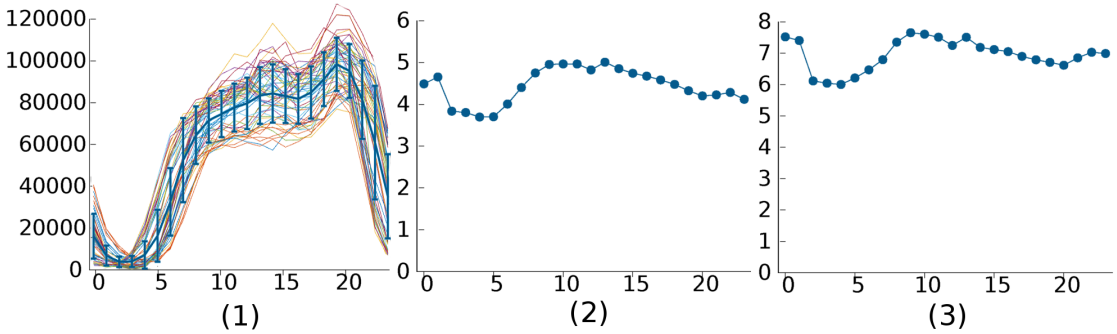



Figure 4.2.2: For each hour of the day, we show the average of three parameters over the 2-month period: (1) average and standard deviation of the number of SMS (bold blue) and distributions for all days (multi-color); (2) average distance in hops of SMS; (3) average distance in km of SMS.

In this section, we make an analysis of the localized SMS according to three parameters: the activity (number of SMS sent), the distance in kilometers and the distance in hops (distance in hops from the source

to the destination of the SMS in the cellular graph, defined by Voronoï diagram, from source base station and destination base station). These parameters are computed according to the time (per hour) and the space (per cell). The temporal analysis consists in studying the number of SMS and the distance of SMS per hour. To deeper analyze the traffic patterns, we use an empirical mode decomposition (EMD) method. Using the EMD method, the time series can be decomposed into a finite and small number of components. These components form a complete and nearly orthogonal basis for the original signal. Without leaving the time domain, EMD is adaptive and efficient. Since the decomposition is based on the local characteristic time scale of the data, it can be applied to nonlinear and non-stationary processes. Once the time series is decomposed, we are able to detect traffic behavior that does not fit the general periodicity and extract the high frequency component that gather abnormal events. The spatial analysis is based on the Voronoï diagram from base stations, and aims to show up the spatial diversity.

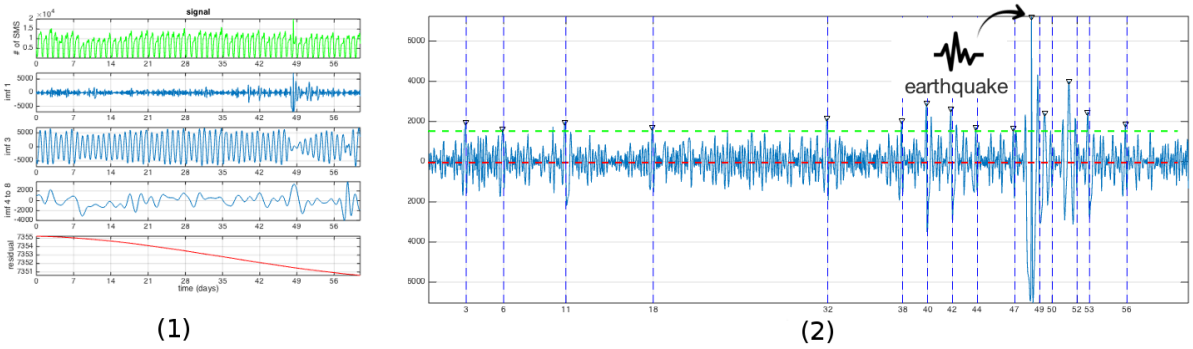



Figure 4.2.3: (1) EMD of the original temporal signal: number of SMS sent per day. One can clearly identify a day oscillation in the IMF 3. IMF 1 is high frequency variation and other IMF (4 to 8) are low frequencies, (2) For IMF 1, the high frequency of EMD is linked to specific events, the earthquake on the 18th April 2014 induces a perturbation on that mode. The red line corresponds to the mean.

**TEMPORAL ANALYSIS AND EVENT DETECTION.** As we expected, the activity varies through time at several scales. During the day (from midday to 8pm), the activity is greater than during the night. When we average the activity on a daily base for each hour or when we superimposed activity for every days (Figure 4.2.2), we observe a period of lower activity during the night and higher activity during the day. The standard deviation is quite small and variations are periodical. Although the density of mobile phones varies during the day, it also varies between different days. For instance, the activity during weekdays is greater than during the weekend, and the typical peak is reached on Friday at 6 pm.

Moreover, the activity sometimes increases because of specific events. In order to detect and validate such observations and have a better insight on the time series representing the SMS activity per hour, we perform an Empirical Mode Decomposition (EMD) on the signal [130]. The signal is the number of SMS per hour during two months. The EMD allows to represent the non-stationary signal as sum of zero-means Intrinsic Mode Function (IMF) and one residue. Figure 4.2.3 gives the decomposition of the global SMS activity. The

IMF 3 clearly gives a daily periodic signal (a spectral analysis also give an harmonic decomposition in days of the signal). Let observe the high frequency IMF 1 plotted on Figure 4.2.3. We plot in red the mean of the IMF 1. The signal is clearly oscillating around the mean in a compact envelope with few extra peaks of activity. In green, we plot the mean plus twice the standard deviation and extract all points greater than this value in order to point out specific events. For example, in our trace, this signal shows some perturbations, the biggest one happened on the 18th of April 2014 (day 47 on the figure) just after an earthquake near the Pacific coast of Mexico occurred in the state of Guerrero, 265 km southwest of Mexico City.

We also studied the distance in km and in hops that quantify the proximity between source and destination. As one can observe in Figure 4.2.2, these distances are not varying a lot through hours but during the night SMS are more local than during the day. One hypothesis might be that people may send SMS at home. To estimate how local the SMS are, we can compare the average distances of our trace to a random case. For instance, the average in km between two random base stations is equal to 16.4 km whereas, in our trace, the average in km of SMS sent is equal to 6.9 km. Similarly, the average of distances in hops in the graph induces by Voronoï diagram is equal to 14.2 for random base stations and 4.7 for localized SMS. The SMS seems to be much more local than the random case. So, it seems that the locality property can be used to deliver SMS without any routing as many destinations are close to the source.

**ANALYSIS PER BASE STATION.** From an operator point of view, it is interesting to study the traffic in each cell as managed by the base stations. In Figure 4.2.4, we can notice that the distribution of density per base station is heterogeneous as the variation coefficient $CV = \frac{\sigma}{\mu} = 0.63$. Some base stations have a high activity according to the surface whereas others have quite low density. The average of the number of SMS sent is 1378 SMS during that period. It also shows that the density is higher in the city center rather than in suburbs. Although the distance in hops is more homogeneous than the density ($CV = 0.44$), one can notice a larger distance in the city center area. It is certainly due to the spatial density of base stations, which is higher in the city center to manage the higher traffic load. The number of hops to send a SMS at a given distance is then greater. Even if the average of the distance in km per SMS sent for each cell is quite constant ($CV = 0.32$), this averaged distance seems to be a bit longer for the SMS sent from the city center.

## 4.3 PROTOCOLS

We describe two protocols to deliver the SMS in another way than the protocol used in the cellular network. Our approach relies on the density and mobility of phone users combined with the locality of SMS. The protocols use the base stations that are close to the base station attached to the source and users that are connected to it. They are not totally decentralized but more centralized than the classic cellular protocol that routes SMS in the infrastructure network. Figure 4.3.1 depicts these protocols.

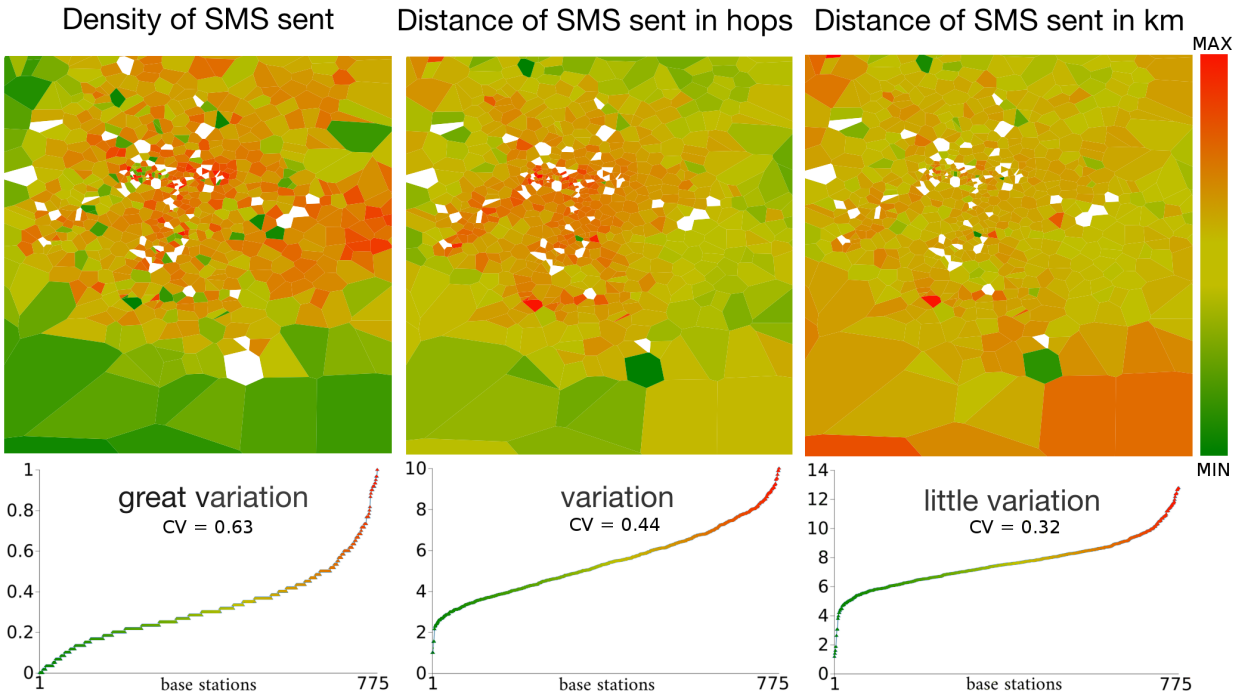



Figure 4.2.4: For each base station, we show the density of SMS sent which corresponds to the normalized ratio between the number of SMS sent and the size of the cell, the distance in hops of SMS sent and the distance in km of SMS sent. The result is plotted in two different ways: (Top) a heat map to give a visualization of the quantities and (Bottom) quantities according to base stations are computed from the lowest to the highest value, the min/max are given in these figures below. (left) The density of a cell is computed as the ratio of the number of SMS sent for a base station and the area size; (middle) the average distance in hops of SMS sent; (right) the average distance in km of the SMS sent.

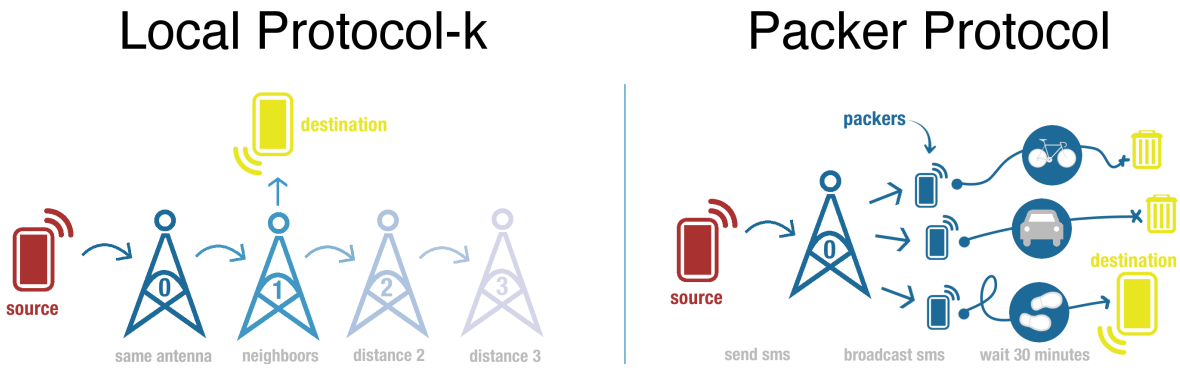



Figure 4.3.1: **Basic protocols:** sketches of the local protocol and packer protocol as defined in this Section 4.3.

**LOCAL PROTOCOL K ($LP_k$).** The source sends the SMS in the usual way to the base station to which it is attached (Figure 4.3.1) (LP-0). Then, this base station transmits to the base stations around itself which

are neighbors at a distance 1 ($LP_1$). This process can be repeated in order to reach the neighbors of the neighbors at distance 2 from the original station ($LP_2$). Considering a fixed distance $k$, the SMS is well received if and only if the destination is attached to a base station at a distance $\leq k$ ($LP_k$). The efficiency of $LP_k$ depends on the locality of the exchanges. During this process, the location of the destination does not have to be known.

**PACKER PROTOCOL ($PP$).** The packer protocol relies on the mobiles attached to the original base station of the sender (Figure 4.3.1). When this base station receives a SMS, it duplicates this message to the mobiles which are attached to that station. In practice, every mobile in that cell gets the SMS, they are called the packers. If the destination is one of these packers, the SMS is already transmitted. The packers who have just received the SMS move through Mexico City and may switch from that antenna to others. As the packers are moving, if a packer and the destination are at the same time attached to the same antenna, the destination will receive the SMS. At this moment, the packer sends the SMS to the base station and the base station sends the SMS to the destination as it is attached to this antenna. The success of this communication depends on the density of packers and thus of mobile users and on their mobility. If many packers are moving randomly all over the city, the probability of reaching the destination should be very high.

## 4.4 RESULTS

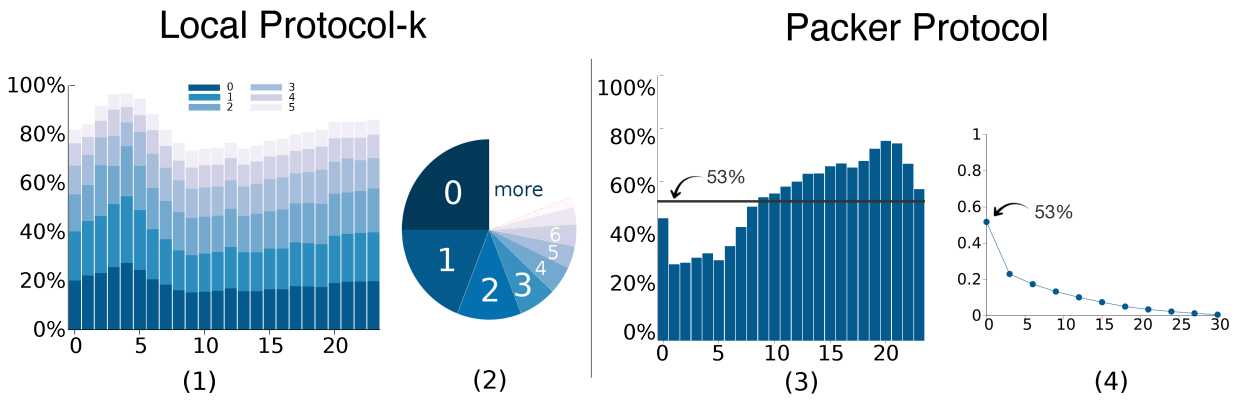



Figure 4.4.1: (1) The average success rate for $LP_k$ as function of the hours of the day. We consider different number of hops $k$ for the transmission of SMS; (2) global transmission success according to the number of hops $k$ averaged over the day. (3) Temporal transmission success for $PP$ averaged by day hours where packers keep the message 30 minutes then remove it from their phones; (4) inverse cumulative distribution of the global transmission success according to the delay ( $\leq$ 30 min ).

We validate our concept by evaluating the feasibility of our approach. It is evaluated through the percentage of SMS that are properly/successfully delivered. We are not competing with classical cellular routing but we show that our protocols can be an alternative for some applications to complement it by relieving the

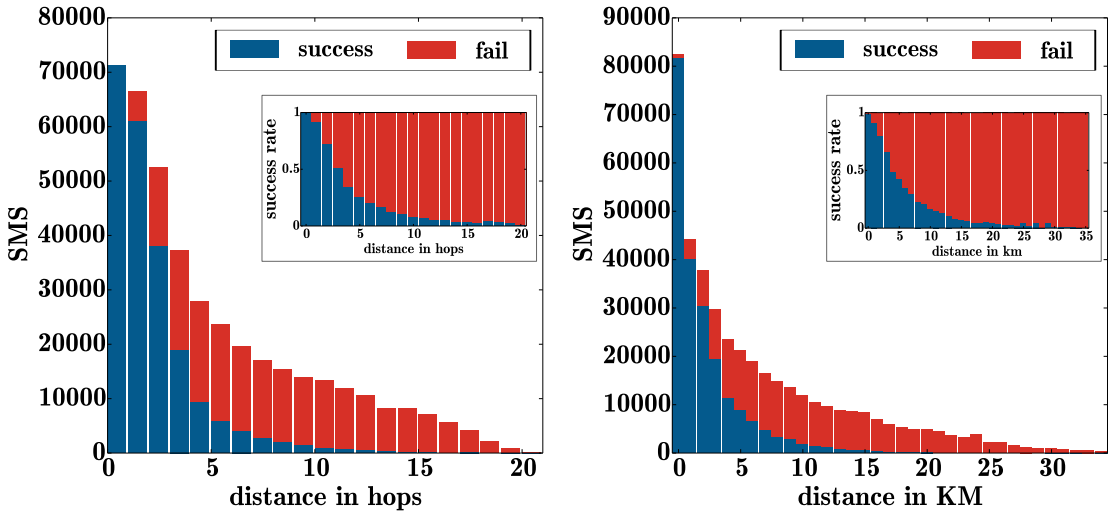
infrastructure network of a part of its load during some challenging time, for instance. For $PP$, we choose to limit the delay to 30 minutes which seems to be a reasonable upper limit for delivering a SMS.

**RESULTS FOR LOCAL PROTOCOL.** We quantified the efficiency of the first protocol in Figure 4.4.1 for which the success is only based on the locality of SMS. We show that 24% of SMS sent reached their destination if we only consider people attached to the same base station ($k = 0$), almost 45% when the base station transmits the SMS to its neighbors ($k = 1$) and 54% if we consider base stations at a two-hop distance ($k = 2$). This result highlights the fact that the destination is usually close to the base station of the source. In comparison, if we consider the base station network of Mexico City, the average of hops between two random stations is 11 and the diameter (maximum distance in hops) of this network is 22. Considering $k \leq 5$, only 13% of the base stations are reachable. In that sense, we can say that there is a high locality of SMS.

**RESULTS FOR PACKER PROTOCOL.** In the traces that we studied for the network of Mexico City (1.5 million SMS over 2 months), one third of SMS are delivered in less than 10 minutes and 53% of them are transmitted in 30 minutes (Figure 4.4.1), this delay is set arbitrarily and reasonable. We consider that after 30 minutes without being delivered, the transmission fails. One can notice that 20% of transmission success is due to a packer that switch from an antenna to another. The mobility is a key point for that protocol. When the activity is high, during rush hours, the protocol has more than 70% of delivering success. It is interesting to notice that the efficiency of our protocol is the best when the operator service is challenged. Moreover, as we miss out some locations (only 9% of the SMS were localized), we just have a part of the packers and have an underestimation of the efficiency. $LP_k$ relies on the locality of SMS sent during the day whereas $PP$ also takes advantage of the mobility and density of phone users leading to a transmission success rate that increases drastically during the rush hours with a maximum rate of 70%.

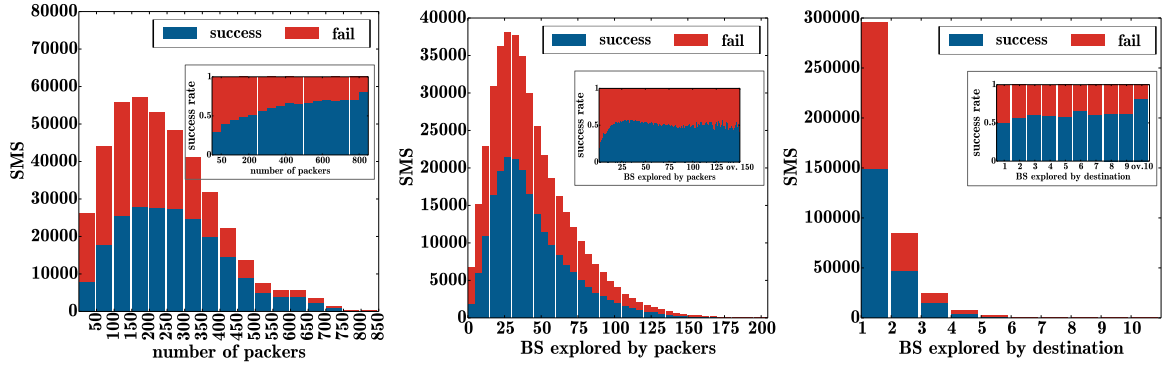
Intuitively, the distance in hops or in km between the source and the destination has a great impact on the success rate. If the source is far from the destination, the packers should be very quick to hopefully reach the destination within 30 minutes. As we can notice on the Figure 4.4.2 (focus), the success rate is decreasing drastically as the original distance between the source and the destination is increasing. When the number of hops is equal to 0, the success rate is 100% whereas it is null for a maximum distance in hops of 21. The distance in km gives the same results. Yet, SMS are very local and the global success rate is boosted by this local property. Few SMS have a long distance, the social interactions between people in Mexico City are local.

In the $PP$, the packers have a key role as they deliver SMS by moving along the city. Our data set contains 6% of these packers. It seems difficult to estimate the underestimation of our success rate due to the lack of geographical information but it seems to be significant. In Figure 4.4.3a, we evaluate the success rate according to the number of packers. When the number of packers is very low, the success rate is approximately 40% and increases with the number of packers. We have to notice that the limit should not




(a) (b)

Figure 4.4.2: Analysis of the success rate of $PP$ according to the original distance between the source and the destination. (left) Number of SMS that reaches the destination, number of SMS that are dropped and total number of SMS according to the distance in hops of the SMS; (left, focus) Success rate according to the distance in hops of SMS; (right) Number of SMS that reaches the destination, number of SMS that are dropped and total number of SMS according to the original distance in km of the SMS; (right, focus) Success rate according to the original distance in km.




(a) (b) (c)

Figure 4.4.3: **For $PP$ protocol.** (a) The average success rate, percentage of SMS delivered, according to the number of packers. (b) The average success rate, percentage of SMS delivered, according to the number of base stations reached by the packers during 30 minutes. (c) Average number of explored base stations according to the hours of the day. It represents the average mobility for each hour of the day.

be 100% because long distance SMS have a very low probability to be delivered even with more packers. Figure 4.4.3a not only shows that our protocol success rate is actually an underestimation but it gives us an indirect measure of the mobility of phone users. Like the density of users and the locality of interactions, the mobility is a capital asset for the mobile users DTN network. More precisely, we quantify the mobility of packers by counting the number of base stations explored during the 30-minute experiment. For each SMS, we count all the base stations explored by at least one packer. In Figure 4.4.3b, we measure the delivering success rate according to the number of base stations explored by the packers. One can intuitively make a link between the mobility of the packers and the amount of base stations explored. When the mobility of packers is very low, the success rate is not suitable, close to 20%. As the mobility of packers is increasing, the rate success is becoming greater than 50%. The difference between having a very small number of explored base stations (less than 20) and having a decent number of explored base stations (greater than 20) has a great impact on the delivery success. We also point out in Figure 4.4.3c that the mobility of the destination has a small effect on the result. When the destination is exploring more than 10 base stations, the success rate is even greater than 80%. Yet, if the destination is not moving, the result are still maintained.

It seems that the number of packers representing the density and their mobility are positively correlated. These two parameters have a direct impact on the success of the protocols. We empirically prove that the correlation between the mobility and the activity is temporally strong (Figure 4.4.4). Users are tending to move and use their mobile phone at the same hours of the day. It shows that we can obtain the full potential of the user DTN network during rush hours when users are active and are moving.

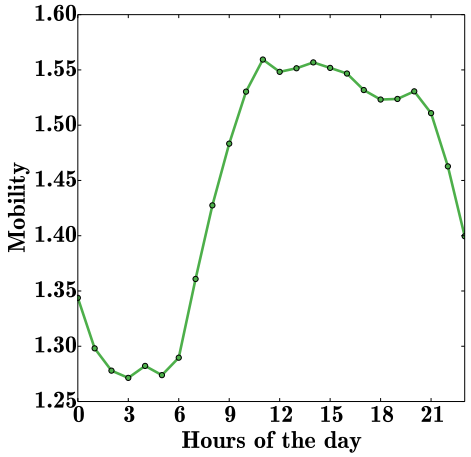



Figure 4.4.4: Average number of explored base stations according to the hours of the day. It represents the average mobility for each hour of the day.

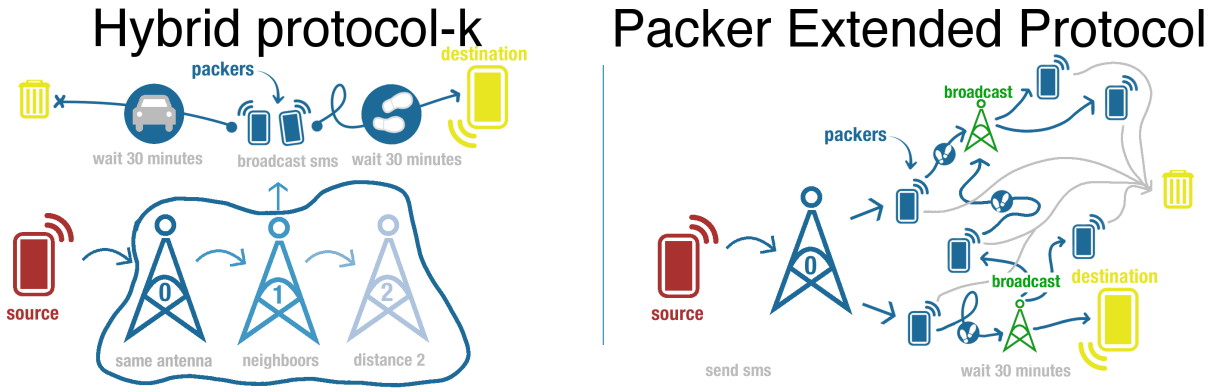



Figure 4.5.1: **Mixed protocols:** sketches of the hybrid protocol-k ($HP_k$) ($k > 0$) and packer extended protocol ($PEP$) as defined in this Section 4.5

## 4.5 Mixed Protocols

In Sections 4.3 and 4.4, we define and analyze two basic protocols. It allowed us to understand the impact of many parameters such as the distance of the SMS, the number of packers around the source and their mobility for these basic mechanics. In this section, two main variations of these two basic protocols are defined and experimented. If some SMS were delivered with $PP$ and not $LP_k$, it means that two protocols are complementary. Here, by defining mixing protocols, we quantify and analyze this possible complementarity.

**Hybrid Protocol-k ($HP_k$).** $HP_k$ (Figure 4.5.1) is a mix between $LP_k$ and $PP$. $LP_k$ locally spreads the message to the base stations at a distance $k$ to the attached base station whereas in $PP$ the source only spreads the message to the attached base station. As we can see in Figure 4.5.1, $HP_k$ is precisely defined by the sequence $(LP_k, PP)$. The $k$ parameter is the distance used in $LP_k$. The process of $PP$ and $HP_k$ are the same except that in $HP_k$, when the base station receives a SMS from a source it send the message not only to the attached packers but also to the base stations that are at a distance k of it. These base stations then send the SMS to the users that are attached to them.

**Packer Extended Protocol ($PEP$).** $PEP$ is a direct extension of $PP$. In $PP$, packers are the users that are attached to the same base station as the source. Here, when one of the packer reaches a new base station, the SMS is broadcasted and all the users that are attached to it become packers. If one of these new packers reaches a new base station, the users' attached to it become also packers. This process continue during the 30 minutes of the experiment. At the end, all the packers drop the SMS. In $PEP$, the number of packers is continuously increasing during the 30-minutes experiment. In average, the packer is keeping the SMS 21 minutes because most of them were not attached to this initial base station and receive the SMS later.

In Figure 4.5.2, we show the success rate of extending protocols compare to the basic one. For protocol 1, the local property of SMS is the only parameter that plays a role on the delivering process. As the locality

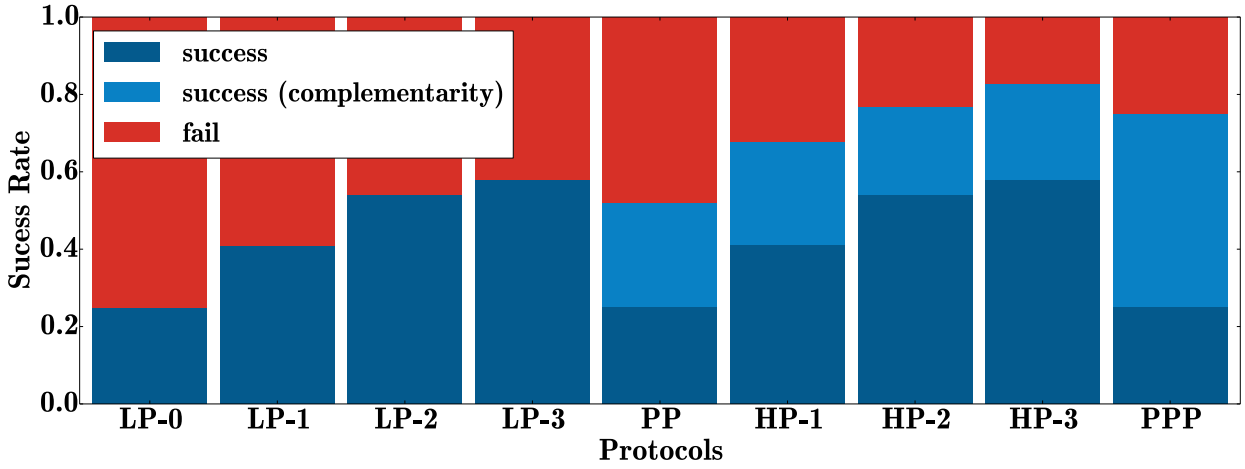



Figure 4.5.2: **Mixed protocols results.** Success rate of the two basic protocols $LP_k$ and $PP$ and of the hybrid or extended versions of $HP_k$ and $PEP$. $k$ varies from 0 to 3 in this figure. The success rate only due to local protocol is depicted in dark blue and the other part due to the mobility of packers in light blue.

diameter is extended, the success rate is increasing. For the protocols $PP$, $HP_k$ and $PEP$, not only the locality but also the mobility of packers is playing an important role. In Figure 4.5.2, we note that the success rate due to the local property and the success rate due to the mobility are not negligeable. Thus, the mobility of packers seems to be supplementary to the local property. We point out that mixing the basic protocols gives a better success rate resulting on a complementarity of local and mobility properties. Furthermore, the average success rate of the broadcast $PEP$ that mainly relies on mobility of packers is equal to 75%. At least 66% of the SMS that reach the destination is due to the mobility of packers. By increasing the locality radius $k$ and the number of packers, the success rate is increasing. These results empirically show it is possible to increase significantly the success rate by using more the mobility of users.

## 4.6 Choices and storage

**Impact of choices.** The spread of the SMS relies on the geo-localization of users. In previous sections, each ego is geo-localized from his phone call activity. Thus, a 30-minute is set to interpolate localization from call to SMS. Yet, one can wonder how this choice is impacting our results. In practice, as one can notice in Figure 4.6.1b, when the interpolation is accurate and so the threshold to interpolate the geo-localization is small, the amount of geo-localized SMS is decreasing. Contrarily, if the threshold is high, the location may be wrong and so it may make no sense to study the SMS. As a good trade off, we decided to set this threshold to 30 minutes. In Figure 4.6.1, we analyze the impact of this threshold. As the threshold is increasing, logically, the number of packers is increasing and so the success rate is improved. This choice is impacting directly the success rate because the number of packers is greater. The choice of a threshold can

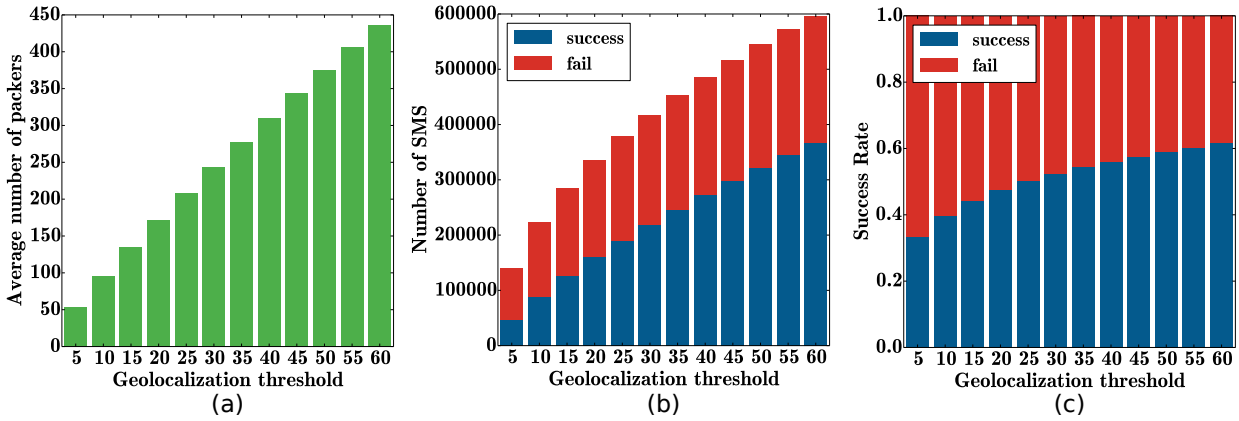



Figure 4.6.1: **Temporal threshold impact in protocol** $PP$ **(a)** Average number of packers according to the geo-localized threshold **(b)** Number of geo-localized SMS depending on the threshold **(c)** Success rate of $PP$ in function of the threshold

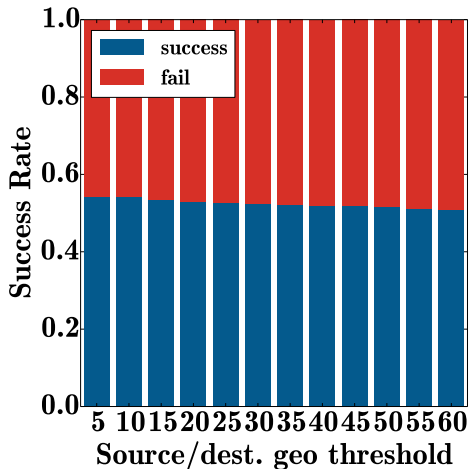



Figure 4.6.2: **Temporal threshold only applied to the source and destination.** Success rate of the $PP$ in function of the geo-localized threshold only applied on the source and the destination. For the packers, we still consider a constant 30-minute threshold.

introduce a bias only if the locations of the source, the destination and the packers are wrong. Nevertheless, when the threshold is smaller, as we miss many packers, we underestimate the success rate. The greatest gap between the experimentation and the reality seems to be due to the missing geo-localizations of users during the diffusion. In order to understand how this threshold is changing the result, we show in the Figure 4.6.2 the success rate of $PP$ when the geo-localized threshold only applied to the source and the destination. Consequently, during the experience, a constant 30-minute threshold is applied to localize the packer. With this experiment, we show that this threshold impacts only the number of packers and that it is this number of packers that influences the success rate.

With this experiment, we show that this threshold only have an impact because choosing a smaller threshold implies a smaller amount of packers.

| Prot. | \|BS\| | Duration | Cost |
|---|---|---|---|
| P1 | 0 | 0mn | 0o |
| P2 | 1 | 30mn | 10kB |
| P3-1 | 5 | 30mn | 63kB |
| P3-2 | 13 | 30mn | 167kB |
| P3-3 | 25 | 30mn | 322kB |
| P4 | 55 | 21mn | 463kB |

Table 4.6.1: Storage cost in kB

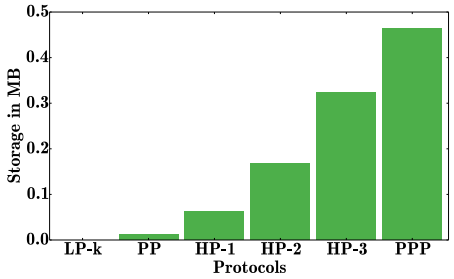



Figure 4.6.3: 30mn storage cost in MB of protocols

**STORAGE COST.** We estimate the capacity of the mobile users DTN network to deliver SMS. We highlighted a link between the number of packers and the success rates. Yet, it is interesting to estimate the cost of these protocols. We estimate the cost of the storage for each user to assess the scalability of our approaches. When a SMS is distributed to packers or potential destinations, the SMS is sent to them. Every SMS last 30 minutes to let the packers moving around the city in order to reach the destination. Each SMS sent to a packer in the same 30-minute slot is stored in its mobile phone. Therefore, the storage cost has to be reasonable. In our experiment, we only have 1.5M geo-localized SMS out of 20M in Mexico City. As the telecommunication is corresponding to 10% of the market share, we estimate to 200M the number of SMS in Mexico City during this 2-month period. We assume that SMS are homogeneously sent through these 775 base stations. We also consider that a SMS costs $140B$. For $PP$, packers are initially linked to the same base station (for a given SMS) and so the number of stored SMS during 30 minutes. The number of SMS stored by a packer is then computed as the mean number of SMS sent from one base station for 30 minutes. The important idea of this work is to make the link between the success rate and some properties like density, mobility, activity of users and communication locality. If we consider the protocol $HP_1$ (resp. $HP_2$ and $HP_3$), the number of SMS stored is equal to 5 (resp. 13 and 25) times the number of SMS sent from one base station. These numbers correspond to the number of initial base stations that are participating to the spread of the SMS over packers that are attached to these base stations. 5 is the mean number of neighbors of a base station and 13 is the number of 2-neighbors, etc. In Figure 4.6.3, for all protocols, the average storage cost does not exceed 1Mo. This storage will be likely be greater during rush hours or in the most active areas like the city center. Furthermore, if we consider $PEP$, it is more demanding on storage as the number of packers is increasing during the whole 30 minutes. For this protocol, the number of base stations that spread SMS to packers is equal to 55 in average. Yet, as some of them are broadcasting at the end, the average time that packers are keeping the SMS with $PEP$ is 21 minutes. In this case, it is also more difficult to predict the cost because it depends on the mobility of packers. As in reality, there are much more packers, this cost

should greatly increase. Yet, our smartphones have a huge capacity of storage, the estimation of the storage is quite reasonable.

## 4.7 Discussion

In this study :

- we have first proposed two basic DTN-like protocols to evaluate the mobile users network. They have been evaluated through an original method as we had the opportunity to replay a large trace of geo-localized SMS in Mexico City. This evaluation has shown that the density of the network, users mobility and locality of SMS play a major role concerning the efficiency of the protocols and so could unload efficiently the backbone network of the operator. Even if there are already some solutions that are close to our protocols, to our knowledge the evaluation of the efficiency of such approaches through the use of huge datasets has never been performed. $LP_k$ offers a transmission success up to 50% whereas it reaches 70% for $PP$ during rush hours, which shows the ability of our approaches to deliver a non negligible part of the SMS and to relieve the infrastructure network of a part of its load. Moreover, we did not used any synthetic model to compute SMS and localizations, our experiments are based on a real traces. We quantified how much the activity and the mobility can improve the success rate of the $PP$. In a large city such as Mexico City, the density and the mobility of people is very high. Our experiment reveals that the mobile users network has an interesting potential,
- we define $HP_k$ and $PP$ that are two variations and combinations of these local and packer protocols. They have been evaluated with this trace. We point out the complementarity of the local property and the mobility of users,
- we finally explain our choice of localizing SMS with a 30-minute window and study its impact. This parameter impacts the success rate because it modifies the number of geo-localized packers. Yet, as we have many missing locations, we already underestimate the number of packers and so the success rate of protocols. The analysis of the storage cost of our solutions show that in a real scenario context, the storage cost of our approach for each user is reasonable. We noticed that the broadcast protocol is the most demanding protocol in terms of storage. This work may be extended in different ways. Concrete proposals on the message exchanges between mobile phones and base stations could be proposed and adapted to the technological context. Also, an analysis in terms of energy consumption and transmission could be performed to estimate globally the costs of our proposals.

Applications to this study are manifold. The comprehension of DTN mobile users network in dense area is the base in order to build smart cities. Understanding the movements of users and above all the spatial and temporal evolution of the social structure is one of the key to improve the transportation services, the communication possibilities and more generally the organization of the city. Yet, locations in our study is obtained at the antenna level. If one can have access at a city level to the communications with more precise locations, it should be possible to refine greatly our experiments.

# Part I: Conclusion & Future Work

Large scale data is offering a great opportunity for studying the dynamic and the nature of social interactions. The temporal and spatial dimensions encompass an important part of the information about the human dynamics and interactions. We provide evidences that the temporal information can be altered dramatically when the time scale of the study is longer than the time scale of the real system. We discussed the link between the event locations of mobile users and their continuous mobility suggesting that mobile activity and mobility are correlated. We provide empirical proofs of the potentiality of the DTN network induced by mobile users in dense areas. The capacity of human DTN network relies on mobility of the users, the density of mobile phones and the locality of interactions.

First, this part raised the issue of making an analysis by taking care of the loss of information while processing large scale CDRs. Each choice made during the analysis should be justified and each limitation should be explained. Second, there are plenty ways in order to undertake a study and answer to a question. We defined new temporal network paths in the Chapter 2 to catch the loss of information, we used the Palm Calculus theory in Chapter 3 to compare two point processes and we experimentally measure protocol efficiency in order to understand properties of the DTN network. The diversity of data-driven approaches is one of the keys to conduct interesting analysis and obtain strong findings. Finally, the richness of the data sets comes from the number of dimensions provided and the distance between each of these dimensions with the real system. The structural aspect, given in this part by mobile interactions, is fundamental in order to verify hypothesis about human structure. Getting not only the interactions but also the time and the locations of mobile events is a great asset to obtain a faithful representation of the real system.

Three extensions of this work can improve the quality of the studies on large scale data sets:

- The spreading of processes may be sensitive to the temporal structures and so, by looking at the way diffusion is evolving over time through the temporal network, it may be possible to detect local and global structures such as clusters.
- Locations at a city level or even a country level over a long period may contain information about the social relations. Having a close relation with someone usually means to share same locations at the same time. It should be interesting to quantify the link between sharing locations and tie existence.
- From location information and relations between cells according to time and intensity, we should be able to model human mobility in dense areas and provide an accurate model for smart cities.

# Part II

# Socioeconomic correlations in mobile networks

# Part II : Introduction

Socioeconomic imbalances, which universally characterize all modern societies [124, 140], are partially induced by the uneven distribution of economic power between individuals. Such disparities are among the key forces behind the emergence of social inequalities [75, 140], which in turn leads to stratification in social structures characterized by correlations between the social network and socioeconomic status of individuals. Although this hypothesis has been drawn long time ago [69], the empirical observation of socioeconomic and structural correlations in large social systems has been difficult as it requires simultaneous access to the social network and economic status for a large number of individuals. Our aim in this set of studies is to understand the social inequalities, find evidence of social stratification and more generally understand the influence of the network position on ego behavior through the analysis of a combined large-scale anonymised dataset that discloses both the social interactions and the economic status of individuals.

The identification of socioeconomic classes is an historical question of the social sciences with several competing hypotheses proposed on their structure and dynamics [65]. One broadly-accepted definition identifies lower, middle, and upper classes [6, 21, 66, 143, 145] based on the socioeconomic status of individuals. These classes can be further used to indicate correlations characterizing the social system. People who live in the same neighborhood may belong to the same class, and may have similar levels of education, jobs, income, ethnic background, and may even share common political views. These similarities together with homophily, i.e. the tendency that people build social ties with similar others [87, 99], strongly influence the structure of social interactions and have indisputable consequences on the global social network as well. The coexistence of social classes and homophily may lead to a strongly stratified social structure where people of the same social class tend to be better connected among each other, while connections between different classes are less frequent than one would expect from structural characteristics only [49, 69, 137].

Social Network Analysis (SNA) provides one promising direction to explore such problems [156], due to its enormous benefit from the massive flow of human behavioral data provided by the digital data revolution [95]. The advent of this era was propagated by some new data collection techniques, which allowed the recording of the digital footprints and interaction dynamics of millions of individuals [5, 88, 112]. On the other hand, although social behavioral data brought us detailed knowledge about the structure and dynamics of social interactions, it commonly failed to uncover the relationship between social and economic positions of individuals. Coarse-grained information about people's economic status are typically provided

as statistical census measures without disclosing the underlying social structure, or by social surveys [24] covering a small and less representative population. Nevertheless, such correlations play important roles in determining one's socioeconomic status (SES) [18], social tie formation preferences due to status homophily [87, 99], and in turn potentially stand behind the emergent stratified structure and segregation on the society level [69, 140]. However until now, the coupled investigation of individual social and economic status remained a great challenge due to lack of appropriate data recording such details simultaneously.

Besides, the consumption of goods and services is a crucial element of human welfare. The uneven distribution of consumption power among individuals goes hand in hand with the emergence and reservation of socioeconomic inequalities in general. Individual financial capacities restrict personal consumer behavior, arguably correlate with one's purchasing preferences, and play indisputable roles in determining the socioeconomic position of an ego in the larger society [41, 75, 124, 140]. It is reflected by sets of commonly purchased products, which are further associated to one's social status [157]. Consumption behavior has been addressed from various angles considering e.g. environmental effects, socioeconomic position, or social influence coming from connected peers [39]. However, large data-driven studies combining information about individual purchasing and interaction patterns in a large population is still rare. Investigation of relations between these characters carries a great potential in understanding better rational social-economic behavior [4], and project to direct applications in personal marketing, recommendation, and advertising.

In addition, cultural reproduction tends to transmit existing cultural values and norms from generation to generation and perpetuate inequalities. The social status of the parents plays a great role on the education and the salary of the children. This phenomenon is also accentuated by the way people are making homophilic connections to each others. When a student is entering the university, new connections are made. It is interesting to quantify how much these new connections, out of the family environment, are random or influenced by socioeconomic parameters. The analysis of the period starting from entering the university and ending when the student gets his or her first job is crucial to comprehend rather if the university period is a key moment in determining the success or if the reproduction is the only base of the social reproduction.

In Chapter 5, we introduce economical parameters and propose a definition of social classes for next chapters. We make the individual link between income, salary, debt and spendings. We show that wealth and debt are unevenly distributed among people in agreement with the Pareto principle. Having access to age and gender of individuals, we propose to point out inequalities according to demographic parameters.

In Chapter 6, we make an analysis of the coupled DS4 recording the mobile phone communications and bank transaction history of one million anonymized individuals living in Mexico. We empirically demonstrate some long-standing hypotheses on socioeconomic correlations which potentially lay behind social segregation, and induce differences in human mobility, or even in communication patterns.

In Chapter 7, by analyzing socioeconomic status, demographic features, and purchasing habits of individuals, we investigate rather if typical consumption patterns are correlated with identified socioeconomic classes leading to a more precise characterization of the stratification in the social structure. In addition, we explore correlations between merchant categories to obtain a big picture of consumption patterns.

In Chapter 8, we provide insights about the effects of marking events on the structure and the dynamics of egocentric networks. More precisely, we study the impact of university admission on the composition and evolution of the egocentric networks of freshmen. The combined dataset gives us the opportunity to estimate the socioeconomic status of students, and to follow their egocentric network evolution starting 6 months before their university admission, up to 18 months after this marking event. Taking the initial before school period as a reference, we study the creation and decay of social ties after the ego was placed in a new social environment. This chapter is part of a bigger project that tends to ask whether university helps to build connections between egos from different socioeconomic classes, or new social ties emerge via homophilic effects between students of similar economic status.

This second and last part is a data-driven longitudinal sociological study that gives answers to long-standing hypotheses on socioeconomic correlations such as social stratification or socioeconomic reproduction which potentially lay behind social inequalities, gender inequity, consumption differences and spatial segregation regarding social classes. To our knowledge, this is the first large-scale data-driven sociological study that explores the social structure with individual socioeconomic parameters of a large population.

# 5

# Social class and inequalities

The full description of one's socioeconomic status is rather difficult as it is characterized not only by quantitative features but also related to one's social or cultural capital [18], reputation, or professional skills. However, we can estimate socioeconomic status by assuming a correlation between one's social position and economic status, which can be approximated by following the network position and financial development of people. This approach in turn not only gives us a measure of an individual's socioeconomic status but can also help us to draw conclusions about the overall distribution of socioeconomic potential in the larger society.

## 5.1 Economic status indicators

Our estimation of an individual's economic status is based on the measurement of consumption power of users extracted from DS4 (presented in Section 1.4). By following the purchase history of each individual, we estimate their economic position from their average amount of debit card purchases. More precisely, for an individual $u$ who spent a total amount of $p_u(t)$ in month $t$, we estimate his/her average monthly purchase (AMP) as

$$P_u = \frac{\sum_{t \in T} p_u(t)}{|T|_u}, \tag{5.1}$$

where $|T|_u$ corresponds to the number of active months of the user (with at least one purchase). In order to verify this individual economic indicator we check its correlations with other indicators, such as the salary $S_u$ (defined as the average monthly salary of individual $u$ over the observation period $T$) and the income

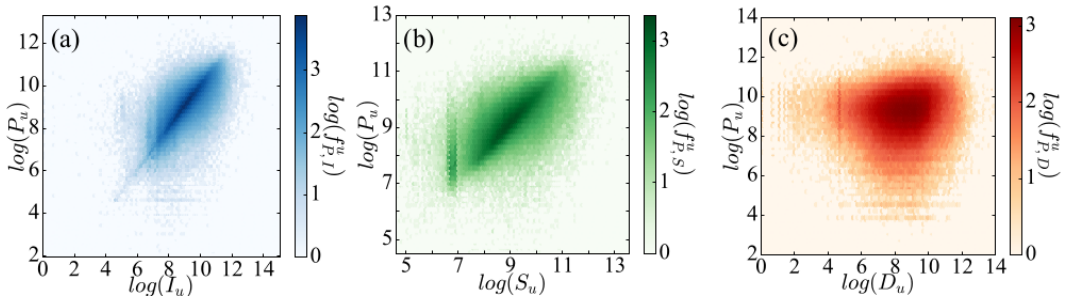



Figure 5.1.1: **Correlations and distributions of individual economic indicators.** The heat maps show correlations between the average monthly purchase $P_u$ and **(a)** average income $I_u$, **(b)** average salary $S_u$, and **(c)** average monthly debt $D_u$ for (a) 625,412 (b) 389,567 and (c) 339,288 customers who have (accordingly) both corresponding measures available. Colors in panels (a-c) depicts the logarithm of the fraction of customers with the given measures.

$I_u$ (defined as the average total monthly income including salary and other incoming bank transfers). We find strong correlations between individual AMP $P_u$ and income $I_u$ with a Pearson correlation coefficient $r \simeq 0.758$ ($p < .001$) (for correlation heat map see Fig.5.1.1a), and also between $P_u$ and salary $S_u$ with $r \simeq 0.691$ ($p < .001$) (see Fig.5.1.1b). Note that direct economic indicators, such as $I_u$ and $S_u$, are available only for a smaller subset of users (for exact numbers see Fig.5.1.1 caption), thus for the present study we decided to use $P_u$ since this measure is available for the whole set of users.

At the same time we are interested in an equivalent indicator which estimates the financial commitments of individuals. We define the average monthly debt (AMD) of an individual $u$ by measuring

$$D_u = \frac{\sum_{t \in T} d_u(t)}{|T|_u}, \tag{5.2}$$

where $d_u(t)$ indicates the debt of individual $u$ in month $t \in T$ and $|T|_u$ is the number of active months where the user had debt. Arguably individual debt could depend on the average income and thus on the AMP of a person due to the loaning policy of the bank. Interestingly, as demonstrated in Fig.5.1.1c, we found weak correlations between AMP and AMD with a small coefficient $r \simeq 0.104$ ($p < .001$), which suggests that it is worth to treat these two indicators independently.

## 5.2 Socioeconomic imbalances

The distribution of an individual economic indicator may disclose signs of socioeconomic imbalances on the population level. This hypothesis was first suggested by V. Pareto and later became widely known as the law named after him [119]. The present data provide a straightforward way to verify this hypothesis through the distribution of individual AMP. We measured the normalized cumulative function of AMP for $f$ fraction

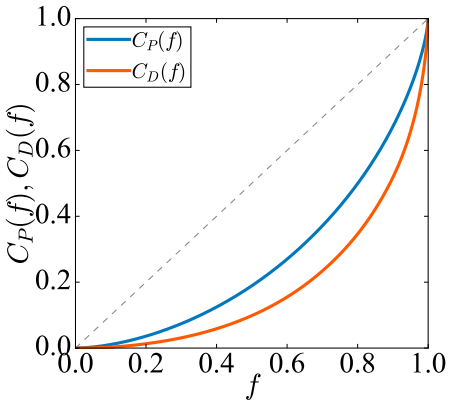



Figure 5.2.1: **Pareto principle.** Cumulative distributions of $P_u$ (blue line) and $D_u$ (orange line) as functions of sorted fraction $f$ of individuals. Dashed line shows the case of the perfectly balanced distribution.

of people sorted by $P_u$ in an increasing order:

$$C_P(f) = \frac{1}{\sum_u P_u} \sum_f P_u \tag{5.3}$$

This function shows (see Fig.5.2.1 blue line) that AMP is distributed with a large variance, i.e., indicating large economical imbalances just as suggested by the Pareto's law. A conventional way to quantify the variation of this distribution is provided by the Gini coefficient $G$ [63], which characterizes the deviation of the $C_P(f)$ function from a perfectly balanced situation, where wealth is evenly distributed among all individuals (diagonal dashed line in Fig.5.2.1). In our case we found $G_P \simeq 0.461$, which is relatively close to the World Bank reported value $G = 0.481$ for Mexico [10], and corresponds to a Pareto index [105] $\alpha = 1.315$. This observation indicates a $0.73 : 0.27$ ratio characterizing the uneven distribution of wealth, i.e., that the $27\%$ of people are responsible for the $73\%$ of total monthly purchases in the observed population. Note that these values are close to the values $G = 0.6$ and $80 : 20$, which were suggested by Pareto.

At the same time we have characterized the distribution of individual AMD by measuring the corresponding $C_D(f)$ function as shown in Fig.5.2.1 (orange line). It indicates even larger imbalances in case of debt with a Gini coefficient $G_D \simeq 0.627$ and $\alpha = 1.140$ indicating $19\%$ of the population to be actually responsible for $81\%$ of overall debt in the country. This observation suggests that Pareto's hypothesis holds not only for the distribution of purchases but for debt as well. Note that similar distribution of debt of bankrupt companies has been reported [7].

## 5.3 Social class definition

We assign each individual into one of $n = 9$ socioeconomic classes based on their individual AMP values. This classification is defined by sorting individuals by their AMP, taking the cumulative function $C_P(f)$

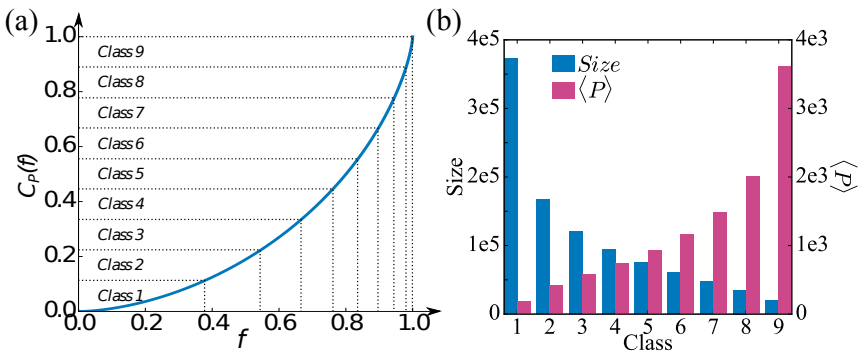



Figure 5.3.1: **Social class definition (a)** Schematic demonstration of user partitions into 9 socioeconomic classes by using the cumulative average monthly purchase (AMP) function $C_P(f)$. Fraction of individuals belonging to a given class ($x$ axis) have the same sum of AMP $(\sum_u P_u)/n$ ($y$ axis) for each class. **(b)** Number of egos (blue), and the average AMP $\langle P \rangle$ (in USD [2]) per individual (pink) in different classes.

of AMP, and cutting it in $n$ segments such that the sum of AMP in each class is equal to $(\sum_u P_u)/n$ (as shown in Fig.5.3.1a). Our selection of nine distinct classes is based on the common three-stratum model [6, 21], which identifies three main social classes (lower, middle, and upper), and three sub-classes for each of them [137]. More importantly, this way of classification relies merely on individual economic estimators, $P_u$, and naturally partitions individuals into classes with decreasing sizes, and increasing $\langle P \rangle$ per capita average AMP values for richer groups (for exact values see Fig.5.3.1b)[2].

## 5.4 Social class characteristics

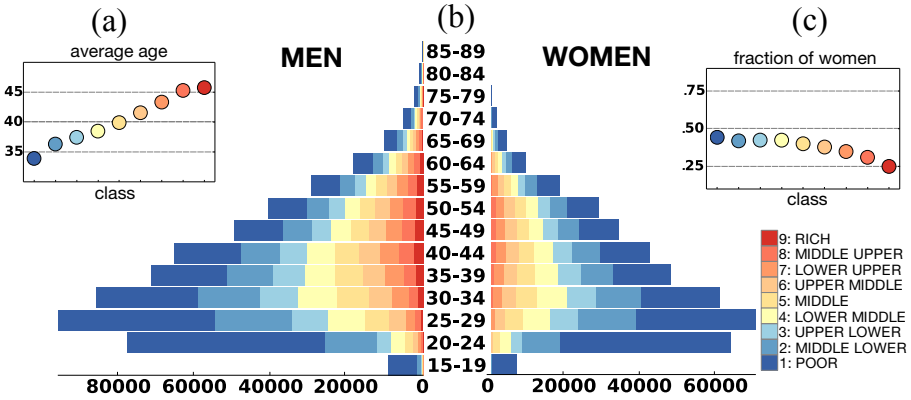



Figure 5.4.1: **Social class characteristics (a)** Average age of different classes. **(b)** Age pyramids for men and women with colors indicating the corresponding socioeconomic groups and with bars proportional to absolute numbers. **(c)** Fraction of women in different classes.

To explore the demographic structure of the classes we used data on the age and gender of customers.

We draw the population pyramids for men and women in Fig.5.4.1b with color-bars indicating the number of people in a given social class at a given age. In our data sets, we have to notice that there are in average less women and older people. The access to bank account is not uniformly distributed in terms of age and gender. The age bias is natural as children do not have income or salary. We found a positive correlation between social class and average age, suggesting that people in higher classes are also older on average (see Fig.5.4.1a). Yet, the gender distribution deviation underestimates the number of women in the country of Mexico meaning that women have less access to bank accounts. In addition, our data verifies the presence of gender imbalance as the fraction of women varies from $0.45$ to $0.25$ by going from lower to upper socioeconomic classes (see Fig.5.4.1c).

## 5.5 Discussion

In this chapter, we defined economic parameters such as the wealth, salary, income or debt. As exploratory results:

- we pointed out that the wealth, the salary and the income are positively correlated and that, surprisingly, the debt and wealth parameters are not correlated,
- we showed that individual economic indicators such as average monthly purchases and also debts are unevenly distributed in the population in agreement with the Pareto principle,
- we proposed a systematic definition for social classes recovering realistic characteristics,
- we observed signatures of gender and age inequalities and confirmed that the fraction of men is increasing for wealthier social classes,
- we confirmed this hypothesis by showing that highest social class is mainly composed by men with a certain age.

There are some ways to refine our definition of social classes. The full description of one's socioeconomic status is rather difficult as it is characterized not only by quantitative features but also related to one's social or cultural capital [18], reputation, or professional skills. Complementary information providing the marital status, the work and the education level of each ego seem to be essential in order to take into account as main criteria.

# 6

# Socioeconomic correlations and stratification in social-communication networks

The uneven distribution of wealth and individual economic capacities are among the main forces which shape modern societies and arguably bias the emerging social structures. However, the study of correlations between the social network and economic status of individuals is difficult due to the lack of large-scale multimodal data disclosing both the social ties and economic indicators of the same set of individuals. In this chapter, we aim to close this gap through the analysis of coupled datasets recording the mobile phone communications and bank transaction history of one million anonymized individuals living in Mexico. We observed social structure is strongly stratified, with people being better connected to others of their own socioeconomic class rather than to others of different classes; the social network appears with assortative socioeconomic correlations and tightly connected "rich-clubs"; and that egos from the same class live closer to each other but commute further if they are wealthier. These results are based on a representative, society-level population, and empirically demonstrate some long-lasting hypotheses on socioeconomic correlations which potentially lay behind social segregation, and induce differences in human mobility, or even in communication patterns.

## 6.1 Introduction

The economic capacity of individuals arguably correlates with their professional occupation, education level, and housing, which in turn determine their social status and environment. At the same time status

homophily [87, 99], i.e., people's tendency to associate with others of similar social status, has been argued to be an important mechanism that drives the creation of social ties. Our hypothesis is that these two effects, diverse socioeconomic status and status homophily, potentially lead to the emergence of a stratified structure in the social network where people of the same social class tend to be better connected among themselves than with people from other classes. A similar hypothesis has been suggested earlier [17] but its empirical verification has been impossible until now as this would require detailed knowledge about the social structure and precise estimators of individual economic status. In the following, our main contribution is to clearly identify signatures of social stratification in a representative society-level dataset, that contains information on both the social network structure and the economic status of people.

## 6.2 Social structural stratification

Using the above-defined socioeconomic classes (defined in Section 5.1) and the social network structure from DS4 (more details in Section 1.4), we turn to look for correlations in the inter-connected class structure. To highlight structural correlations, such as the probability of connectedness, we use a randomized reference system. It is defined as the corresponding configuration network model structure where we take the original social network, select random pairs of links and swap them without allowing multiple links and self loops. In order to remove any residual correlations we repeated this procedure $5 \times |E|$ times. This randomization keeps the number of links, individual economic indicators $P_u$, and the assigned class of people unchanged, but destroys any structural correlations in the social structure and consequently between socioeconomic layers as well. In each case, we repeat this procedure for 100 times and present results averaged over the independent random realizations. Taking the original (resp. randomized) network we count the number of links $|E(s_i, s_j)|$ (resp. $|E_{rn}(s_i, s_j)|$) connecting people in different classes $s_i$ and $s_j$. After repeating this procedure for each pair of classes in both networks, we take the fraction:

$$L(s_i, s_j) = \frac{|E(s_i, s_j)|}{|E_{rn}(s_i, s_j)|}, \tag{6.1}$$

which gives us how many times more (or less) links are present between classes in the original structure as compared to the randomized one. Note that in the randomized structure the probability that two people from given classes are connected depends only on the number of social ties of the individuals and the size of the corresponding classes, but is independent of the effect of potential structural correlations. This way the comparison of the original and random structures highlights only the effect of structural correlations induced by status homophily or other tie creation mechanisms.

From the chord diagram visualization of this measure in Fig.6.2.1c, we can draw several conclusions. Note that for better visual presentation in Fig.6.2.1a we have normalized $L(s_i, s_j)$ and thus chord width indicates relative values $\widetilde{L}_{s_i}(s_j) = L(s_i, s_j)/\sum_{s_j} L(s_i, s_j)$ as compared to the origin class $s_j$ (as also explained in the figure caption). First, after sorting the chords of a given class $s_i$ in a decreasing $L(s_i, s_j)$

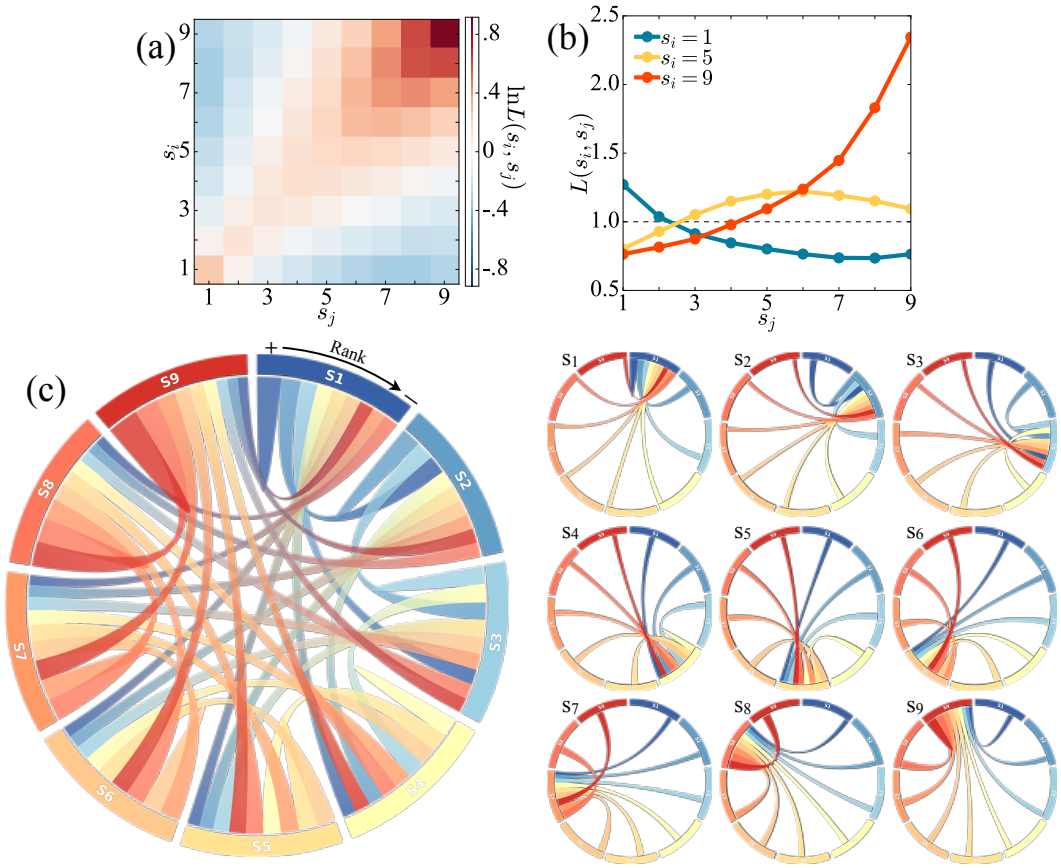



Figure 6.2.1: **Structural correlations in the socioeconomic network (a)** Matrix representation of $L(s_i, s_j)$ (for definition see Eq.6.1) with logarithmic color scale. **(b)** The $L(s_i, s_j)$ function extracted for three selected classes (1 (blue), 5 (yellow), and 9 (red)). Panels (a)-(d) provide quantitative evidence on the stratified structure of the social network and the upward-biased connections of middle classes. **(c)** Chord diagram of connectedness of socioeconomic classes $s_i$, where each segment represents a social class $s_i$ connected by chords with width proportional to the corresponding inter-class link fraction $\tilde{L}_{s_i}(s_j)$, and using gradient colors matched with opposite ends $s_j$. Note that the $\tilde{L}_{s_i}(s_j) = L(s_i, s_j)/\Sigma_{s_j} L(s_i, s_j)$ normalized fraction of $L(s_i, s_j)$ (in Eq.6.1) was introduced here to assign equal segments for each class for better visualization. Chords for each class are sorted in decreasing width order in the direction shown above the main panel. On the minor chord diagrams of panel **(a)**, graphs corresponding to each class are shown with non-gradient link colors matching the opposite end other than the selected class.

order, chords connecting a class to itself (self-links) appear always at top (or top 2nd) positions of the ranks. At the same time other top positions are always occupied by chords connecting to neighboring social classes. These two observations (better visible in Fig.6.2.1c insets) indicate strong effects of status homophily and the existence of stratified social structure where people from a given class are the most connected with similar others from their own or from neighboring classes, while connections with individuals from remote classes are least frequent. A second conclusion can be drawn by looking at the sorting of links in the middle and lower upper classes ($S4 - S8$). As demonstrated in the inset of Fig.6.2.1c, people prefer to connect upward and tend to hold social ties with others from higher social classes rather than with people from lower classes.

These conclusions can be further verified by looking at other representations of the same measure. First we show a heat map matrix representation of Eq.6.1 (see Fig.6.2.1a), where $L(s_i, s_j)$ values are shown with logarithmic color scales. This matrix has a strong diagonal component verifying that people of a given class are always better connected among themselves (red) and with others from neighboring groups, while social ties with people from remote classes are largely underrepresented (blue) as compared to the expected value provided by the random reference model. This again indicates the presence of homophily and the stratified structure of the socioeconomic network. The upward-biased inter-class connectivity can also be concluded here from the increase of the red area around the diagonal by going towards richer classes. These conclusions are even more straightforward from Fig.6.2.1b where the $L(s_i, s_j)$ is shown for three selected classes (1-poor, 5-middle, and 9-rich). These curves clearly indicate the connection preferences of the selected classes. Moreover, they show that richest people appear with the strongest homophilic preferences as their class is $\sim 2.25$ times better connected among each other than expected by chance, on the expense of weaker connectivity to remote classes. This effect is somewhat weaker for middle classes, which function as bridges between poor and rich classes, but apparently upward biased towards richer classes. This set of results directly verifies our earlier conjectures that the structure of the socioeconomic network is strongly stratified and builds up from social ties, whose creation is potentially driven by status homophily, and determined by the socioeconomic characteristics of individuals.

The above observations further suggest that the social structure may show assortative correlations in terms of socioeconomic status on the individual level. In other words, richer people may be better connected among themselves than one would expect them by chance and this way they form tightly connected "rich-clubs" in the structure similar to the suggestion of Mills [102]. This can be actually verified by measuring the "rich-club" coefficient [34, 167], after we adjust its definition to our system as follows. We take the original social network structure, sort individuals by their AMP value $P_u$ and remove them in an increasing order from the network (together with their connected links). At the same time we keep track of the density of the remaining network defined as

$$\phi(P_>) = \frac{2L_{P_>}}{N_{P_>}(N_{P_>} - 1)} \tag{6.2}$$

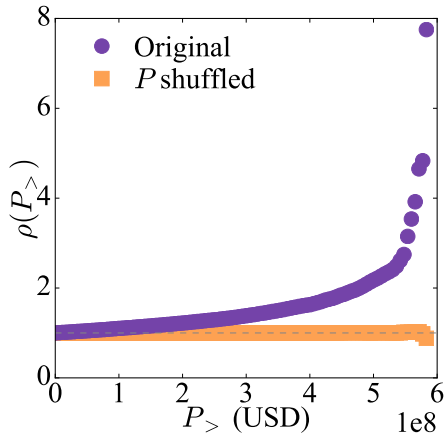



Figure 6.2.2: **Rich clubs.** "Rich-club" coefficient $\rho(P_>)$ (definition see Eq.6.3) based on the original (purple), and shuffled (orange) networks. On the individual level the richest people of the population appear to be eight times more densely connected than expected randomly.

where $L_{P_>}$ and $N_{P_>}$ are the number of links and nodes remaining in the network after removing nodes with $P_u$ smaller than a given value $P_>$. In our case, we consider $P_>$ as a cumulative quantity going from 0 to $\sum_u P_u$ with values determined just like in case of $C_P(f)$ in Fig.6.2.2 but now using 100 segments. At the same time, we randomize the structure using a configuration network model and by removing nodes in the same order we calculate an equivalent measure $\phi_{rn}(P_>)$ as defined in Eq.6.2 but in the uncorrelated structure. For each randomization process, we used the same parameters as earlier and calculated the average density $\langle\phi_{rn}\rangle(P_>)$ of the networks over 100 independent realizations. Using the two density functions we define the "rich-club" coefficient as

$$\rho(P_>) = \frac{\phi(P_>)}{\langle\phi_{rn}\rangle(P_>)}, \tag{6.3}$$

which indicates how many times the remaining network of richer people is denser connected than expected from the reference model. In our case (see Fig.6.2.2 (purple symbols)) the "rich-club" coefficient increases monotonously with $P_>$ and grows rapidly once only the richer people remain in the network. At its maximum it shows that the richest people are $\sim 8$ times more connected in the original structure than in the uncorrelated case. This provides a direct evidence about the existence of tightly connected "rich-clubs" [102], and the presence of strong assortative correlations in the social structure on the level of individuals in terms of their socioeconomic status. To rule out the possibility that our observation was induced by positive correlations between the wealth and observed number of links of people, we performed an alternative measurement where we first shuffled the $P_u$ values among individuals and repeated the same procedure by calculating average coefficient values over 100 independent realizations. In this case the $\rho(P_>)$ function appeared approximately as a constant around 1 (orange symbols in Fig.6.2.2) verifying that our original observation was exclusively induced by correlations indicating that the richest people in the social network actually prefer to build social ties with other rich individuals.

## 6.3 SPATIAL CORRELATIONS BETWEEN SOCIOECONOMIC CLASSES

As we discussed earlier, the economic capacity of an ego strongly determines the possible places he or she can afford to live, arguably leading to somewhat homogeneous neighborhoods, districts, towns, and regions occupied by people from similar socioeconomic classes. This effect may translate to correlations in the spatial distribution of socioeconomic classes in relation with each other. To study such correlations, we use DS5 (for details see chapter 1).

### 6.3.1 SOCIAL CLASS AND COMMUTING DISTANCE

Socioeconomic status of people may also correlate with their typical commuting distances (between home and work), a question which has been studied thoroughly during the last few decades. Some of these studies suggest a positive correlation between economical status (income) and the distance people travel every day between their home and work locations [126, 158, 159]. Such correlations were partially explained by the positive payoff between commuting farther for better jobs, while keeping better housing conditions. On the other hand recent studies suggest that such trends may change nowadays as in central metropolitan areas, where the better job opportunities are concentrated, became more expensive to live and thus occupied by people from richer classes [92, 132]. Without going into details we looked for overall signs of such correlations by using the estimated home ($\ell_h$) and work ($\ell_w$) locations of individuals from different classes. For each ego we measure a commuting distances as $d_{hw} = |\ell_h - \ell_w|$ and compute the $P_{s_i}(d_{hw})$ distributions for everyone in a given $s_i$ class, together with the $P_{all}(d_{hw})$ distribution considering all individuals. For each class we are interested in

$$d_{\Delta}^{s_i}(d_{hw}) = P_{s_i}(d_{hw}) - P_{all}(d_{hw}), \tag{6.4}$$

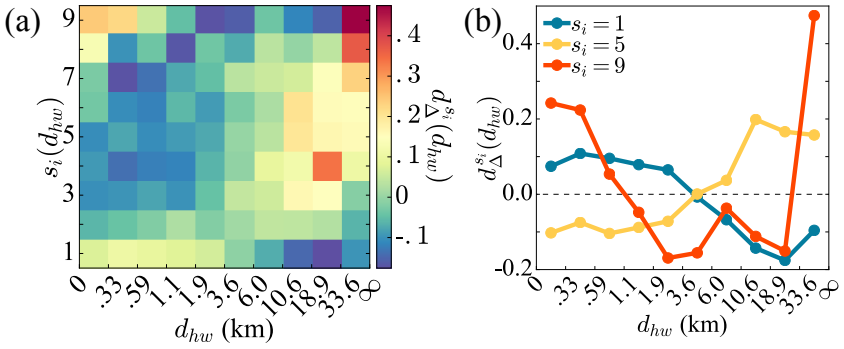



Figure 6.3.1: **Social class and commuting distance. (a)** $d_{\Delta}^{s_i}(d_{hw})$ differences between commuting distance distributions calculated for different classes and for the whole population. $x$ scale depicts in logarithmic values of $d_{hw}$ commuting distances. **(b)** The same $d_{\Delta}^{s_i}(d_{hw})$ functions as on panel (a) shown for a selected set of classes (1-poor (blue), 5-middle (yellow), 9-rich (red)).

i.e., the difference of the corresponding distributions at each distance $d_{hw}$. This measure is positive (resp.

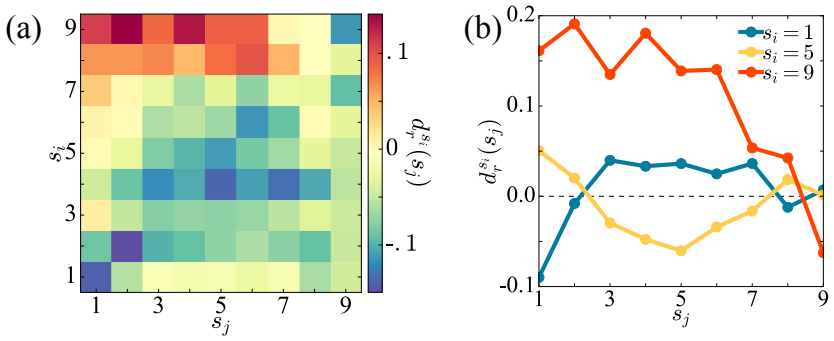



Figure 6.3.2: **Spatial socioeconomic correlations.** **(a)** Relative average geodesic distances for different classes using the measure $d_r^{s_i}(s_j)$ defined in Eq.6.6. **(b)** The same $d_r^{s_i}(s_j)$ functions as on panel (a) shown for a selected set of classes (1-poor (blue), 5-middle (yellow), 9-rich (red)).

negative) if more (resp. less) people commute at a distance $d_{hw}$ as compared to the overall distribution, thus indicating whether people of a given class are over (under)represented at a given distance. Interestingly, our data is in agreement with both above mentioned hypotheses, as seen in Fig.6.3.1a where we show $d_\Delta^{s_i}(d_{hw})$ for each class as a heat map. There, poorer people are over represented in shorter distances while this trend is shifted towards larger distances (see right skewed yellow component in Fig.6.3.1a) as going up in the class hierarchy. This continues until we reach the richest classes (8 and 9) where the distance function becomes bimodal assigning that more people of these classes tend to live very far or very close to their work places as compared to expectations considering the whole population. This is even more visible in Fig.6.3.1a where selected $d_\Delta^{s_i}(d_{hw})$ functions are depicted for selected classes.

### 6.3.2 Spatial correlations in socioeconomic network

To quantify spatio-socioeconomic correlations, we measure the relative average geodesic distance between classes. More precisely, we take all connected egos $(u, v) \in E$ belonging to classes $u \in s_i$ and $v \in s_j$ respectively and measure the geodesic distance $d_{geo}^{zip}(a, b)$ between their zip locations. Using these values we calculate the average geodesic distance between any pairs of socioeconomic classes as

$$\langle d_{geo}(s_i, s_j)\rangle = \frac{1}{L(s_i, s_j)} \sum_{\substack{(u,v)\in E \\ u\in s_i, v\in s_j}} d_{geo}^{zip}(u, v) \tag{6.5}$$

where $L(s_i, s_j)$ assigns the number of links between nodes in classes $s_i$ and $s_j$. Subsequently we calculate the average distance between nodes from class $s_i$ and any of their neighbors $\langle d_{geo}(s_i)\rangle$ to derive

$$d_r^{s_i}(s_j) = \frac{\langle d_{geo}(s_i, s_j)\rangle - \langle d_{geo}(s_i)\rangle}{\langle d_{geo}(s_i)\rangle}. \tag{6.6}$$

This measure gives us the relative average geodesic distance between egos in $s_i$ to egos in other classes $s_j$ as compared to the average distance of egos $s_i$ from any of their connected peers. Results are presented as a heat map matrix in Fig.6.3.2a where the diagonal component suggests a peculiar correlation. It shows that the relative average distance is always minimal (and negative) between egos of the same class $s_i$. This means that people tend to live relatively the closest to similar others from their own socioeconomic class as to egos from different classes, independently in which class they belong to. This is even more visible in Fig.6.3.2b after extracting the $d_r^{s_i}(s_j)$ curves (corresponding to rows in Fig.6.3.2a) for three selected classes. It highlights that while people of the poorest class live relatively the closest to each other, rich people tend to leave relatively the furthest from anyone from lower socioeconomic classes. These correlations are very similar to ones we already observed in the social structure suggesting that the stratified structure and spatial segregation may have similar roots. They are determined by the entangled effects of economic status and status homophily, together with other factors such as ethnicity or other environmental effects, which we cannot consider here.

## 6.4 DISCUSSION

In this chapter, we have investigated socioeconomic correlations through the analysis of a coupled dataset of mobile phone communication records and bank transaction history for $1$ million individuals over $8$ months. After mapping the social structure and estimating individual economic capacities, we addressed three different aspects of their correlations:

- after grouping people into nine socioeconomic classes we detected effects of status homophily and showed that the socioeconomic network is stratified as people most frequently maintain social ties with people from their own or neighboring social classes,
- we observed that the social structure is upward-biased towards wealthier classes and show that assortative correlations give rise to strongly connected "rich-clubs" in the network,
- finally, we demonstrated that people of the same socioeconomic class tend to live closer to each other as compared to people from other classes, and found a positive correlation between their economic capacities and the typical distance they use to commute.

There are interesting ways for extending our findings. The social stratification is quantified for the undirected and unweighted graph. One can imagine that the strength of relations may have an impact on the social stratification. Including direction, total duration of calls and number of SMS and calls may refine our correlations. In general, the impact of certain parameters such as the strength of ties, the spatiality, the temporality (daily or monthly), the personal parameters (age, gender), the type of links (family, friends, co-workers) should be even more interesting that the data-driven approach is the only way of catching it.

# 7

# Correlations of consumption patterns in socio-economic networks

In this chapter, we analyze a coupled dataset collecting the mobile phone communication and bank transactions history of a large number of individuals living in Mexico. From the mapping of the social structure with individual indicators of socioeconomic status, demographic features, and purchasing habits of individuals we show that typical consumption patterns are strongly correlated with identified socioeconomic classes leading to patterns of stratification in the social structure. In addition we measure correlations between merchant categories and introduce a correlation network, which emerges with a meaningful community structure. We detect multivariate relations between merchant categories and show correlations in purchasing habits of individuals. Our work provides novel and detailed insight into the relations between social and consuming behavior with potential applications in recommendation system design.

## 7.1 Introduction

The consumption of goods and services is a crucial element of human welfare. The purchasing power is strongly linked to the financial capacities of individuals and the socioeconomic stratification as well. Earlier hypotheses on the relation between consumption patterns and socioeconomic inequalities, and their correlations with demographic features such as age, gender, or social status were drawn from specific sociological studies [29] and from cross-national social surveys [40]. These studies show that personal social interactions, social influence [39], or homophily [162] in terms of age or gender [78] have strong effects on

purchase behavior, knowledge which led to the emergent domain of online social marketing [58]. Yet it is challenging to measure correlations between individual social status, social network, and purchase patterns simultaneously. Although socioeconomic parameters can be estimated from communication networks [48] or from external aggregate data [52] usually they do not come together with individual purchase records. In this chapter, we propose to explore this question through the analysis of a combined dataset proposing simultaneous observations of social structure, economic status and purchase habits of millions of individuals.

To set this combined data-driven approach , we make the analysis of DS5, which simultaneously records the mobile-phone communication, bank transaction history, and purchase sequences of millions of inhabitants of Mexico over several months. This corpus, one among the firsts at this scale and level of details, allows us to infer the socioeconomic status, consumption habits, and the underlying social structure of millions of connected individuals. Using this information our overall goal is to identify people with certain financial capacities, and to understand *how much money they spend, on what they spend, and whether they spend like their friends?* More precisely, we formulate our study around two research questions:

- Can one associate typical consumption patterns to people and to their peers belonging to the same or different socioeconomic classes, and if yes how much such patterns vary between individuals or different classes?
- Can one draw relations between commonly purchased goods or services in order to understand better individual consumption behavior?

Starting from DS5 described in Chapter 1 and social classes definition introduced in Chapter 5, we propose spending indicators in Section 7.2. In Section 7.3 we show how typical consumption patterns vary among classes and relate them to structural correlations in the social network. In Section 7.4 we draw a correlation network between consumption categories to detect patterns of commonly purchased goods and services. Finally we present some concluding remarks and future research ideas in section 7.5.

## 7.2 Socioeconomic correlations in purchasing patterns

All this study is based on DS3 and DS4 (presented in Chapter 1). DS3 provides spending information about 3,680,652 bank users whereas DS4 gives, in addition, the ego network of 992,538 bank and mobile users. In order to address our first research question, we compare the purchasing patterns of the nine socioeconomic classes. In Fig.7.2.1a, without considering cash (*Service Providers*) that represents around 68% of the total spending, we show the percentage of total amount of money spent on each PCG. In average, bank clients spend 26.5% of the total spending on *Retail Stores*, 17.8% on *High Risk Personal Retail* and 9.5% on *Restaurants*. For each line of the histogram, percentages for the poorest, middle and richest social class are shown colored dots. It roughly points out major purchasing differences between social classes. We note the poorest spend 31.8% on *Retail Stores* whereas the richest spend only 19.6% on the same category.

To understand better the link between social classes and PCGs, we formally quantify the spending rate of each class on each PCG. For each class $s_j$ we take every users $u \in s_j$ and calculate the $m_u^k$ total amount

of purchases they spent on a purchase category group $k \in K_{17}$ (see more details in Section 1.3). Then we measure a fractional distribution of spending for each PCGs as:

$$r(k, s_j) = \frac{\sum_{u \in s_j} m_u^k}{\sum_{u \in s} m_u^k}, \tag{7.1}$$

where $s = \bigcup_j s_j$ assigns the complete set of users. In Fig.7.2.1b each line shows the $r(k, s_j)$ distributions for a PCG as the function of $s_j$ social classes, and lines are sorted (from top to bottom) by the total amount of money spent on the actual PCG[1]. Interestingly, people from lower socioeconomic classes spend more on PCGs associated to essential needs, such as *Retail Stores (St.)*, *Gas Stations*, *Service Providers* (cash) and *Telecom*, while in the contrary, other categories associated to extra needs such as *High Risk Personal Retail* (Jewelry, Beauty), *Mail Phone Order*, *Automobiles*, *Professional Services (Serv.)* (extra health services), *Whole Trade* (auxiliary goods), *Clothing St.*, *Hotels* and *Airlines* are dominated by people from higher socioeconomic classes. Also note that concerning *Education* most of the money is spent by the lower middle classes, while *Miscellaneous St.* (gift, merchandise, pet St.) and more apparently *Entertainment* are categories where the lowest and highest classes are spending the most.

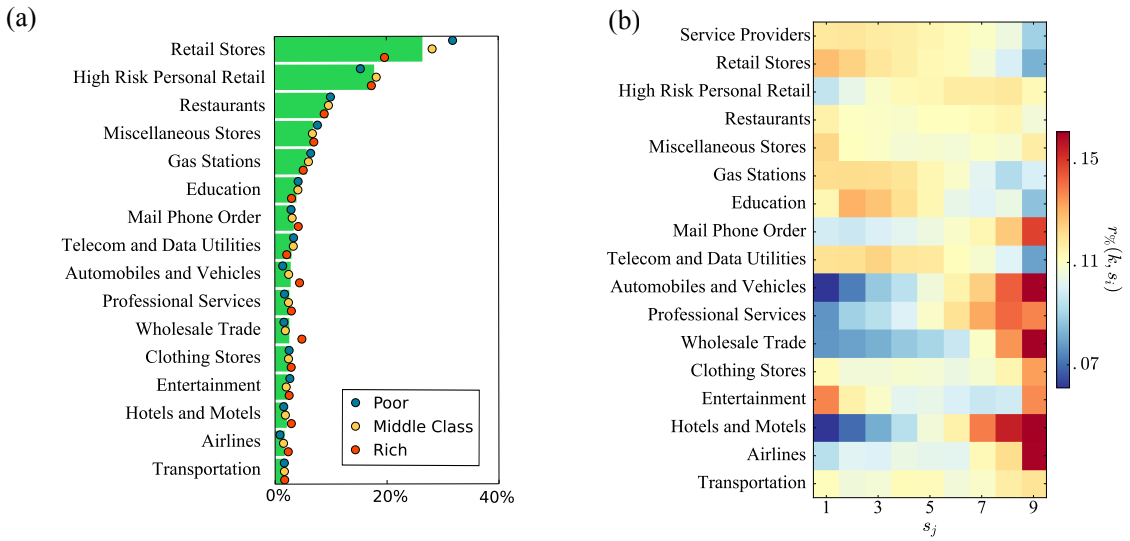



Figure 7.2.1: **Social classes spending on purchase category groups. (a)** The histogram in green shows the distribution of the total amount of money spent on the PCGs $K_{2\text{-}17}$. Blue, yellow and red dots represents the values for the social class 1 (poorest), 5 (middle) and 9 (richest). **(b)** $r(k, s_i)$ distribution of spending in a given purchase category group $k \in K_{17}$ by different classes $s_j$. Distributions are normalized as in Eq.7.1, i.e. sums up to 1 for each category.

[1]Note that in our social class definition the cumulative AMP is equal for each group and this way each group represents the same economic potential as a whole. Values shown in Fig.7.2.1b assign the total purchase of classes. Another strategy would be to calculate per capita measures, which in turn would be strongly dominated by values associated to the richest class, hiding any meaningful information about other classes.

From this first analysis we can already identify large differences in the spending behavior of people from lower and upper classes. To further investigate these dissimilarities on the individual level, we consider the $K_{2\text{-}17}$ category set as defined in Section 1.3 (category $k_1$ excluded) and build (as we did in Section 1.3 for MCCs) a spending vector $SV(u) = [SV_2(u), ..., SV_{17}(u)]$ for each ego $u$. Here each item $SV_k(u)$ determines the fraction of money $m_u^k/m_u$ what user $u$ spent on a category $k \in K_{2\text{-}17}$ out of his/her $m_u = \sum_{k \in K} m_u^k$ total amount of purchases. Using these individual spending vectors we calculate the average spending vector of a given socioeconomic class as $\overline{SV}(s_j) = \langle SV(u) \rangle_{u \in s_j}$. We associate $\overline{SV}(s_j)$ to a representative consumer of class $s_j$ and use this average vector to quantify differences between distinct socioeconomic classes as follows.

The Euclidean metric between average spending vectors is:

$$d_{SV}(s_i, s_j) = \|\overline{SV}_k(s_i) - \overline{SV}_k(s_j)\|_2, \tag{7.2}$$

where $\|\vec{v}\|_2 = \sqrt{\sum_k v_k^2}$ assigns the $L^2$ norm of a vector $\vec{v}$. Note that the diagonal elements of $d_{SV}(s_i, s_i)$ are equal to zero by definition. However, in Fig.7.2.2c the off-diagonal green component around the diagonal indicates that the average spending behavior of a given class is the most similar to neighboring classes, while dissimilarities increase with the gap between socioeconomic classes. We repeated the same measurement separately for the single category of cash purchases (PCG $k_1$). In this case Euclidean distance is defined between average scalar measures as $d_{k_1}(s_i, s_j) = \|\langle SV_1 \rangle(s_i) - \langle SV_1 \rangle(s_j)\|_2$. Interestingly, results shown in Fig.7.2.2d. indicates that here the richest social classes appear with a very different behavior. This is due to their relative under-spending in cash, which can be also concluded from Fig.7.2.1b (first row). On the other hand as going towards lower classes such differences decrease as cash usage starts to dominate.

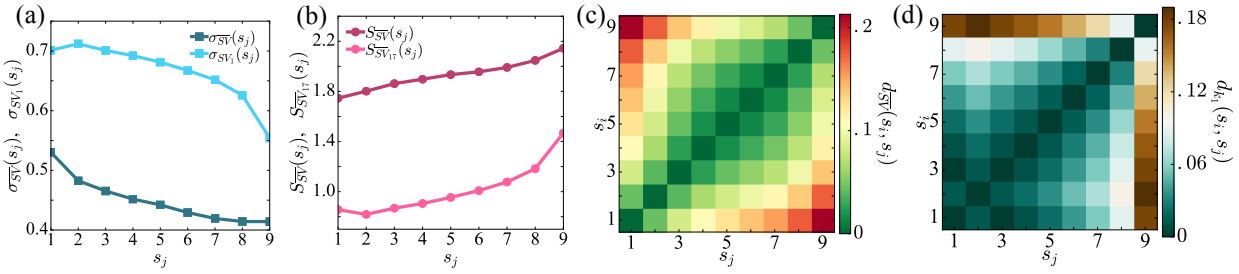



Figure 7.2.2: **Purchasing dissimilarities between social classes (a)** Dispersion $\sigma_{SV}(s_j)$ for different socioeconomic classes considering PCGs in $K_{2\text{-}17}$ (dark blue) and the single category $k_1$ (light blue). **(b)** Shannon entropy measures for different socioeconomic classes considering PCGs in $K_{2\text{-}17}$ (dark pink) and in $k_{17}$ (light pink). **(c)** (resp. **(d)**) Heat-map matrix representation of $d_{\overline{SV}}(s_i, s_j)$ (resp. $d_{k_1}(s_i, s_j)$) distances between the average spending vectors of pairs of socioeconomic classes considering PCGs in $K_{2\text{-}17}$ (resp. $k_1$).

To explain better the differences between socioeconomic classes in terms of purchasing patterns, we introduce two additional scalar measures. First, we introduce the dispersion of individual spending vectors

as compared to their class average as

$$\sigma_{SV}(s_j) = \langle \|\overline{SV}_k(s_j) - SV_k(u)\|_2 \rangle_{u \in s_j}, \tag{7.3}$$

which appears with larger values if people in a given class allocate their spending very differently. Second, we also calculate the Shannon entropy of spending patterns as

$$S_{SV}(s_j) = \sum_{k \in K_{2\text{-}17}} -\overline{SV}_k(s_j) \log(\overline{SV}_k(s_j)) \tag{7.4}$$

to quantify the variability of the average spending vector for each class. This measure is minimal if each ego of a class $s_j$ spends exclusively on the same single PCG, while it is maximal if they equally spend on each PCG. As it is shown in Fig.7.2.2a (light blue line with square symbols) dispersion decreases rapidly as going towards higher socioeconomic classes, richer people tends to be more similar in terms of their purchase behavior. On the other hand, surprisingly, in Fig.7.2.2b (dark pink line with square symbols) the increasing trend of the corresponding entropy measure suggests that even richer people behave more similar in terms of spending behavior they used to allocate their purchases in more PCGs. These trends are consistent even in case of $k_1$ cash purchase category (see $\sigma_{SV_1}(s_j)$ function depicted with dark blue line in Fig.7.2.2a) or once we include category $k_1$ into the entropy measure $S_{SV_{17}}(s_j)$ (shown in Fig.7.2.2b with light pink line).

## 7.3 PURCHASE CORRELATIONS IN SOCIOECONOMIC NETWORK

To complete our investigation we characterize the effects of social relationships on the purchase habits of individuals. We address this problem through an overall measure quantifying differences between individual purchase vectors of connected egos positioned in the same or different socioeconomic classes. More precisely, we consider each social tie $(u, v) \in E$ connecting individuals $u \in s_i$ and $v \in s_j$, and for each purchase category $k$ we calculate the average absolute difference of their purchase vector items as

$$d^k(s_i, s_j) = \langle |SV_k(u) - SV_k(v)| \rangle_{u \in s_i, v \in s_j}. \tag{7.5}$$

Following that, as a reference system we generate a corresponding configuration network by taking randomly selected edge pairs from the underlying social structure and swap them without allowing multiple links and self loops. In order to remove any residual correlations we repeated this procedure in $5 \times |E|$ times. This randomization keeps the degree, individual economic estimators $P_u$, the purchase vector $SV(u)$, and the assigned class of each people unchanged, but destroys any structural correlations between egos in the social network, consequently between socioeconomic classes as well. After generating a reference structure we computed an equivalent measure $d^k_{rn}(s_i, s_j)$ but now using links $(u, v) \in E_{rn}$ of the randomized network. We repeated this procedure 100 times and calculated an average $\langle d^k_{rn} \rangle (s_i, s_j)$. In order to quantify

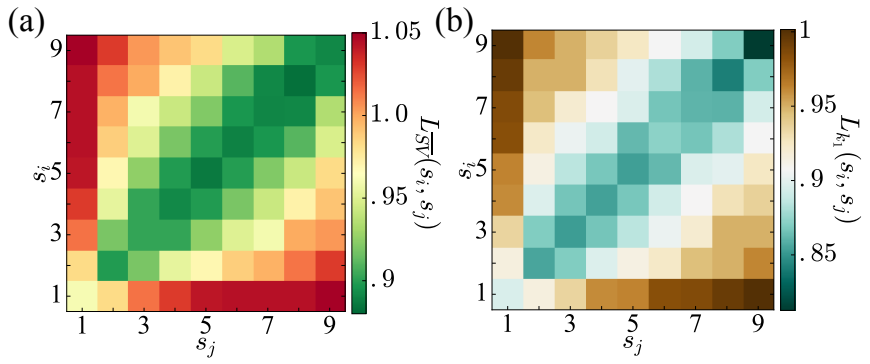



Figure 7.3.1: **Consumption correlations in the socioeconomic network.** **(a)** (resp. **(b)**) Heat-map matrix representation of the average $L_{\overline{SV}}(s_i, s_j)$ (resp. $L_{k_1}(s_i, s_j)$) measure between pairs of socioeconomic classes considering PCGs in $K_{2\text{-}17}$ (resp. $k_1$).

the effect of the social network we simply take the ratio

$$L_k(s_i, s_j) = \frac{d^k(s_i, s_j)}{\langle d^k_{rn} \rangle (s_i, s_j)} \tag{7.6}$$

and calculate its average $L_{SV}(s_i, s_j) = \langle L_k(s_i, s_j) \rangle_k$ over each category group $k \in K_{2\text{-}17}$ or respectively $k_1$. This measure shows whether connected people have more similar purchasing patterns than one would expect by chance without considering any effect of homophily, social influence or structural correlations. Results depicted in Fig.7.3.1a and 7.3.1b for $L_{SV}(s_i, s_j)$ (and $L_{k_1}(s_i, s_j)$ respectively) indicates that the purchasing patterns of individuals connected in the original structure are actually more similar than expected by chance (diagonal component). On the other hand people from remote socioeconomic classes appear to be less similar than one would expect from the uncorrelated case (indicated by the $L_{SV}(s_i, s_j) > 1$ values typical for upper classes in Fig.7.3.1a). These observations do not clearly assign whether homophily [87, 99] or social influence [39] induce the observed similarities in purchasing habits but undoubtedly clarifies that social ties (i.e. the neighbors of an ego) and socioeconomic status play deterministic roles in the emerging similarities in consumption behavior.

Note that we found the same correlation trends in cash purchase patterns as shown in Fig.7.3.1b and more generally, as we observe in Fig.A.3.1 (see appendix), the correlation holds for each of the PCGs. Some category like *Education* seems to be more influenced by the ties than others like *Retail Stores*. In order to quantify the influence of ties on purchase habits at the PCG level, giving a network $G = (V, E)$ and a PCG $c_i$, we introduce a correlation measure $\rho$ defined as follow :

$$\rho(c_i) = |E| \frac{(\sum_{(u,v)\in E} r(c_i, u) r(c_i, v))}{(\sum_{(u,v)\in E} r(c_i, u))(\sum_{(u,v)\in E} r(c_i, v))}. \tag{7.7}$$

$\rho(c_i)$ quantifies the tendency that two connected egos $(u, v)$ spend commonly in a same category $c_i$.

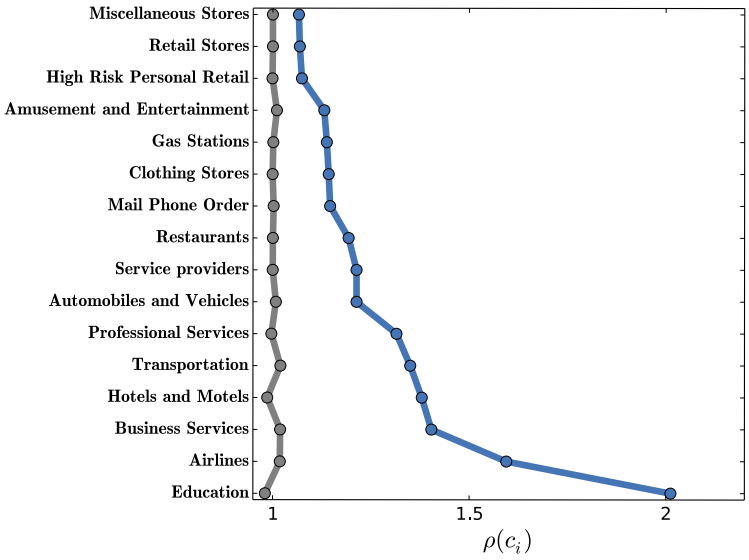



Figure 7.3.2: **Network influence quantification.** Correlation measure $\rho(c_i)$ computed for real communication network (in blue) and for configuration model (in grey) for each PCG $k \in K_{2-17}$ sorted from smallest to greatest values of real network.

If ties do not have any influence on purchase habits, $r(c_i, u)$ and $r(c_i, v)$ are independent. It follows $\frac{\sum_{(u,v)\in E} r(c_i,u)r(c_i,v))}{|E|} = \frac{(\sum_{(u,v)\in E} r(c_i,u))}{|E|} \frac{(\sum_{(u,v)\in E} r(c_i,v))}{|E|}$ and so $\rho(c_i) = 1$. Finally, an influence of the network on purchase behavior is characterized by $\rho(c_i) > 1$. In Fig.7.3.2, given the communication network, the values of the correlation measure $\rho(c_i)$ are all greater than 1, meaning that purchases are correlated with communication network. As intended, for a configuration model graph that contain random ties, the measure $\rho$ is always close to 1 as the ties are not correlated to purchases. At the PCG level, we demonstrate that the intensity of the correlation between ties and purchase patterns greatly depend on the category. Some purchase categories groups like *Education*, *Airlines*, *Business Services* or *Hotels* are well correlated with ties ($\rho \approx 2$). Interestingly, ties have not much impact on daily supermarket spendings whereas it plays an important role on education, business and traveling.

## 7.4 PURCHASE CATEGORY CORRELATIONS

To study consumption patterns of single purchase categories PCGs provides a too coarse grained level of description. Hence, to address our second question we use DS2 and we downscale from the category group level to the level of single merchant categories. We are dealing with 271 categories after excluding some with less than 100 purchases and the categories linked to money transfer and cash retrieval (for a complete list of IDs and name of the purchase categories considered see Table A.2.1). As explained in Chapter 1 we assign to each ego $u$ a personal vector $PV(u)$ of four socioeconomic features: the age, the gender, the social economic group, and the distribution $r(c_i, u)$ of purchases in different merchant categories made by the central ego. Our aim here is to obtain an overall picture of the consumption structure at the level of

merchant categories and to understand precisely how personal and socioeconomical features correlate with the spending behavior of individuals and with the overall consumption structure.

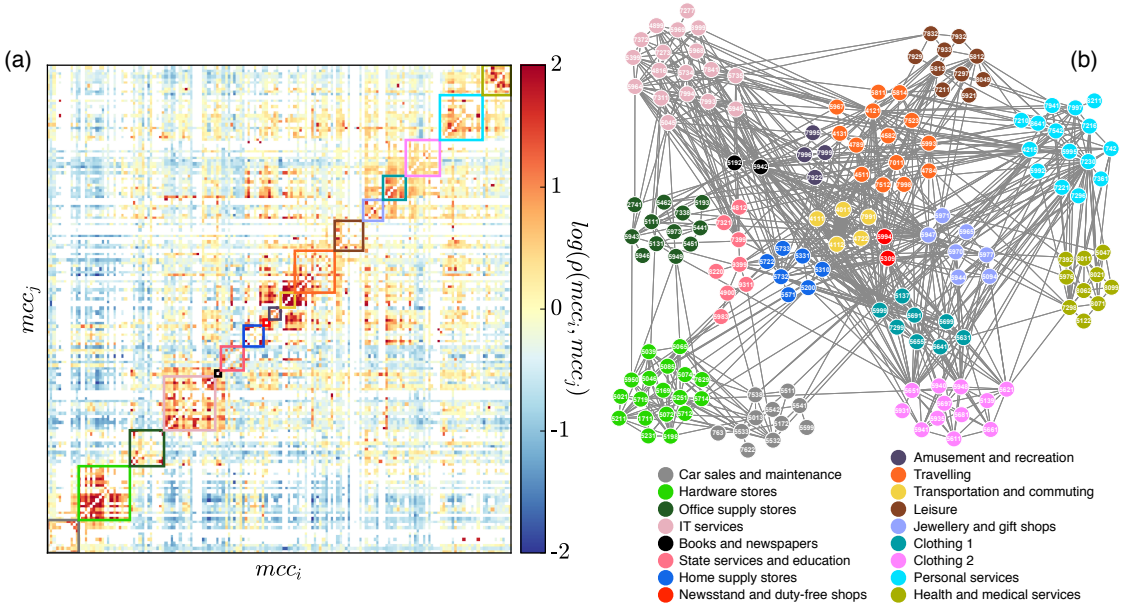



Figure 7.4.1: **Merchant category correlation matrix and graph. (a)** 163×163 matrix heatmap plot corresponding to $\rho(c_i, c_j)$ correlation values (see Eq. 7.8) between categories. Colors scale with the logarithm of correlation values. Positive (resp. negative) correlations are assigned by red (resp. blue) colors. Diagonal components represent communities with frames colored accordingly.**(b)** Weighted $G_\rho^>$ correlation graph with nodes annotated with MCCs (see Table A.2.1). Colors assign 17 communities of merchant categories with representative names summarized in the figure legend.

As we introduced in chapter 1, the purchase spending vector $r(c_i, u)$ of an ego quantifies the fraction of money spent on a category $c_i$. Using the spending vectors of $n$ number of individuals we define an overall correlation measure between categories as

$$\rho(c_i, c_j) = \frac{n(\sum_u r(c_i, u) r(c_j, u))}{(\sum_u r(c_i, u))(\sum_u r(c_j, u))}. \tag{7.8}$$

This symmetric formula quantifies how much people spend on a category $c_i$ if they spend on an other $c_j$ category or vice versa. Therefore, if $\rho(c_i, c_j) > 1$, the categories $c_i$ and $c_j$ are positively correlated and if $\rho(c_i, c_j) < 1$, categories are negatively correlated. Using $\rho(c_i, c_j)$ we can define a weighted correlation graph $G_\rho = (V_\rho, E_\rho, \rho)$ between categories $c_i \in V_\rho$, where links $(c_i, c_j) \in E_\rho$ are weighted by the $\rho(c_i, c_j)$ correlation values. The weighted adjacency matrix of $G_\rho$ is shown in Fig.7.4.1a as a heat-map matrix with logarithmically scaling colors. Importantly, this matrix emerges with several block diagonal components suggesting present communities of strongly correlated categories in the graph.

To identify categories which were commonly purchased together we consider only links with positive correlations. Furthermore, to avoid false positive correlations, we consider a $10\%$ error on $r$ that can induce,

in the worst case $50\%$ overestimation of the correlation values. In addition, to consider only representative correlations we take into account category pairs which were commonly purchased by at least $1000$ consumers. This way we receive a $G_\rho^>$ weighted sub-graph of $G_\rho$, shown in Fig.7.4.1b, with 163 nodes and 1664 edges with weights $\rho(c_i, c_j) > 1.5$.

To identify communities in $G_\rho^>$ indicated by the correlation matrix in Fig.7.4.1a we applied a graph partitioning method based on the Louvain algorithm [14]. We obtained 17 communities depicted with different colors in Fig.7.4.1b and as corresponding colored frames in Fig.7.4.1a. Interestingly, each of these communities group a homogeneous set of merchant categories, which could be assigned to similar types of purchasing activities (see legend of Fig.7.4.1b). In addition, this graph indicates how different communities are connected together. Some of them, like *Transportation, IT* or *Personal Serv.* playing a central role as connected to many other communities, while other components like *Car sales and maintenance* and *Hardware St.*, or *Personal* and *Health and medical Serv.* are more like pairwise connected. Some groups emerge as stand-alone communities like *Office Supp. St.*, while others like *Books and newspapers* or *Newsstands and duty-free Shops (Sh.)* appear as bridges despite their small sizes.

Note that the main categories corresponding to everyday necessities related to food (*Supermarkets*, *Food St.*) and telecommunication (*Telecommunication Serv.*) do not appear in this graph. Since they are responsible for the majority of total spending, they are purchased necessarily by everyone without obviously enhancing the purchase in other categories, thus they do not appear with strong correlations.

Finally we turn to study possible correlations between purchase categories and personal features. An average feature set $AFS(c_i) = \{\langle age(c_i)\rangle, \langle gender(c_i)\rangle, \langle SEG(c_i\}\rangle)$ is assigned to each of the 271 categories. The average $\langle v(c_i)\rangle$ of a feature $v \in \{age, gender, SEG\}$ assigns a weighted average value computed as:

$$\langle v(c_i)\rangle = \frac{\sum_{u\in\{u\}_i} \alpha_i(v_u) v_u}{\sum_{u\in\{u\}_u} \alpha_i(v)}, \tag{7.9}$$

where $v_u$ denotes a feature of a user $u$ from the $\{u\}_i$ set of individuals who spent on category $c_i$. Here

$$\alpha_i(v_u) = \sum_{(u\in\{u\}_i | v_u = v)} \frac{r(c_i, u)}{n_i(v_u)} \tag{7.10}$$

corresponds to the average spending on category $c_i$ of the set of users from $\{u\}_i$ sharing the same value of the feature $v$. $n_i(v_u)$ denotes the number of such users. In other words, e.g. in case of $v = age$ and $c_{742}$, $\langle age(c_{742})\rangle$ assigns the average age of people spent on Veterinary Services ($mcc = 742$) weighted by the amount they spent on it. In case of $v = gender$ we assigned 0 to females and 1 to males, thus the average gender of a category can take any real value between $[0, 1]$, indicating more females if $\langle gender(c_i)\rangle \leq 0.5$ or more males otherwise.

We visualize this multi-modal data in Fig.7.4.2a as a scatter plot, where axes scale with average age and SEG, while the shape and size of symbols correspond to the average gender of each category. To

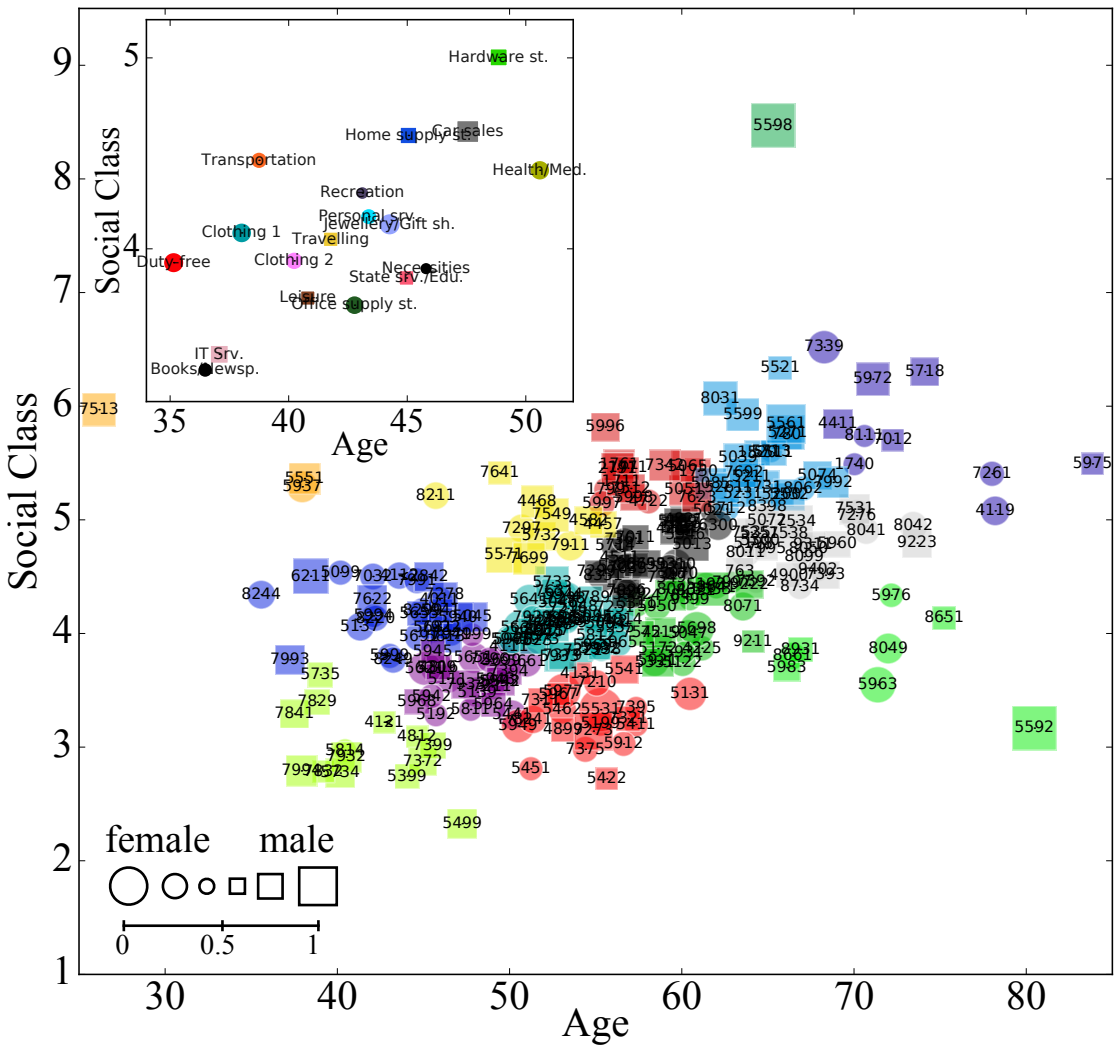



Figure 7.4.2: **Socioeconomic parameters of merchant categories. (a)** Scatter plot of $AFS(c_i)$ triplets (for definition see Eq. 7.9 and text) for 271 merchant categories summarized in Table A.2.1. Axis assign average age and SEG of purchase categories, while gender information are assigned by symbols. The shape of symbols assigns the dominant gender (circle-female, square-male) and their size scales with average values. **(b)** Similar scatter plot computed for communities presented in Fig.7.4.1b. Labels and colors are explained in the legend of Fig.7.4.1a.

further identify correlations we applied k-means clustering [12] using the $AFS(c_i)$ of each category. The ideal number of clusters was 15 according to several criteria: Davies-Bouldin Criterion, Calinski-Harabasz criterion (variance ratio criterion) and the Gap method [150]. Colors in Fig.7.4.2a assign the identified k-mean clusters.

The first thing to remark in Fig.7.4.2a is that the average age and SEG assigned to merchant categories

are positively correlated with a Pearson correlation coefficient $0.42$ ($p < 0.01$). In other words, elderly people used to purchase from more expensive categories, or alternatively, wealthier people tend to be older, in accordance with our intuition. At the same time, some signs of gender imbalances can be also concluded from this plot. Wealthier people appear to be commonly males rather than females. A Pearson correlation measure between gender and SEG, which appears with a coefficient $0.29$ ($p < 0.01$) confirmed it. On the other hand, no strong correlation was observed between age and gender from this analysis.

To have an intuitive insight about the distribution of merchant categories, we take a closer look at specific category codes (summarized in Table A.2.1). As seen in Fig.7.4.2a elderly people tend to purchase in specific categories such as *Medical Serv.*, *Funeral Serv.*, *Religious Organizations*, *Motorhomes Dealers*, *Donation*, *Legal Serv.*. Whereas categories such as *Fast Foods*, *Video Game Arcades*, *Cinema*, *Record St.*, *Educational Serv.*, *Uniforms Clothing*, *Passenger Railways*, *Colleges-Universities* are associated to younger individuals on average. At the same time, wealthier people purchase more in categories as *Snowmobile Dealers*, *Secretarial Serv.*, *Swimming Pools Sales*, *Car Dealers Sales*, while poorer people tend to purchase more in categories related to everyday necessities like *Food St.*, *General Merch.*, *Dairy Products St.*, *Fast Foods* and *Phone St.*, or to entertainment as *Billiard* or *Video Game Arcades*. Typical purchase categories are also strongly correlated with gender as categories more associated to females are like *Beauty Sh.*, *Cosmetic St.*, *Health and Beauty Spas*, *Women Clothing St.* and *Child Care Serv.*, while others are preferred by males like *Motor Homes Dealers*, *Snowmobile Dealers*, *Dating/Escort Serv.*, *Osteopaths*, *Instruments St.*, *Electrical St.*, *Alcohol St.* and *Video Game Arcades*.

Finally we repeated a similar analysis on communities shown in Fig.7.4.1b, but computing the $AFS$ on a set of categories that belong to the same community. Results in Fig.A.3.1b disclose positive age-SEG correlations as observed in Fig.7.4.2a, together with somewhat intuitive distribution of the communities.

## 7.5 Discussions

In this chapter, we analyzed a multi-modal dataset collecting the mobile phone communication and bank transactions of a large number of individuals living in Mexico. This corpus allowed for an innovative global analysis both in term of social network and its relation to the economical status and merchant habits of individuals. We introduced several measures to estimate the socioeconomic status of each individual together with their purchasing habits. Using these information:

- we identified distinct socioeconomic classes, which reflected strongly imbalanced distribution of purchasing power in the population. After mapping the social network of egos from mobile phone interactions,
- we showed that typical consumption patterns are strongly correlated with the socioeconomic classes and the social network behind. We observed these correlations on the individual and social class level.

In the second half of our study:

- we detected correlations between merchant categories commonly purchased together and introduced a correlation network which in turn emerged with communities grouping homogeneous sets of categories,
- we further analyzed some multivariate relations between merchant categories and average demographic and socioeconomic features, and found meaningful patterns of correlations giving insights into correlations in purchasing habits of individuals.

We identified several new directions to explore in the future. One possible track would be to better understand the role of the social structure and interpersonal influence on individual purchasing habits, while the exploration of correlated patterns between commonly purchased brands assigns another promising directions. Beyond our general goal to better understand the relation between social and consuming behavior these results may enhance applications to better design marketing, advertising, and recommendation strategies, as they assign relations between co-purchased product categories.

# 8

# Impact of university admission on freshmen's egocentric network

The social network is evolving over time. Some social relations are appearing and others are disappearing as a sequence of personal events. Even if such dynamical changing network describes how people divide their communication effort and exchange between each others, there is still much to discover. In this chapter, from DS6 (presented in Section 1.6), we are tracking 1675 students that are entering in the university. We point out that life events such as the entrance to university has a direct impact on the evolution of the individual social network. Our study shows that the way how students communicate, quantified as their social signature, differs from one to another and is persistent.

## 8.1 Introduction

Social relations are well-known to bring significant individual and collective benefits. The intimate, relational and collective connectedness respectively lets people affirm their personality, mutually has rewarding contacts and feels being a part of a group, and so, increases his or her social capital [56]. On the contrary, feeling isolated lowers overall subjective well-being. More generally, the communication behavior is a key component in the characterization of social personality types of humans [103] and even animals [161]. Recently due to Internet and the improvements of transportation services, the number of communication means to maintain real or virtual contacts has been blowing up in developed and developing countries.

Yet, maintaining social relations require a lot of effort and time while human individual capacities are

limited in time [104, 131] and space [101]. More precisely, the social brain hypothesis [50] says that there is a quantitative relationship between primate brain size and social group size. For these reasons, individuals have to make choices and adopt a social strategy. The intensity and number of social relations can be specific to anyone but limitations bring people replace old relations by new interactions to maintain a reasonable and manageable amount of relations [103, 134].

In this chapter, we discuss two main questions about communication efforts. As individual social strategy may evolve according to time, turnovers are the major witnesses of this evolution. Like the social strategy, the frequency of social changes can differ from an ego to another. Here, we quantify how much social strategies of egos are specific and persistent in time. We establish the causal link between the network position and the social strategy by observing the influence of a major life event (entering the university) on the growth of the social network and its turnover.

An interesting study already catch the persistency of the social strategy, called social signature, for a small number of egos [134]. In [103], the time scale of the turnovers is analyzed but to our knowledge, there is no such data-driven approach that investigate the influence of the network and life major event on communication effort.

In this chapter, we have two axis. We validate hypothesis on the large scale dataset and at the same time bring new insights concerning the social communication strategy. Starting from DS6 (presented in section 1.6), we present our results concerning the social signature and its link with the network position in section 8.2. Creations and decays of ties in the student ego network while entering the university is investigated in section 8.3 giving insight on the speed and the temporal evolution of turnovers. We finally discuss our conclusions and future works in section 8.4.

## 8.2 Social signature

### 8.2.1 Definitions and notations

The first part of this chapter is dealing with social signature that we define by introducing some notations for the communication network. Considering the weighted communication graph $G = (E, V, w)$ where $V$ is the set of phone users in the datasets and $E$ the set of links such as $(u, v) \in E$ if and only if $u$ and $v$ has at least one contact during the considered period. For this link $(u, v)$, we note $w_{(u,v)}$ the number of undirected communication events (calls and SMS) between $u$ and $v$. For each ego $u$, $N(u) = \{v|(u, v) \in E\}$ represents the set of nodes that have at least one communication with $u$. As the ego network changes over time, we may induce a graph $G_T$ for each specific period $T$. For example, by taking a set of one-month periods $T_1, ..., T_{22}$, we induce a set of graphs observed in these periods $G_{T_1}, ..., G_{T_{22}}$ and a set of neighborhood for each node $u$, $N_{T_1}(u), ..., N_{T_{22}}(u)$.

The social signature is defined for a client $u$ and for a period $T_i$, and noted $\sigma_u^{T_i}$, as the distribution of weights decreasingly sorted by weights $w_{(u,v)}$. In Fig. 8.2.1, there is an example of a social signature distribution (in (b)) generated from an ego network (in (a)). The social signature is a normalized distribution,

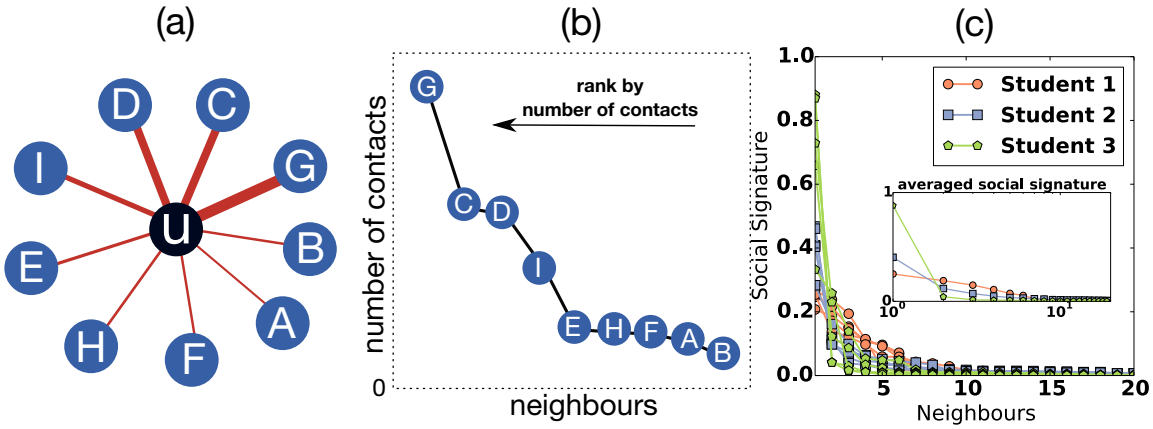



Figure 8.2.1: **Social signature definition and examples.** An example of an ego network with different weight on links is depicted in **(a)**. The social signature generated from the considered ego network is shown in **(b)**. All the contacts are decreasingly sorted according to the number of interactions with the ego. The ego exchanges more with $G$ than with the others. In **(c)**, social signatures of three random students for 5 months (15 curves) show the shape diversity. The average of the social signatures of three random students per month in log-lin scale is visible in the inset for the three students.

the first point represents the percentage of the total communication assigned to the most intensive relation and it is a decreasing function from the most intensive relation to the less involved one.

### 8.2.2 Temporal persistence of the social signature

To validate the hypothesis about the persistency of social signature of egos, as shown in [134], we compare the average distance of social signatures of a same ego over two consecutive periods to the average distance of the social signature of different egos (Fig. 8.2.2a). In practice, each period represents 3 months of the trace. In this experiment, we use 18 months from April 2014 to September 2015. The social signature is directly evaluated over these 6 periods of 3 months periods $P_1, .., P_6$.

Because of the shape and properties of social signature distributions, we use the Jensen-Shannon Divergence (JSD) for comparing two social signatures $\sigma_1$ and $\sigma_2$:

$$\delta(\sigma_1, \sigma_2) = H(\frac{1}{2}\sigma_1 + \frac{1}{2}\sigma_2) - \frac{1}{2}(H(\sigma_1) + H(\sigma_2)) \tag{8.1}$$

For a probability distribution $\sigma$, the entropy $H(\sigma)$ is expressed by the formula:

$$H(\sigma) = -\sum_u \sigma(u) log(\sigma(u)) \tag{8.2}$$

The JensenShannon divergence is an extension of the KullbackLeibler divergence (KLD). $\delta(\sigma_1, \sigma_2) = 0$ if and only if $\sigma_1 = \sigma_2$. In this chapter, distributions do not have the same support range and so there are many zeros if we compare them on the union support. Interestingly, the JensenShannon divergence is not

too sensitive to the zeros of the distribution compare to classical Euclidean distance. Thus, the difference between social signatures of two different egos $\delta(\sigma_u^{P_i}, \sigma_v^{P_i})$ and between two consecutive periods of the same ego $\delta(\sigma_u^{P_i}, \sigma_u^{P_{i+1}})$. Low value of $\delta(\sigma_u^{P_i}, \sigma_v^{P_i})$ means a small difference between social signatures of students $u$ and $v$. In this case, communication efforts of $u$ and $v$ are quite similar. Low value of $\delta(\sigma_u^{P_i}, \sigma_u^{P_{i+1}})$ reveals social strategy similarities among consecutive periods $P_i$ and $P_{i+1}$ for the ego $u$. In Fig. 8.2.2a, on average, $\delta(\sigma_u^{P_i}, \sigma_v^{P_i}) = 2.11 \times \delta(\sigma_u^{P_i}, \sigma_u^{P_{i+1}})$. As the gap between self distance and student-student distance is marked, we confirm empirically that the social strategy of individuals tends to persist in time. Even if there is turnover, the shape of social signatures is a personal characteristic which does not vary much. The social signature is a specific characteristc of each individuals. Furthermore, this result holds for each of the periods $P_i$. The social strategy of an individual is changing and evolve from a period to another. This observed variation can be induced by turnovers or even intensity fluctuations. This question will be pushed in Section 8.3.

### 8.2.3 Typical social signatures

In Fig. 8.2.2a, we noted that the social signature is persistent in time for an ego and changes a lot from an ego to another. Some students are focusing more on a small number of intense relations and others are spreading their effort over their neighborhood. The social signature determines the specific way of sharing the communication effort over the ties. As we observed earlier, social signature is a personal characteristic that does not change in time. In Fig. 8.2.2b, we generate the distance matrix that represents a full pairwise comparison of social signatures between 1675 students. Rows and columns are identically sorted according to a hierarchical clustering algorithm. Finally the matrix is organized with diagonal blocks revealing clusters of students. To identify non-overlapping clusters from the distance matrix, we used hierarchical tree and obtain 20 groups highlighted with white squares. As we can note, the major color is blue inside biggest white boxes, meaning that pairwise distances are small between people of the same group and red outside the group. Social signatures can be clustered in groups revealing maybe several interesting behaviors. We did not go further in this direction even if it opens up interesting questions.

### 8.2.4 Network position and social signature

The network position is used in order to define the social signatures of egos but it may also have an influence on how egos spread their communication effort. More precisely, social relations are well-known to not to be random, the network certainly contains overlapping communities of people that shared the same interests. We discuss in this section about the influence of the network position on the social signature of an ego. Is there a correlation between the social signature of an ego and the social signature of his or her neighborhood? Even if we just point out the diversity of social strategies, some students are closer to others. If an ego communicate only with very close relations, these neighbors may have the same habits.

To verify this hypothesis, we consider the network $G = (E, V)$ defined above. $E$ is the set of connected

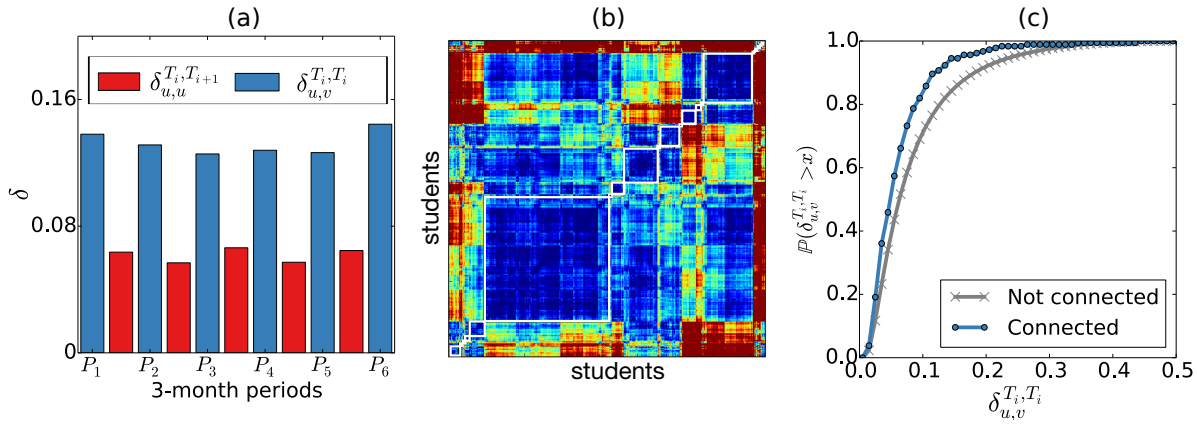



Figure 8.2.2: **Persistence and network influence of the social signature.** As depicted in **(a)**, the distance $\delta$ between two egos using JSD (in blue) and between two consecutive 3-month periods $P_i$ (in red) are evaluated for each period and student. The histogram shows the average values for each pair of students and for each consecutive 3-month periods. **(b)** Distance matrix representing Jensen-Shannon distances between social signatures of each pair of students. Blue color means small distance while red color corresponds to large distance. **(c)** Considering a unique social signature for each student that consists in the average over the six 3-month periods. Curves represent the cumulative distribution of distances of these averaged social signatures between connected pairs (in blue) and not connected pairs (in grey).

pairs $(u, v)$ and we note $\overline{E}$, the complement of $E$, the set of not connected pairs i.e. $\forall (u, v) \notin E, (u, v) \in \overline{E}$. Then, we compute the distances between averaged social signature for all pairs of $E$ and $\overline{E}$ separately. In Fig. 8.2.2c, we plot the cumulative distribution $\mathbb{P}(\delta_{u,v}^{T_i,T_i} > x)$ of distance values of connected pairs and not connected pairs. The connected pairs have smaller distances than not connected pairs as the cumulative function of connected pairs is constantly above. On average, the social signature of an ego is $42\%$ closer to his or her neighbors' social signatures than to others.

## 8.3 Study of turnovers

As we noticed in Section 8.2, the social signature of individuals varies over time. The variation is partly due to turnover: new ties that replace old relations. Therefore, decay and formation of communication ties traduce the evolution of the ego social network. As the time scale of turnovers is considered very long, studies usually neglect decays and formations of ties for studying diffusion processes or community detection.

### 8.3.1 Monthly ego network variations

To study the impact of entering the university, we take the selected set of students and compare our results to a set of a random sample of non student clients. This sample is extracted from the datasets by taking randomly the same number of egos with a similar activity over the whole period such as the distributions

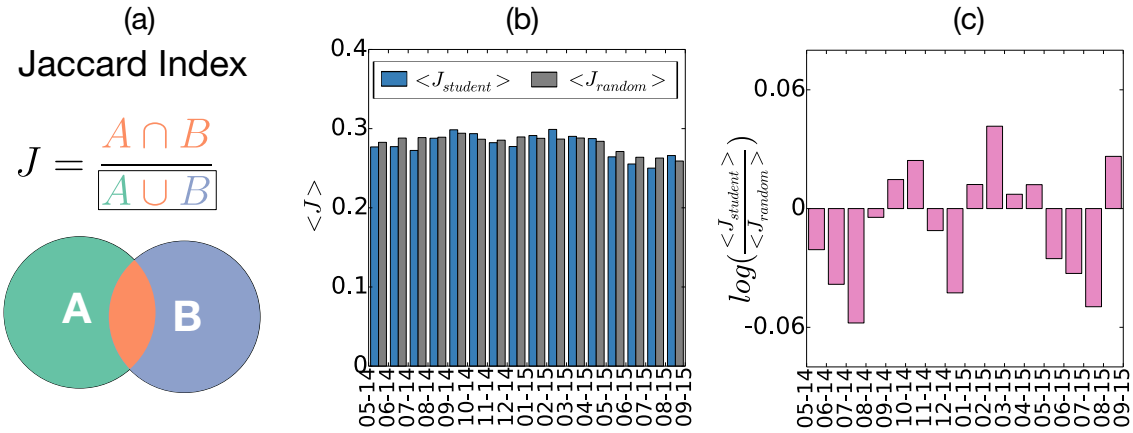



Figure 8.3.1: **Turnovers over time measure with Jaccard Index.** The definition of the Jaccard Index depicted in **(a)** is very intuitive as it only computes intersection between two samples of items like nodes of the neighborhood of $u$ in **(b)** where $A = N_{T_i}(u)$ and $B = N_{T_{i+1}}(u)$. The average of the Jaccard Index quantifies, for a student (in blue) or for a random node (in grey), how much the listing of contacts has changed between two consecutive months. In **(c)**, changes for students and for the random sample are compared by taking the logarithm of the averaged Jaccard Index. In particular, when the comparison value is greater than 0, there are more turnovers for students.

of the degree and activity without considering the temporarily are similar. This method is possible because the number of students is way smaller than the total number of clients, and so, taking an ego with the same degree and activity is feasible. The egos contained in this sample are not students and are not evidently going through a major life event (more details in Section 1.6).

The first study consists of quantifying the difference between the previous and the present ego network. By cutting time in months, we measure the similarity of two consecutive ego networks for each student and for each ego of the sample. A bunch of measures are available from the literature to quantify the difference between two different sets of nodes that we note $N_1$ and $N_2$. The Jaccard Index $J(N_1, N_2)$, shown in Fig. 8.3.1a, is one the simplest as it is the ratio of common nodes divided by the total number of nodes:

$$J(N_1, N_2) = \frac{N_1 \bigcap N_2}{N_1 \bigcup N_2} \tag{8.3}$$

The Jaccard Index is a simple and intuitive measure. When $N_1 = N_2$, $J(N_1, N_2)$ is equal to 1 and when $\forall u \in N_1, n \notin N_2$, $J(N_1, N_2)$ is equal to 0. For an ego, the Jaccard Index between two consecutive month is the number of common relations divided by the total number of relations, it is equal to 0 if two sets have a null intersection and 1 if there are the identical. In Fig. 8.3.1b, the averaged Jaccard Index for each month is computed for the students (in blue) and the sample of egos (in grey). Here, we see that for each month, the value of the Jaccard Index are quite high around 0.26. On average, 26% of relations are active in two consecutive months. With this static analysis it is difficult to conclude if it is due to a quick change of the ego network, or the opposite, if the chosen time window (1 month) is too short and induce too much noise.

In Fig. 8.3.1c (in pink), we compare the results for students and for the random set. The logarithm of the ratio of averaged Jaccard Indexes is positive if the similarity of ego network between consecutive months larger for students and negative if not. We note that during the beginning of the semesters (08-2014, 02-2015, 08-2016), the students not change as often their ties between two months as the Jaccard Index is greater. Again here, it does not means that the global evolution is greater or smaller as we only have results at a 1-month level. During periods where information is needed from other student peers or students concentrate more on their studies, similarity is large while during periods before they entering the university or the exam period, the turnover is large. For example, if an ego is in a new environment during a 3-month period, he or she may be in contact exclusively with people around him or her but it does not imply that he or she has no other contacts out of these persons before and after this 3-month event. Therefore having a greater or smaller Jaccard Index during several periods does not give insights rather the ego network globally change a lot or not. Yet, it testifies periodical differences between the two samples that are interesting to note.

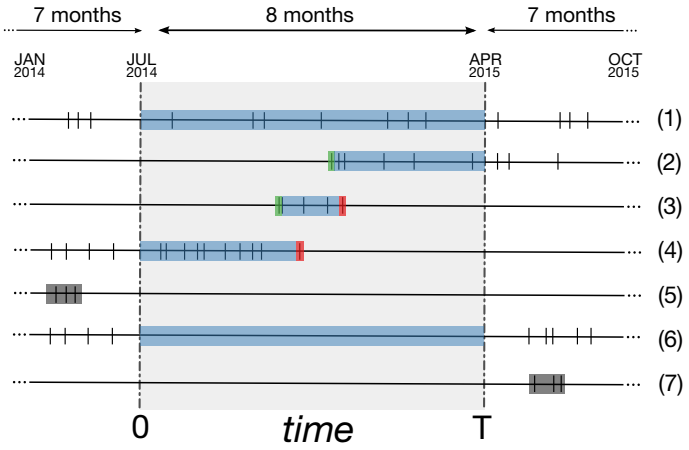



Figure 8.3.2: **Two referenced periods and one measurement period.** The above schema shows the two 7-month referenced periods that determine a stable set of contacts for each ego before and after the experimental period. Ties can be considered as stable if there are interactions in both previous and late referenced periods (case 1 and 6). If there are only contacts in one referenced period, the tie is dropped (case 5 and 7). Yet, if there are contacts only in the previous (respectively late) period and during the 8-month experiment, the tie is deactivated (resp. activated) like in case 4 (resp. case 2). We can note that if there are only contacts during the experimental period, the contacts is activated till the latest contact when it becomes deactivated (case 3). Permanent ties are represented in blue, activations of ties in green, deactivations in red and dropped ties in grey.

### 8.3.2 Temporal formations and decays of communication ties

To catch the ego network evolution, we make a continuous temporal analysis of decays and creations of ties. We consider three main periods depicted in Fig. 8.3.2: one experiment period and two reference periods. The first reference period, from January 2014 to end of July 2014, fixes an initial ego network. This period

has to be quite long in order to avoid missing strong ongoing relations. Surely, in some cases, we are missing some relations and it seems difficult to quantify the ratio of lost previous relations. The longer is the period, the more complete is our initial set of relations. The second reference period positioned at the end of the experiment period of 7 months, from April 2015 to October 2015, gives the opportunity to confirm permanent interactions . To validate creation of links that we have registered during the experimental period or to detect decays. Then, from these two referenced periods, we make a temporal analysis over the experimental period, counting creations and decays of social ties.

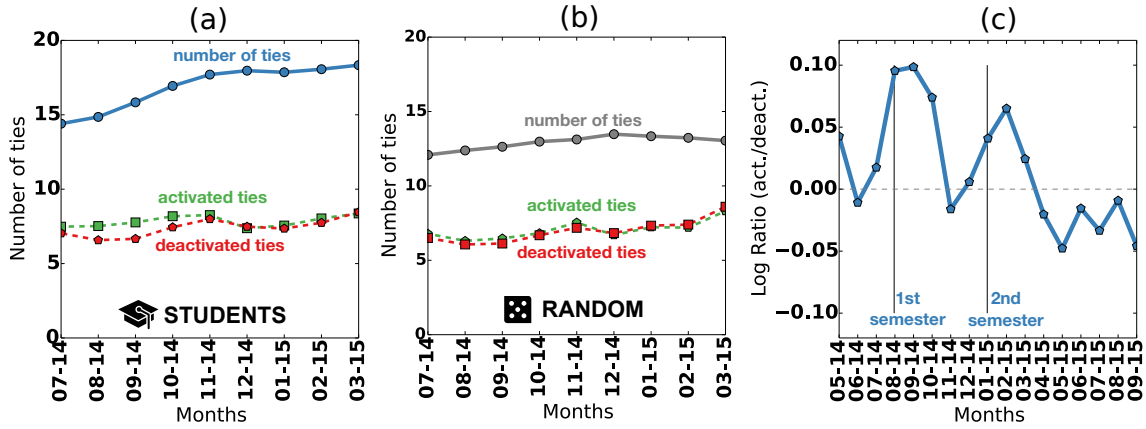



Figure 8.3.3: **Formation and decay of communication ties. (a)** For the 1675 students and for each month, the total number of ties (in blue), the number of tie activations (in green) and the number of tie deactivations (in red) **(b)** For the random sample of clients and for each month, the total number of ties (in grey), the number of tie activations (in green) and the number of tie deactivations (in red) **(c)** The pink line compare the amount of turnovers between the student sample and the random sample for each month. As we compute the logarithm of the average of the ratios of formation over decays, when the compared value is positive, there are more turnovers in the student ego network during two consecutive months.

More precisely, when a tie is active during the initial reference period and during the experimental period but not during the final referenced period, we identify a decay. At the opposite, if the observation of a social tie is during the experimental and continues in the final reference period it is a novel tie. With a global point of view, it is possible to categorize ties as a decay, a creation or a permanent tie. Yet, there are way more categories as we can note in Fig. 8.3.2. A tie that is active only during the experimental period can be considered as a created tie till the last interaction from where it will be considered as a decay.

Counting decays and formations of ties for each day of the experimental period is possible like in Fig. 8.3.3(a,b). For the students and for the random sample, the number of activated links (in green) and deactivated links (in red) are close over the whole period. It confirms the hypothesis that consists of replacing a deactivated relation by creating a new one in order to keep a stable amount of relations. The shapes of the curve that represent the number of ties also shows that at a year level, the global number of ties is quite stable and is around 15. Each month, on average, people tend to create and destroy 7 ties. It seems also that the rhythm of turnover is quite high. Going further, we can point out small increasing of the number of ties

for students and a stable number of ties for the random set.

In Fig. 8.3.3c, we compare for each month the activated-deactivated ratio of ties between the students and the random sample. The logarithm comparison is defined as follow:

$$D(t) = log(\frac{l_{+}^{students}(t)}{l_{-}^{students}(t)} \times \frac{l_{-}^{random}(t)}{l_{+}^{random}(t)}) \tag{8.4}$$

where $t$ represents the considered period and $l_{+}^{N}$ (resp. $l_{-}^{N}$) represents the averaged number of activated (resp. deactivated) ties during the period $t$ for the set of egos $N$. In our case, the contacts are aggregated by months. When $D(t)$ is positive, the students are generating more ties than the random sample. We can note two positive peaks in Fig. 8.3.3c denoting a greater progression of the number of ties for students. Interestingly, these peaks are happening just after the beginning of the first and the second semesters in August 2014 and January 2015. Our experiment shows not only that during this freshman year, the total amount of ties can increase but also that the evolution is not uniform over the year and is characterized by several peaks. On average, during the first year, a student gain three new relations whereas it is stable for the egos of the sample.

## 8.4 Discussion

Our results confirm four insights from previous works [103, 134] on a large scale datasets:

- the communication effort of people is not uniformly distributed, people focus their effort on a small number of relations that may be family or best friends. The shape of the social signature is not well-balanced and reveal an unequal spread of effort,
- at the individual level, the social signature is constantly changing. As we noticed, the variation of the ego network per month is quite important,
- the social signature of an ego is persistent over time as it varies less as a function of time than between two individuals. The communication effort is specific to each ego,
- the amount of relations is increasing/decreasing very slowly over time. For the random sample, it is kept constant during the whole 8-month experimental period and is equal to 15,

and establish three new findings:

- social signatures can be clustered in groups. We point out typical behaviors even if we do not further investigate this result in order to understand the groups that we manage to extract,
- on average, the social signature of egos are correlated with the social signature of the neighborhood showing signs of homophily. People who are connected through the communication network have closer social signatures than people who are not connected,

- During the year that students are entering to the university, the number of relations is increasing slightly (from 15 to 18) suggesting that this special year gives social benefits. The most of new relations are created just after the entrance of the first and the second semester revealing that the evolution is not uniform. In this chapter, we established a first causal link between a major event of ego life and evolution of the ego network.

There are many ways to extend this study. We intend to characterize typical social signatures like social keepers and social explorers defined in [103]. As we have pointed out the correlation between the network position and the social signature, an experiment can be done to decide whether connected people become more similar (effect of social influence) or similar people tend to connect each other (homophily). In parallel, a deeper study of deactivated ties may reveal interesting facts on turnovers. If we include the social classes in this study, we can make the link between the way that an ego network is evolving for student and its effort on the success after the university. Understanding why some students manage to get a greater income than the others is crucial to understand inequalities and social reproduction.

# Part II : Conclusion & Future Work

In this part, we presented a data-driven approach and made a sociological study by looking at combined communication and bank data sets at Mexico. This data-driven approach has many advantages. As we proposed, we can automatically define social classes of millions of users both having socioeconomic status and demographic attributes available. As an evidence, it is impossible to have access to such a large sample by only personal questionnaires. To our knowledge, it is the first work that proposes a sociological study at a country population from the analysis of the combination of social structure and socioeconomic status of millions of users. First, we showed how data-driven approaches give the possibility to put in evidence higher-order correlations, impossible to catch with a conventional approach. Second, we have validated many long lasting sociological and economical hypotheses:

- we pointed out inequalities on the distribution of wealth and debts,
- we validate and quantify the effects of social stratification in the society,
- we showed spatial segregation concerning both commuting distance from home to work and home distances between social classes,
- we made a transversal analysis of human purchasing behavior to understand better its relation to social structure,
- we finally provided some initial results on the role of the university studies in the social reproduction.

Even if our studies are built on large and detailed data, the utilized data cover the population of Mexico only partially. However, as we demonstrated above, for population-level measures, such as the Gini coefficient and spatial distribution, we obtained values close to independently reported cases, and thus our observations may generalize in this sense. In addition, the question remains how well mobile phone call networks approximate the real social structure. A recent study [54] demonstrated that real social ties can be effectively mapped from mobile call interactions with precision up to $95\%$. However, it is important to keep in mind that the poorest social class of the society is probably under-represented in the data as they may have no access to bank services and/or do not hold mobile phones. Datasets simultaneously disclosing the social structure and the socioeconomic indicators of a large number of individuals are still very rare.

However, several promising directions have been proposed lately to estimate socioeconomic status from communication behavior on regional level [16, 96, 138] or even for individuals [15], just to mention a few.

In future works these methods could be used to generalize our results to other countries using mobile communication datasets. Here, our aim was to report some general observations in this direction using directly estimated individual economic indicators. Our overall motivation was to empirically verify some long-standing hypotheses and to explore a common ground between hypothesis-driven and data-driven research addressing social phenomena.

The main message of these studies is about to consider data-driven approaches as a great opportunity to complement classical sociological approaches. There are many advantages that reduce some biases such as the bias induced by the choice of the question in an interview and the scale of the sample is very large compare to interview processes. Yet, as it concerns social sciences, computer science thus highlights the importance of collaborations between these disciplines to build relevant studies. The difference between only processing data and undertake a full sociological analysis is big. Moreover, the limitation of the parameters with a data-driven approach is the main limitation. Therefore getting an access to rich and diverse data open avenues to addressing novel problems. In addition, as the time period of the bank and communication data is quite long, it can be very interesting to make temporal economical analysis. By looking at personal events such as relocations or working events combining with the evolution of personal parameters like the ties and spendings and the external economical or political events (crisis, conflicts), we can answer a lot of questions. The time aspect can be pushed in order to answer to novel questions like the understanding of the temporal convergence and equilibrium between the expenditure and the income or the temporal social stratification structure when one changes his or her social status.

# Big data approach : the pros and the cons

In this thesis, from the data sets, I tried to push the limit of the data-driven approach in order to understand their capacities. Always limited in space, in time and in number of features, there are still great advantages of big data compared to classical survey studies :

- Completeness: dealing with a high percentage of the global system improves the quality and precision of the study, of the predictions and the recommendation processes that emerge from the analysis. Furthermore, it let me consider a large connected representative part of the global network instead of analyzing a sample of it.
- Validity: the registration of the activity of users are coming from real actions. In this case, the user cannot lie, attenuates or exaggerates on the answer. There are no biases coming from the answer of the user as during a typical sociological interview.
- Novel information: automatically, many data are registered and stored. In many cases, the ego does not even have in mind parameters that are stored. How many times did you call your family last year? Do you call more your rich friends rather than your poor friends? Are your mobility and mobile events related? Big data not only insure the validity but also brings new information on users.
- Efforts: when tools are already known and seted, the effort in order to collect the data is quite small compare to the interviewing process. It is impossible to compare one interview to one client in the data but at the end, dealing with data helps research to gain time. For example, for transportation, it is easy to register origin and destination locations of people automatically. With a sociological approach, one may ask randomly in the subway where people are going.

Yet, these pros are not free. By applying data-driven approaches with a modern dataset, I tried to understand the challenges and criteria of a good study. Four main challenges point out how data-driven approaches are still complementary to social science:

- Interdisciplinarity: combining the techniques of computer scientists and the expertise of social scientists is a veritable challenge as different fields have to be confident and speak in a same language. However, interdisciplinarity is a helpful asset in every field.

- In-depth details : as we noticed in the studies of this thesis, even if we have access to a large number of features, it seems impossible to ask in-depth questions to an ego as we only have access to his or her actions through logs. In a close future, if the number of combined features increases, it may offer the possibility of having access to in-depth details without interviewing such as opinions and feelings.

- Gap to reality: data are mainly collected by companies or research projects with a limited context. Most of the time, the considered sample is a biased sample of the population. In this PhD, as we consider people that have a bank account and a mobile account, we introduced unquantifiable biases that may create a small gap between the results and reality. In some studies, the methodology also increases the gap between the studied object and the real system.

- Access and privacy : having a privileged access to a good data set is demanding. For some reasons, this access is mainly restricted to private companies. In addition, the anonymization is partial in most of the case. We commonly called anonymized data sets when the ids are anonymized by hashed phone numbers and the individual identity is not given. However, the locations, age and gender may give the possibility of uniquely identify an ego as discussed in [37].

As we noticed, big data offers the possibility to get access to new features at an individual level according to time and space like communications, demographic attributes or spendings. Moreover, as it is recorded, the biases induced by the data-driven approaches are quite different and in this way can complement the classical interviewing approach. However, it is not easy to guess what will be the evolution of future data-driven studies. In one hand, it is becoming very easy to record actions and contents, the diversity of such data is increasing. Therefore, the refinement of the knowledge at the user level of personal traits and their relations may overcome the quality of the answer obtained during a traditional interviewing process. In the other hand, the user is becoming less naive regarding the data recorded on him or her. Each person, as a user, would take care about what kind of data he or she gives to the companies in order to protect his or her privacy. In addition, the access to the big data is very unequal as very few organizations have access to multimodal big data. In this context, I wonder rather if big data approaches will increase the knowledge and improve our living or if, on the contrary, they will increase inequalities by being a new opportunity for big organizations to confirm their supremacy on clients and their influence on personal and collective decisions.

# Final Conclusion

In this thesis I have carried out data-driven studies based on rich, large-scale combined data sets. In addition, I have presented diverse findings in the fields of computer networks and computational social science. In computer networks, my results may have an impact on the study of dynamic graphs, but will more likely influence applied domains such as telecommunication networks and mobile data mining. Having said this, the second part is even more interdisciplinary, as I have presented a longitudinal social study and have given results relevant to the fields of sociology, economics and geography.

The first benefit of this three-year-long work is the interdisciplinary approach. Working with researchers from various fields has been the key to tackling interesting questions with original approaches. The variety of approaches drawn from various disciplines has provided the possibility to improve comprehension of systems at the level of human dynamics. As I have presented, a single rich dataset can help us to develop tools that improve structured data-driven methodology, to experimentally evaluate the connectedness of mobile users and to answer to long-standing sociological questions such as those regarding social stratification.

In this thesis, I have pointed out the main properties of a dataset that increase its richness. First, the number and the quality of personal attributes such as gender and age are crucial for a sociological study, because for many questions, results will depend heavily on such dimensions. Second, the evolution in time and space of each individual is key to the study of urban areas and design models for smart cities. For example, the spatial daily rhythm has an impact on transportation and communication services as we noticed in Chapter 4. Third, spontaneous individual choices and actions such as purchases make the link between individuals and short-term activity (such as shopping or practice of sport) and long-term attributes (social status or preferences). Fourth, the network position is fundamental. The social structure stemming from social interactions is key to understanding human behavior. I pointed out many correlations at the network level which reveal that personality traits depend greatly on network position. Consequently, as the homophily of ties is very strong, the personality traits of an ego can be inferred by the analysis of his or her network position. Finally, in practice, the context of the data conditions the validity of the findings. For example, having a large representative sample of users and making an analysis over a long period improve the analysis significantly.

In addition, applications to these findings are manifold. Correlations are essential in making predictions about clients and propose recommendations to users. Operators may adapt their protocols in order to anticipate peaks and congestion. By sharing their mobile data, they can understand the evolution of the usage

and gain a precise idea of the market and the infrastructure they have to improve. In order to create new protocols, they can run experiments over real traces and get precise estimates of their efficiency and cost. Besides, the understanding of personal attributes such as age, gender, favourite locations and social status is invaluable for companies that may recommend specific products and activities to clients according to personal traits, their location and their consumption power. As I showed several times in this thesis, the network position plays an important role in the way the ego lives and interacts with others. Combining correlations at the network level and personal traits for a sample of users should help obtain a good estimation of unknown personality traits. This aspect is crucial in designing a specific experiment for new clients, for whom we know his or her ego network, by extending the personality traits and preferences of his or her neighbors.

The eight chapters of this thesis are diversified examples of big data studies. As is fundamental to understanding social structure, these studies have advanced the knowledge of social aspects such as temporal and spatial activity, interactions and inequalities. I started this thesis by synthesizing different definitions of the term big data, I would like to end it by proposing my own definition:

> "Big data is a term describing the analysis of rich and multi-dimensional data sets using diverse, innovative tools and following a precise methodology that provides justification of each step. Consideration of every bias and explanation of the limits are indispensable in order to obtain meaningful quantified correlations that reveal faithful findings that improve our understanding of real systems."

# References


[1] Instituto nacional de estadśtica geografá (www.inegi.org.mx) - 2015 census (date of access: 22 february 2016).

[2] To assign purchase values in usd we used the daily average currency rate (17.90 mxn/usd) on the 2nd march 2016.

[3] Wikipedia - mexican states by population density.

[4] Social class and consumer behavior: the relevance of class and status. *NA - Advances in Consumer Research*, 14, 1987.

[5] A. Abraham, A.-E. Hassanien, V. Sná, et al. *Computational social network analysis: Trends, tools and research advances*. Springer Science & Business Media, 2009.

[6] T. Akbar-Williams. Black class structure. In *Encyclopedia of African American Popular Culture, edited by Jessie Carney Smith,*. Westport, CT: Greenwood Press,, 2009.

[7] H. Aoyama, W. Souma, Y. Nagahara, M. P. Okazaki, H. Takayasu, and M. Takayasu. Pareto's law for income of individuals and debt of bankrupt companies. *Fractals*, 8(03):293–300, 2000.

[8] F. Baccelli and P. Brémaud. *Elements of queueing theory : Palm-martingale calculus and stochastic recurrences*. Springer, Berlin; New York, 2nd ed. edition, c2003. (TIT) Palm-martingale calculus and stochastic recurrences.

[9] J. P. Bagrow and Y.-R. Lin. Mesoscopic structure and social aspects of human mobility. *PloS one*, 7(5):e37676, 2012.

[10] W. Bank. Gini index estimates. To assign purchase values in USD we used the daily average currency rate (17.90 MXN/USD) on the 2nd March 2016.

[11] S. Bender-deMoll and D. A. McFarland. The art and science of dynamic network visualization. *Journal of Social Structure*, 7, 2006.

[12] C. M. Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

[13] V. D. Blondel, A. Decuyper, and G. Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):1, 2015.

[14] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[15] J. Blumenstock, G. Cadamuro, and R. On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

[16] J. Blumenstock and N. Eagle. Mobile divides: gender, socioeconomic status, and mobile phone use in rwanda. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, page 6. ACM, 2010.

[17] W. Bottero. *Stratification: social division and inequality*. Routledge, 2004.

[18] P. Bourdieu. *Distinction: A social critique of the judgement of taste*. Harvard University Press, 1984.

[19] D. Braha and Y. Bar-Yam. From centrality to temporary fame: Dynamic centrality in complex networks. *Complexity*, 12(2):59–63, 2006.

[20] J. Brea, J. Burroni, M. Minnoni, and C. Sarraute. Harnessing mobile phone social network topology to infer users demographic attributes. In *Proceedings of the 8th Workshop on Social Network Mining and Analysis*, page 1. ACM, 2014.

[21] D. Brown. Social class and status. *Mey, Jacob. Coincise Encyclopedia of pragmatics*, 2009.

[22] S. Butt and J. G. Phillips. Personality and self reported mobile phone use. *Computers in Human Behavior*, 24(2):346–360, 2008.

[23] R. S. Caceres, T. Berger-Wolf, and R. Grossman. Temporal scale of processes in dynamic networks. In *IEEE 11th International Conference on Data Mining Workshops (ICDMW 2011)*, pages 925–932. IEEE, 2011.

[24] K. E. Campbell, P. V. Marsden, and J. S. Hurlbert. Social resources and socioeconomic status. *Social networks*, 8(1):97–117, 1986.

[25] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.

[26] D. Cardon. *A quoi rêvent les algorithmes: Nos vies à lheure des big data*. Seuil, 2015.

[27] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, and A. Vespignani. Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE*, 5:e11596, 2010.

[28] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*, 6(6):606–620, 2007.

[29] T. W. Chan. *Social status and cultural consumption*. Cambridge University Press, 2010.

[30] Y. Chi, S. Zhu, X. Song, J. Tatemura, and B. L. Tseng. Structural and temporal analysis of the blogosphere through community factorization. In *13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007)*, pages 163–172. ACM, 2007.

[31] G. Chittaranjan, J. Blom, and D. Gatica-Perez. Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, 17(3):433–450, 2013.

[32] K. Church and R. de Oliveira. What's up with whatsapp?: Comparing mobile instant messaging behaviors with traditional sms. In *Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services*, MobileHCI '13, pages 352–361, New York, NY, USA, 2013. ACM.

[33] A. Clauset and N. Eagle. Persistence and periodicity in a dynamic proximity network. In *DIMACS Workshop on Computational Methods for Dynamic Interaction Networks*, 2007.

[34] V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani. Detecting rich-club ordering in complex networks. *Nature physics*, 2(2):110–115, 2006.

[35] B. C. Csáji, A. Browet, V. A. Traag, J.-C. Delvenne, E. Huens, P. Van Dooren, Z. Smoreda, and V. D. Blondel. Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 392(6):1459–1473, 2013.

[36] H. N. Curto, J. Caetano, J. Almeida, A. Ziviani, C. H. S. Malab, and H. T. Marques-Neto. Using sms to transfer small data packets during periods of high workload on mobile data networks. In *XXXIII Simpsio Brasileiro de Redes de Computadores e Sistemas Distribudos*, 2015.

[37] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.

[38] Y.-A. de Montjoye, J. Quoidbach, F. Robic, and A. S. Pentland. Predicting personality using novel mobile phone-based metrics. In *Social computing, behavioral-cultural modeling and prediction*, pages 48–55. Springer, 2013.

[39] A. Deaton. *Understanding consumption*. Oxford University Press, 1992.

[40] A. Deaton. *The analysis of household surveys: a microeconometric approach to development policy*. World Bank Publications, 1997.

[41] A. Deaton and J. Muellbauer. *Economics and consumer behavior*. Cambridge university press, 1980.

[42] C. Déglise, L. S. Suggs, and P. Odermatt. Short message service (sms) applications for disease prevention in developing countries. *Journal of medical Internet research*, 14(1):e3, 2012.

[43] P. Desikan and J. Srivastava. Mining temporally changing web usage graphs. In *Advances in Web Mining and Web Usage Analysis: 6th International Workshop on Knowledge Discovery on the Web (WebKDD 2004)*, number 3932 in LNAI, pages 1–17. Springer, 2006.

[44] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014.

[45] C. Dickens. *A tale of two cities*. Vintage, 2012.

[46] A. Dobra, N. E. Williams, and N. Eagle. Spatiotemporal Detection of Unusual Human Population Behavior Using Mobile Phone Data. *PLOS ONE*, 10(3):e0120449+, Mar. 2015.

[47] Y. Dong, F. Pinelli, Y. Gkoufas, Z. Nabi, F. Calabrese, and N. V. Chawla. Inferring unusual crowd events from mobile phone call detail records. In *Machine Learning and Knowledge Discovery in Databases*, pages 474–492. Springer, 2015.

[48] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla. Inferring user demographics and social strategies in mobile social networks. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 15–24. ACM, 2014.

[49] C. B. Doob. *Social inequality and social stratification in US society*. Routledge, 2015.

[50] R. Dunbar. The social brain hypothesis. *brain*, 9(10):178–190, 1998.

[51] N. Eagle. *Machine perception and learning of complex social systems*. PhD thesis, Massachusetts Institute of Technology, Department of Media Arts and Sciences, 2005.

[52] N. Eagle, M. Macy, and R. Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.

[53] N. Eagle and A. Pentland. Reality mining: Sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, 2006.

[54] N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, 106(36):15274–15278, 2009.

[55] B. Efron and T. Hastie. *Computer Age Statistical Inference*, volume 5. Cambridge University Press, 2016.

[56] R. M. Emerson. Social exchange theory. *Annual review of sociology*, pages 335–362, 1976.

[57] M. Faulkner, M. Olson, R. Chandy, J. Krause, K. M. Chandy, and A. Krause. The next big one: Detecting earthquakes and other rare events from community-based sensors. In *Information Processing in Sensor Networks (IPSN), 2011 10th International Conference on*, pages 13–24. IEEE, 2011.

[58] R. Felix, P. A. Rauschnabel, and C. Hinsch. Elements of strategic social media marketing: A holistic framework. *Journal of Business Research*, 2016.

[59] M. Ficek and L. Kencl. Inter-call mobility model: A spatio-temporal refinement of call data records using a gaussian mixture model. In *INFOCOM, 2012 Proceedings IEEE*, pages 469–477. IEEE, 2012.

[60] B. Fish and R. S. Caceres. Handling oversampling in dynamic networks using link prediction. In *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2015)*, volume 9285, part II of *LNAI*, pages 671–686. Springer, 2015.

[61] B. S. Fjeldsoe, A. L. Marshall, and Y. D. Miller. Behavior change interventions delivered by mobile telephone short-message service. *American journal of preventive medicine*, 36(2):165–173, 2009.

[62] A. Furno, R. Stanica, and M. Fiore. A comparative evaluation of urban fabric detection techniques based on mobile traffic data. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 689–696, New York, NY, USA, 2015. ACM.

[63] J. L. Gastwirth. The estimation of the lorenz curve and gini index. *The Review of Economics and Statistics*, pages 306–316, 1972.

[64] A. Gautreau, A. Barrat, and M. Barthlemy. Microdynamics in stationary complex networks. *PNAS*, 106:8847–8852, 2009.

[65] A. Giddens, F. Ociepka, and W. Zujewicz. *The class structure of the advanced societies*. Hutchinson London, 1973.

[66] D. Gilbert. *The American class structure in an age of growing inequality*. Sage Publications, 2014.

[67] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[68] M. Grossglauser and D. Tse. Mobility increases the capacity of ad-hoc wireless networks. In *INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 3, pages 1360–1369. IEEE, 2001.

[69] D. B. Grusky. Theories of stratification and inequality. *The Blackwell Encyclopedia of Sociology. Oxford: Basil Blackwell*, pages 4809–4818, 2007.

[70] G. Heine and M. Horrer. *GSM networks: protocols, terminology, and implementation*. Artech House, Inc., 1999.

[71] R. G. Hollands. Will the real smart city please stand up? intelligent, progressive or entrepreneurial? *City*, 12(3):303–320, 2008.

[72] P. Holme. Network dynamics of ongoing social relationships. *Europhysics Letters*, 64(3):427, 2003.

[73] P. Holme, S. Min Park, B. Kim, and C. Edling. Korean university life in a network perspective: Dynamics of a large affiliation network. *Physica A: Statistical Mechanics and Its Applications*, 373:821–830, 2007.

[74] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket switched networks and human mobility in conference environments. In *Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, pages 244–251. ACM, 2005.

[75] C. E. Hurst. *Social inequality: Forms, causes, and consequences*. Routledge, 2015.

[76] R. Kitchin. The real-time city? big data and smart urbanism. *GeoJournal*, 79(1):1–14, 2014.

[77] B. Klimt and Y. Yang. The enron corpus: A new dataset for email classification research. In *15th European Conference on Machine Learning (ECML 2004)*, pages 217–226. Springer, 2004.

[78] L. Kovanen, K. Kaski, J. Kertész, and J. Saramäki. Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. *Proceedings of the National Academy of Sciences*, 110(45):18070–18075, 2013.

[79] G. Krings, M. Karsai, S. Bernharsson, V. D. Blondel, and J. Saramäki. Effects of time window size and placement on the structure of aggregated networks. *EPJ Data Science*, 1(4):1–16, 2012.

[80] M. Lahiri, A. S. Maiya, R. Sulo, Habiba, and T. Y. Berger-Wolf. The impact of structural changes on predictions of diffusion in networks. In *IEEE International Conference on Data Mining Workshops (ICDMW 2008)*, pages 939–948, 2008.

[81] T. L. Lai. Service quality and perceived value's impact on satisfaction, intention and usage of short message service (sms). *Information Systems Frontiers*, 6(4):353–368, 2004.

[82] R. Lambiotte, V. D. Blondel, C. De Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008.

[83] D. Laney. 3d data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6:70, 2001.

[84] M. Latapy, A. Hamzaoui, and C. Magnien. Detecting events in the dynamics of ego-centered measurements of the internet topology. *Journal of Complex Networks*, 2013.

[85] M. Latapy and C. Magnien. Complex network measurements: Estimating the relevance of observed properties. In *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*. IEEE, 2008.

[86] M. Latapy and C. Magnien. Complex network measurements: Estimating the relevance of observed properties. In *27th IEEE Conference on Computer Communications (INFOCOM 2008)*, pages 1660–1668. IEEE, 2008.

[87] P. F. Lazarsfeld, R. K. Merton, et al. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, 18(1):18–66, 1954.

[88] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[89] J.-Y. Le Boudec. Understanding the simulation of mobility models with palm calculus. *Perform. Eval.*, 64(2):126–147, Feb. 2007.

[90] Y. Leo, C. Sarraute, A. Busson, and E. Fleury. Taking benefit from the user density in large cities for delivering sms. In *Proceedings of the 12th ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, & Ubiquitous Networks*, PE-WASUN '15, pages 55–61, New York, NY, USA, 2015. ACM.

[91] K. Lerman, R. Ghosh, and J. H. Kang. Centrality metric for dynamic networks. In *8th Workshop on Mining and Learning with Graphs (MLG 2010)*, pages 70–77. ACM, 2010.

[92] S. F. LeRoy and J. Sonstelie. Paradise lost and regained: Transportation innovation, income, and residential location. *Journal of Urban Economics*, 13(1):67–89, 1983.

[93] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[94] M.-X. Li, Z.-Q. Jiang, W.-J. Xie, S. Miccichè, M. Tumminello, W.-X. Zhou, and R. N. Mantegna. A comparative analysis of the statistical properties of large mobile phone calling networks. *arXiv preprint arXiv:1402.6573*, 2014.

[95] S. Lohr. The age of big data. *New York Times*, 11, 2012.

[96] H. Mao, X. Shuai, Y.-Y. Ahn, and J. Bollen. Quantifying socio-economic indicators in developing countries from mobile phone communication data: applications to côte divoire. *EPJ Data Science*, 4(1):1–16, 2015.

[97] J. B. M. M. M. T. C. S. Martin Fixman, Ariel Berenstein. A bayesian approach to income inference in a communication network. *ASONAM*.

[98] L. Martinet, C. Crespelle, and E. Fleury. Dynamic contact network analysis in hospital wards. In *5th Workshop on Complex Networks (CompleNet 2014)*, number 549 in Studies in Computational Intelligence, pages 241–249. Springer, 2014.

[99] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[100] R. Michalski, S. Palus, and P. Kazienko. Matching organizational structure and social network extracted from email communication. In *14th International Conference on Business Information Systems (BIS 2011)*, volume 87 of *LNBIP*, pages 197–206. Springer, 2011.

[101] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[102] C. Mills. Wright: The power elite. *New York*, 1956.

[103] G. Miritello, R. Lara, M. Cebrian, and E. Moro. Limited communication capacity unveils strategies for human interaction. *Scientific reports*, 3, 2013.

[104] G. Miritello, E. Moro, R. Lara, R. Martínez-López, J. Belchamber, S. G. Roberts, and R. I. Dunbar. Time as a limited resource: Communication strategy in mobile phone networks. *Social Networks*, 35(1):89–95, 2013.

[105] T. Mizuno, M. Katori, H. Takayasu, and M. Takayasu. Empirical science of financial fluctuations-the advent of econophysics, 2002.

[106] J. Moody. The importance of relationship timing for diffusion. *Social Forces*, 81(1):25–56, 2002.

[107] J. Moody. Static representations of dynamic networks. Technical report, Duke Population Research Institute, Duke University, Durham, 2008.

[108] J. Moody, D. McFarland, and S. Bender-deMoll. Dynamic network visualizations. *American Journal of Sociology*, 110(4):1206–1241, 2005.

[109] D. Naboulsi, R. Stanica, and M. Fiore. Classifying call profiles in large-scale mobile traffic datasets. In *INFOCOM, 2014 Proceedings IEEE*, pages 1806–1814. IEEE, 2014.

[110] V. Neiger, C. Crespelle, and E. Fleury. On the structure of changes in dynamic contact networks. In *Workshop on Complex Networks and their Applications (Complex Networks 2012). In 8th International Conference on Signal Image Technology and Internet Based Systems (SITIS 2012)*, pages 731–738. IEEE, 2012.

[111] M. Newman. *Networks: an introduction*. Oxford university press, 2010.

[112] M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[113] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: universal patterns in human urban mobility. *PloS one*, 7(5):e37027, 2012.

[114] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, and C. Sarraute. Measurement-driven mobile data traffic modeling in a large metropolitan area. In *Pervasive Computing and Communications (PerCom), 2015 IEEE International Conference on*, pages 230–235. IEEE, 2015.

[115] E. Oliver. Characterizing the transport behaviour of the short message service. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, MobiSys '10, pages 223–238, New York, NY, USA, 2010. ACM.

[116] R. N. Onody and P. A. de Castro. Complex network study of Brazilian soccer players. *Phys. Rev. E*, 70:037103, 2004.

[117] P. Panzarasa, T. Opsahl, and K. M. Carley. Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *J. Am. Soc. Inf. Sci. Technol.*, 60(5):911–932, 2009.

[118] P. Paraskevopoulos, T.-C. Dinh, Z. Dashdorj, T. Palpanas, and L. Serafini. Identification and characterization of human behavior patterns from mobile phone data. *Proc. of NetMob*, 2013.

[119] V. Pareto. Manual of political economy. 1971.

[120] U. Paul, A. Subramanian, M. Buddhikot, and S. Das. Understanding traffic dynamics in cellular data networks. In *INFOCOM, 2011 Proceedings IEEE*, pages 882–890, April 2011.

[121] L. Peel and A. Clauset. Detecting change points in the large-scale structure of evolving networks. In *29th AAAI Conference on Artificial Intelligence (AAAI 2015)*, pages 2914–2920. AAAI Press, 2015.

[122] N. Perra, B. Goncalves, R. Pastor-Satorras, and A. Vespignani. Activity driven modeling of time varying networks. *Scientific Reports*, 2, 2012.

[123] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *International Workshop on Human Behavior Understanding*, pages 14–25. Springer, 2010.

[124] T. Piketty, A. Goldhammer, and L. Ganser. Capital in the twenty-first century. 2014.

[125] L. Ponomarenko, C. S. Kim, and A. Melikov. *Performance Analysis and Optimization of Multi-Traffic on Communication Networks*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition, 2010.

[126] D. L. Poston. Socioeconomic status and work-residence separation in metropolitan america. *The Pacific Sociological Review*, 15(3):367–380, 1972.

[127] L. L. Putnam and A. M. Nicotera. *Building theories of organization: The constitutive role of communication*. Routledge, 2009.

[128] F. Radicchi, S. Fortunato, B. Markines, and A. Vespignani. Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, 80:056103+, 2009.

[129] B. Ribeiro, N. Perra, and A. Baronchelli. Quantifying the effect of temporal resolution on time-varying networks. *Scientific Reports*, 3, 2013.

[130] G. Rilling, P. Flandrin, P. Goncalves, et al. On empirical mode decomposition and its algorithms. In *IEEE-EURASIP workshop on nonlinear signal and image processing*, volume 3, pages 8–11. IEEER, 2003.

[131] S. G. Roberts and R. I. Dunbar. The costs of family and friends: an 18-month longitudinal study of relationship maintenance and decay. *Evolution and Human Behavior*, 32(3):186–197, 2011.

[132] S. S. Rosenthal and S. L. Ross. Change and persistence in the economic status of neighborhoods and cities. *Forthcoming in The Handbook of Regional and Urban Economics*, 5, 2014.

[133] S. Saganowski, P. Brodka, and P. Kazienko. Influence of the dynamic social network timeframe type and size on the group evolution discovery. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 679–683. IEEE, 2012.

[134] J. Saramäki, E. A. Leicht, E. López, S. G. Roberts, F. Reed-Tsochas, and R. I. Dunbar. Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences*, 111(3):942–947, 2014.

[135] P. Sarkar and A. W. Moore. Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter*, 7(2):31–40, 2005.

[136] C. Sarraute, P. Blanc, and J. Burroni. A study of age and gender seen through mobile phone usage patterns in mexico. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 836–843. IEEE, 2014.

[137] P. Saunders. *Social class and stratification*. Routledge, 1990.

[138] S. Šćepanović, I. Mishkovski, P. Hui, J. K. Nurminen, and A. Ylä-Jääski. Mobile phone call data as a regional socio-economic proxy indicator. *PloS one*, 10(4):e0124160, 2015.

[139] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.

[140] S. Sernau. *Social inequality in a global age*. SAGE Publications, 2013.

[141] C. Song, T. Koren, P. Wang, and A.-L. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010.

[142] S. Soundarajan, A. Tamersoy, E. B. Khalil, T. Eliassi-Rad, D. H. Chau, B. Gallagher, and K. Roundy. Generating graph snapshots from streaming edge data. In *WWW'16 Companion*. ACM, 2016.

[143] R. Stark. *Sociology*. Thompson Wadsworth, 2007.

[144] 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.

[145] J. E. Stiglitz. *The price of inequality: How today's divided society endangers our future*. WW Norton & Company, 2012.

[146] D. Stoyan, W. S. Kendall, and J. Mecke. *Stochastic geometry and its applications*. Wiley series in probability and mathematical statisitics. Wiley, Chichester, W. Sussex, New York, 1987. Rev. translation of: Stochastische Geometrie.

[147] R. Sulo, T. Berger-Wolf, and R. Grossman. Meaningful selection of temporal resolution for dynamic networks. In *8th Workshop on Mining and Learning with Graphs (MLG 2010)*, pages 127–136. ACM, 2010.

[148] J. Sun, S. Papadimitriou, P. S. Yu, and C. Faloutsos. Graphscope: Parameter-free mining of large time-evolving graphs. In *13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007)*, pages 687–696. ACM, 2007.

[149] J. Sun, D. Tao, and C. Faloutsos. Beyond streams and graphs: Dynamic tensor analysis. In *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, pages 374–383. ACM, 2006.

[150] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[151] C. K. Toh. *Ad hoc mobile wireless networks: protocols and systems*. Pearson Education, 2001.

[152] V. A. Traag, A. Browet, F. Calabrese, and F. Morlot. Social event detection in massive mobile phone data using probabilistic location inference. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 625–628. IEEE, 2011.

[153] P. Vanhems, A. Barrat, C. Cattuto, J.-F. Pinton, N. Khanafer, C. Rgis, B.-a. Kim, B. Comte, and N. Voirin. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS ONE*, 8(9):e73970, 2013.

[154] J. Viard and M. Latapy. Identifying roles in an ip network with temporal and structural density. In *6th IEEE International Workshop on Network Science for Communication Networks (NetSciCom 2014). In INFOCOM IEEE Conference on Computer Communications Workshops*, pages 801–806. IEEE, 2014.

[155] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in facebook. In *2nd ACM SIGCOMM Workshop on Online Social Networks (WOSN 2009)*, pages 37–42. ACM, 2009.

[156] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[157] P. West and I. for the Study of Civil Society. *Conspicuous compassion: why sometimes it really is cruel to be kind*. Civitas, 2004.

[158] J. O. Wheeler. Occupational status and work-trips: A minimum distance approach. *Social Forces*, 45(4):508–515, 1967.

[159] J. O. Wheeler. Some effects of occupational status on work trips. *Journal of Regional Science*, 9(1):69–77, 1969.

[160] J. Wiese, J.-K. Min, J. I. Hong, and J. Zimmerman. You never call, you never write: Call and sms logs do not always indicate tie strength. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 765–774. ACM, 2015.

[161] A. D. Wilson, S. Krause, N. J. Dingemanse, and J. Krause. Network position: a key component in the characterization of social personality types. *Behavioral Ecology and Sociobiology*, 67(1):163–173, 2013.

[162] W. Wood and T. Hayes. Social influence on consumer decisions: Motives, modes, and consequences. *Journal of Consumer Psychology*, 22(3):324–328, 2012.

[163] G. Yavaş, D. Katsaros, Ö. Ulusoy, and Y. Manolopoulos. A data mining approach for location prediction in mobile environments. *Data & Knowledge Engineering*, 54(2):121–146, 2005.

[164] W. C. Young, J. E. Blumenstock, E. B. Fox, and T. H. McCormick. Detecting and classifying anomalous behavior in spatiotemporal network data. In *Proceedings of the*

*2014 KDD Workshop on Learning about Emergencies from Social Information (KDD-LESI 2014)*, pages 29–33, 2014.

[165] P. Zerfos, X. Meng, S. H. Wong, V. Samanta, and S. Lu. A study of the short message service of a nationwide cellular network. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pages 263–268. ACM, 2006.

[166] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Towards mobile intelligence: Learning from gps history data for collaborative recommendation. *Artificial Intelligence*, 184:17–37, 2012.

[167] S. Zhou and R. J. Mondragón. The rich-club phenomenon in the internet topology. *IEEE Communications Letters*, 8(3):180–182, 2004.

[168] M. Zonoozi, P. Dassanayake, and M. Faulkner. Mobility modelling and channel holding time distribution in cellular mobile communication systems. In *Global Telecommunications Conference, 1995. GLOBECOM '95., IEEE*, volume 1, pages 12–16 vol.1, Nov 1995.

# A

# Appendix

## A.1 Degree and wealth correlations

In the main text of Chapters 5 and 6 we studied a social network where each individual was assigned with a socioeconomic indicator defined as their average monthly purchase (AMP) (see Eq. 5.1 in the main text). We used these indicators to estimate the socioeconomic status of individuals and group them into 9 exclusive socioeconomic classes. By analysing the social network and the assigned socioeconomic classes we observed that individuals tend to connect to similar others from their own or neighbouring socioeconomic classes, while social ties with people from remote classes are less frequent. We argued that this observed stratification in the social structure (see in Fig.3 a-c in the main text) is due to the entangled effects of socioeconomic imbalances and status homophily, i.e. the tendency of people to connect to others with similar socioeconomic status.

However, one can argue that the observed stratified structure can be simply the consequence of simultaneously present degree-degree and degree-wealth correlations. More precisely, if the degree of an individual is highly correlated with its economic status (wealth) and the network is strongly assortative (i.e. people prefer to connect to other people with similar degrees) we may observe similar effects as in Fig.1 (in the main text). To close out this possibility we present here a correlation analysis and a null model study where we carefully define random reference models to remove the correlations in focus in a controlled way and check their effects on the quantitative observations.

### A.1.1 Degree-degree correlations

The simplest way to characterise degree-degree correlations in a network is by computing the Pearson correlation coefficient between two random variables identified as the degrees of

nodes connected in the network structure [111]. After calculating the Pearson correlation coefficient in the investigated social network we found that it is $r \approx -0.00813$ ($p < 0.001$, $SE = 7.13 \times 10^{-4}$), suggesting that the network shows no (or very weak disassortative) degree-degree correlations. However, since the Pearson correlation coefficient gives only an overall characteristic measure and assumes that correlations are linear, we further investigate degree-degree correlations with another metric, which is conventionally used to characterise degree-degree correlations. We measure the $k_{nn}(k)$ average nearest neighbour degree for each degree class $k$ in the network [111]. This function (shown in Fig.SA.1.1a) disclosed a somewhat more sophisticated picture about degree-degree correlations. First of all it shows that it is not a monotonous function but it assigns mixed effects of assortative and disassortative mixing. It shows positive (assortative) correlations up to $k \simeq 10$, which after it indicates negative (disassortative) correlations, and becomes flat for the largest degrees. Consequently our network does not show strong assortative correlations over its whole degree range, which suggest that degree correlations may not evidently play a deterministic role in the observed stratified structure even if they are correlated with wealth. Note that this type of complex functional scaling of $k_{nn}(k)$ commonly characterises non-mutualised directed networks as discussed in [94]. This is in line with our case where the network was not mutualised were kept in the structure, which were assumed to be indirected after the network construction for the convenience of our study.

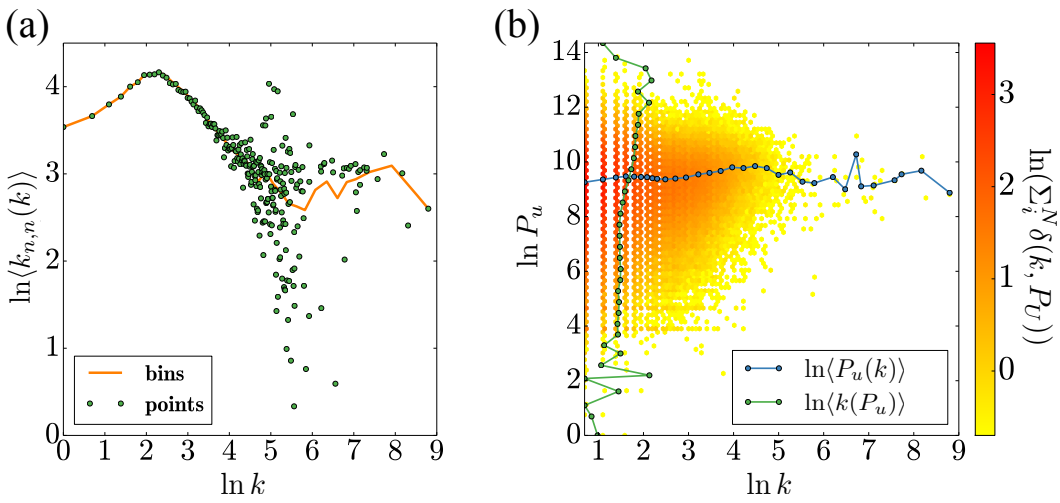



Figure A.1.1: Degree-degree and degree-wealth correlations in the social-economical network. (a) The $k_{nn}(k)$ function computed in the social network. (b) Correlation plot (shown as a heat map) of degree-wealth correlations. Blue (resp. green) horizontal (resp. vertical) solid line and symbols show the average wealth (resp. degree) as the function of degree (resp. wealth).

### A.1.2 Degree-wealth correlations

Entangled with degree correlations, dependencies between node degree and economic status (wealth) can also contribute to the emergence of the observed stratified structure. To characterise this correlation, first we measure again the Pearson correlation coefficient between the degree $k$ and AMP value $P_u$ of each node. This correlation turns out to be small with coeffi-

cient $r \approx 0.0357$ ($p < 0.001$, $SE = 9.71 \times 10^{-1}$). To obtain a more complete picture about their dependencies we simply show in Fig.A.1.1b the binned scatter plot as a heat map of these two variables and in addition calculate the average value of wealth (resp. degree) as the function of degree (resp. wealth). These results indicates the weak dependencies between these variables for the whole range of variables and although un-disclose some non-monotonous dependencies between node degree and economic status but do not indicate evident strong positive correlations between these variables.

### A.1.3 NULL MODEL STUDY

We concluded that the social structure is stratified by socioeconomic status by measuring the fraction $L(s_i, s_j) = |E(s_i, s_j)|/|E_{rn}(s_i, s_j)|$ (see also in Eq. 6.1 in the main text) of the number of links connecting people from different classes $s_i$ and $s_j$ in the original structure and in a null model structure. In this case the null model structure is defined as the configuration network model of the empirical network (here we call null model 1 (NM1)), where we take the original social network, select random pairs of links and swap them without allowing multiple links and self loops. In order to remove any residual correlations we repeated this procedure $5 \times |E|$ times (where $|E|$ being the number of links in the social network). This randomisation keeps the number of links, individual economic indicators $P_u$, and the assigned class of people unchanged, but destroys degree-degree correlations, possibly present community structure (for a summary of present correlations see Table SA.1.1). Since it destroys all possible structural correlations in the social network (apart from correlations due to unavoidable finite size effects) as a consequence it eliminates the socioeconomic layers as well. In each case, we repeat this procedure for 100 times and present results averaged over the independent random realisations. Results shown in Fig.SA.1.2b appears with a diagonal component, which evidently assigns strong connectivity between neighbouring socioeconomic classes, i.e. a stratified structure. This measure simply assigns *how many times people of different classes are more connected as compared to the case where they are connected by chance*. The observed diagonal component indeed assigns the significance of this correlations what we associated to status homophily.

To further investigate potential reasons behind this observation, let's take a null model to directly address the possible effects of degree homophily. In this null model (NM2) we destroy the potential community structure and all other structural correlations including socioeconomic layers, but conserve degree-degree and degree-wealth correlations (see Table SA.1.1). NM2 is defined as a modification of the configuration network model, where instead of selecting link pairs randomly to swap, we select a link and one of its end randomly, and choose another link randomly where the degree of one of the ending nodes is equal to the degree of the selected end of the first link. Swapping the other ends of the links (with potentially different degrees) will result yet two links between nodes of the original degrees but connected randomly otherwise. We do not allow self loops and multiple links and skip to swap links where both node degrees are unique in the network. We found 22 such cases from $\sim 2M$ links thus we assume this condition will not bias our shuffling considerably. We swapped randomly link pairs $5 \times |E|$ just as for the NM1 model and computed averages over 100 realisations.

Our hypothesis is that if the present degree-degree correlations in NM2 would explain the observed stratified structure, then after using the corresponding $|E_{rn}^{NM2}(s_i, s_j)|$ link density matrix

| | Original | NM1 | NM2 |
|---|---|---|---|
| degree-degree | ✓ | × | ✓ |
| degree-wealth | ✓ | ✓ | ✓ |
| communities | ✓ | × | × |

Table A.1.1: Correlations present in different null models. We consider degree-degree, degree-wealth, and higher-order structural (communities) correlations in the original network (Orig) and two null models (for definitions see text).

in the normalisation of $L(s_i, s_j)$ (see Eq. 6.1 in the main text) the resulting matrix should become flat. This would mean that the actually present correlations could explain (reproduce) the empirical observations. However, this is not the case here as seen in Fig.SA.1.2b. The $L(s_i, s_j)$ matrix normalised by the corresponding NM2 matrix appears to be almost identical than the one normalised by the NM1 matrix. This suggests that degree-degree correlations and degree homophily do not play a role here, thus it cannot explain the emergence of the stratified structure.

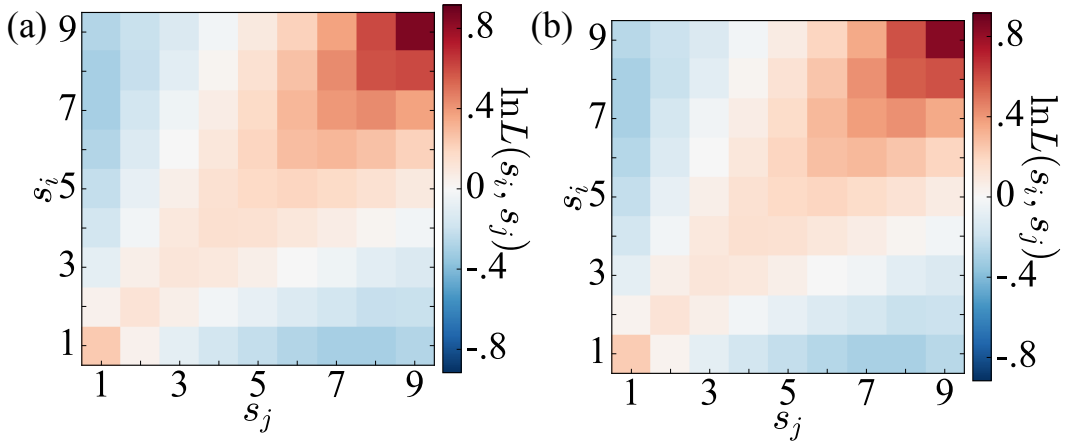



Figure A.1.2: $L(s_i, s_j)$ normalised socioeconomic class connectivity matrices in case of two null models (for definition see Eq. 6.1 in the main text). In each case the numerator was taken as the socioeconomic class connectivity matrices of the original network, while the denumerator was measured from a null model structure of (a) NM1, (b) NM2.

## A.2 MERCHANT CATEGORY CODES AND NAMES

| | | | | |
|---|---|---|---|---|
| 742: Veterinary Serv. | 5139: Commercial Footwear | 5719: Home Furnishing St. | 6211: Security Brokers | 7699: Repair Sh. |
| 763: Agricultural Cooperative | 5169: Chemicals Products | 5722: House St. | 6300: Insurance | 7829: Picture/Video Production |
| 780: Landscaping Serv. | 5172: Petroleum Products | 5732: Elec. St. | 7011: Hotels | 7832: Cinema |
| 1520: General Contr. | 5192: Newspapers | 5733: Music Intruments St. | 7012: Timeshares | 7841: Video Tape Rental St. |
| 1711: Heating, Plumbing | 5193: Nursery & Flowers Supp. | 5734: Comp.Soft. St. | 7032: Sporting Camps | 7911: Dance Hall & Studios |
| 1731: Electrical Contr. | 5198: Paints | 5735: Record St. | 7033: Trailer Parks, Camps | 7922: Theater Ticket |
| 1740: Masonry & Stonework | 5199: Nondurable Goods | 5811: Caterers | 7210: Laundry, Cleaning Serv. | 7929: Bands, Orchestras |
| 1750: Carpentry Contr. | 5200: Home Supp. St. | 5812: Restaurants | 7211: Laundries | 7932: Billiard/Pool |
| 1761: Sheet Metal | 5211: Materials St. | 5813: Drinking Pl. | 7216: Dry Cleaners | 7933: Bowling |
| 1771: Concrete Work Contr. | 5231: Glass & Paint St. | 5814: Fast Foods | 7217: Upholstery Cleaning | 7941: Sports Clubs |
| 1799: Special Trade Contr. | 5251: Hardware St. | 5912: Drug St. | 7221: Photographic Studios | 7991: Tourist Attractions |
| 2741: Publishing and Printing | 5261: Nurseries & Garden St. | 5921: Alcohol St. | 7230: Beauty Sh. | 7992: Golf Courses |
| 2791: Typesetting Serv. | 5271: Mobile Home Dealers | 5931: Secondhand St. | 7251: Shoe Repair/Hat Cleaning | 7993: Video Game Supp. |
| 2842: Specialty Cleaning | 5300: Wholesale | 5932: Antique Sh. | 7261: Funeral Serv. | 7994: Video Game Arcades |
| 4011: Railroads | 5309: Duty Free St. | 5933: Pawn Shops | 7273: Dating/Escort Serv. | 7995: Gambling |
| 4111: Ferries | 5310: Discount St. | 5935: Wrecking Yards | 7276: Tax Preparation Serv. | 7996: Amusement Parks |
| 4112: Passenger Railways | 5311: Dep. St. | 5937: Antique Reproductions | 7277: Counseling Serv. | 7997: Country Clubs |
| 4119: Ambulance Serv. | 5331: Variety St. | 5940: Bicycle Sh. | 7278: Buying/Shopping Serv. | 7998: Aquariums |
| 4121: Taxicabs | 5399: General Merch. | 5941: Sporting St. | 7296: Clothing Rental | 7999: Recreation Serv. |
| 4131: Bus Lines | 5411: Supermarkets | 5942: Book St. | 7297: Massage Parlors | 8011: Doctors |
| 4214: Motor Freight Carriers | 5422: Meat Prov. | 5943: Stationery St. | 7298: Health and Beauty Spas | 8021: Dentists, Orthodontists |
| 4215: Courier Serv. | 5441: Candy St. | 5944: Jewelry St. | 7299: General Serv. | 8031: Osteopaths |
| 4225: Public Storage | 5451: Dairy Products St. | 5945: Toy,-Game Sh. | 7311: Advertising Serv. | 8041: Chiropractors |
| 4411: Cruise Lines | 5462: Bakeries | 5946: Camera and Photo St. | 7321: Credit Reporting Agencies | 8042: Optometrists |
| 4457: Boat Rentals and Leases | 5499: Food St. | 5947: Gift Sh. | 7333: Graphic Design | 8043: Opticians |
| 4468: Marinas Serv. and Supp. | 5511: Cars Sales | 5948: Luggage & Leather St. | 7338: Quick Copy | 8049: Chiropodists, Podiatrists |
| 4511: Airlines | 5521: Car Repairs Sales | 5949: Fabric St. | 7339: Secretarial Support Serv. | 8050: Nursing/Personal Care |
| 4582: Airports, Flying Fields | 5531: Auto and Home Supp. St. | 5950: Glassware, Crystal St. | 7342: Exterminating Serv. | 8062: Hospitals |
| 4722: Travel Agencies | 5532: Auto St. | 5960: Dir Mark - Insurance | 7349: Cleaning and Maintenance | 8071: Medical Labs |
| 4784: Tolls/Bridge Fees | 5533: Auto Access. | 5962: Direct Marketing - Travel | 7361: Employment Agencies | 8099: Medical Serv. |
| 4789: Transportation Serv. | 5541: Gas Stations | 5963: Door-To-Door Sales | 7372: Computer Programming | 8111: Legal Serv., Attorneys |
| 4812: Phone St. | 5542: Automated Fuel Dispensers | 5964: Dir. Mark. Catalog | 7375: Information Retrieval Serv. | 8211: Elem. Schools |
| 4814: Telecom. | 5551: Boat Dealers | 5965: Dir. Mark. Retail Merchant | 7379: Computer Repair | 8220: Colleges Univ. |
| 4816: Comp. Net. Serv. | 5561: Motorcycle Sh. | 5966: Dir Mark - TV | 7392: Consulting, Public Relations | 8241: Correspondence Schools |
| 4821: Telegraph Serv. | 5571: Motorcycle Sh. | 5967: Dir. Mark. | 7393: Detective Agencies | 8244: Business Schools |
| 4899: Techno St. | 5592: Motor Homes Dealers | 5968: Dir. Mark. Subscription | 7394: Equipment Rental | 8249: Training Schools |
| 4900: Utilities | 5598: Snowmobile Dealers | 5969: Dir. Mark. Other | 7395: Photo Developing | 8299: Educational Serv. |
| 5013: Motor Vehicle Supp. | 5599: Auto Dealers | 5970: Artists Supp. | 7399: Business Serv. | 8351: Child Care Serv. |
| 5021: Commercial Furniture | 5611: Men Cloth. St. | 5971: Art Dealers & Galleries | 7512: Car Rental Agencies | 8398: Donation |
| 5039: Constr. Materials | 5621: Wom Cloth. St. | 5972: Stamp and Coin St. | 7513: Truck/Trailer Rentals | 8641: Associations |
| 5044: Photographic Equip. | 5631: Womens Accessory Sh. | 5973: Religious St. | 7519: Mobile Home Rentals | 8651: Political Org. |
| 5045: Computer St. | 5641: Childrens Wear St. | 5975: Hearing Aids | 7523: Parking Lots, Garages | 8661: Religious Org. |
| 5046: Commercial Equipment | 5651: Family Cloth. St. | 5976: Orthopedic Goods | 7531: Auto Body Repair Sh. | 8675: Automobile Associations |
| 5047: Medical Equipment | 5655: Sports & Riding St. | 5977: Cosmetic St. | 7534: Tire Retreading & Repair | 8699: Membership Org. |
| 5051: Metal Service Centers | 5661: Shoe St. | 5978: Typewriter St. | 7535: Auto Paint Sh. | 8734: Testing Lab. |
| 5065: Electrical St. | 5681: Furriers Sh. | 5983: Fuel Dealers (Non Auto) | 7538: Auto Service Shops | 8911: Architectural Serv. |
| 5072: Hardware Supp. | 5691: Cloth. St. | 5992: Florists | 7542: Car Washes | 8931: Accounting Serv. |
| 5074: Plumbing, Heating Equip. | 5697: Tailors | 5993: Cigar St. | 7549: Towing Serv. | 8999: Professional Serv. |
| 5085: Industrial Supplies | 5698: Wig and Toupee St. | 5994: Newsstands | 7622: Electronics Repair Sh. | 9211: Courts of Law |
| 5094: Precious Objects/Stones | 5699: Apparel Accessory Sh. | 5995: Pet Sh. | 7623: Refrigeration Repair | 9222: Government Fees |
| 5099: Durable Goods | 5712: Furniture | 5996: Swimming Pools Sales | 7629: Small Appliance Repair | 9223: Bail and Bond Payments |
| 5111: Printing, Office Supp. | 5713: Floor Covering St. | 5997: Electric Razor St. | 7631: Watch/Jewelry Repair | 9311: Tax Payments |
| 5122: Drug Proprietaries | 5714: Window Covering St. | 5998: Tent and Awning Sh. | 7641: Furniture Repair | 9399: Government Serv. |
| 5131: Notions Goods | 5718: Fire Accessories St. | 5999: Specialty Retail | 7692: Welding Repair | 9402: Postal Serv. |
| 5137: Uniforms Clothing | | | | |

Table A.2.1: **Codes and names of** 271 **merchant categories used in our study.** MCCs were taken from the Merchant Category Codes and Groups Directory published by American Express [1]. Abbreviations correspond to: Serv. - Services, Contr. - Contractors, Supp. - Supplies, St. - Stores, Equip. - Equipment, Merch. - Merchandise, Prov. - Provisioners, Pl. - Places, Sh. - Shops, Mark. - Marketing, Univ. - Universities, Org. - Organizations, Lab. - Laboratories.

## A.3 CONSUMPTION CORRELATIONS IN THE SOCIOECONOMIC NETWORK BY PURCHASE CATEGORY GROUP

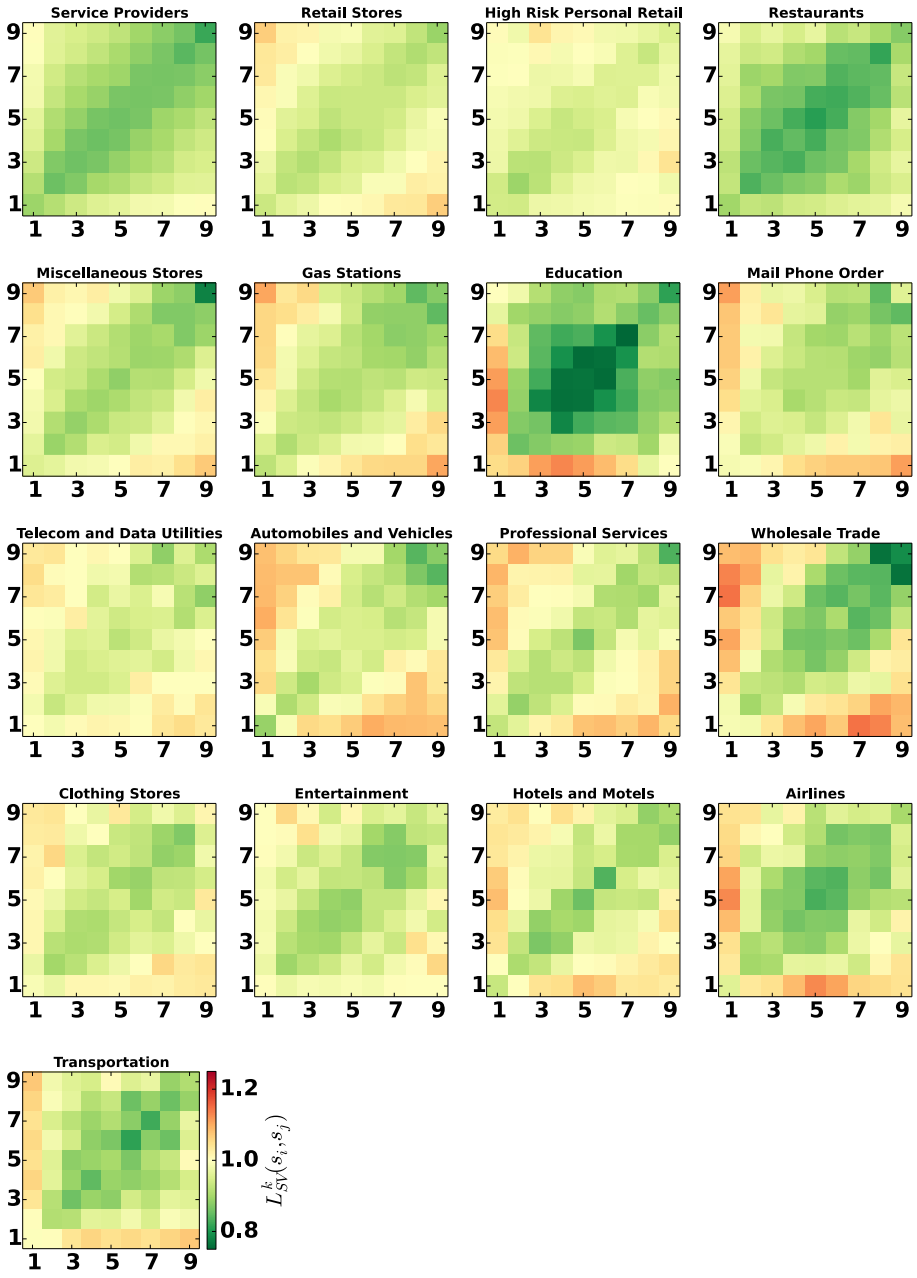



Figure A.3.1: **Consumption correlations in the socioeconomic network by PCG** Heat-map matrix representation of the average $L^k_{SV}(s_i, s_j)$ measure between pairs of socioeconomic classes for each PCG in $K_{17}$.