UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

Quantifying Urban Social Well-Being using Mobile Phone Data

Didem Gündoğdu

Advisor

Dr.Bruno Lepri

Fondazione Bruno Kessler

April 2018

"He remarks that, while the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician."

- Sherlock Holmes, The sign of the four. Sir Arthur Conan Doyle

Acknowledgements

In that thesis two different mobile phone datasets from Orange Telecom and Telecom Italia (TIM) have been used. Both of these datasets are compliant with the EU data security and privacy regulations. Here, I would like to thank Orange Telecom and Telecom Italia to make this work happen, by organizing a challenge to exploit new type of data, in the scale of country. United Nations Office for the Coordination of Humanitarian Affairs (UNOCHA) and The Humanitarian Data Exchange (HUMDATA) for accessing regional level data.

I would like to thank my supervisor Dr. Bruno Lepri for his guidance and support. My master supervisors Assoc.Prof. Albert Ali Salah and Assoc.Prof. Özlem Durmaz İncel for encouraging me to have a PhD, Prof.Dr. Nicu Sebe for choosing Trento and Prof.Dr. Lale Akarun, in the very beginning accepting me as a Master student to Boğaziçi University. Big hugs to Andrea Stenico and Francesca Belton, for surviving under my endless questions.

I am indebted to my many of my colleagues in MobS to support me, especially Simone (Cimone), Marco (il principesso), Gianni (CEO), Paolo (my best friend for today), Lorenzone, Lorenzino, Christos. I am sorry but I will put parmigiano on top of fish pasta.

My co-authors, Prof. Pietro Panzarasa and Dr. Nuria Oliver, I learned many things during that journey, thanks for your precious support and mentoring. I owe my deepest gratitude to the Intelligent Social System Lab at University College London. Particularly Prof. Mirco Musolesi for accepting to his group, and encourage me to be a part of his big family. I learned a lot from Benjamin Baron during my visit in UCL and I want to express my thankfulness for his support. The other family members Beatrice Perez, Paolo Bellagente, Abhinav Mehrotra, Alessio Pagani, Mariflor Vega, you make me feel like I have been your friend for a long time, which was so precious.

I owe everything to my family. My parents, Neşe and Gürol Gündoğdu, without your support I would probably give up in the beginning, and could not be where I am standing right now. My aunt Nihal Çanakcı (teyzoş), your being is so valuable to me, thanks for the chats we had which makes me feel not alone. My brother, Oktay Gündoğdu, you are always there when I need you. I feel so lucky to have you in my life, with your family, Ege and Esra. Ege, my precious, whenever I forgot my reason and ask for myself what I had been doing in Trento, I remember his question, “Aunt, you do not need to go to Italy to buy presents for me.”. So should I. :)

My Turkish gang in Trento: Begüm, Zafer, Serra, Orkan, Müge, Başak, Fatih, Emre, Elem, Gözde, Sinan, Ailbhe (you will soon officially become one of us), and Duygu. It was easier with you to embrace Trento, I can not imagine the scale of my suffer in your absence. I would like to particularly mention Orkan’s help for proofreading my thesis, it is great to know that at least one person has read it. :)

Big hugs to my overseas support: Binnur, Şule, Şakire, Zeynep, Emre, Eyüp, Sinan, Suat, Ziya and all not mentioned friends. I have to mention *Kahvaltı, yürüyüş, yakalamaç* gang Özge ve Ersin Özçıtak, Handan, İrem, Arzu, Nurdan, Deniz, which makes London my second home.

I also would like to thank my virtual friends Sheldon, Amy, Leonard, Penn, Rajes, Howard and all the people who take part in The Big Bang Theory. Sir Arthur Conan Doyle, for creating Sherlock Holmes.

And at last, I want to send my greatest appreciation to people, who are not with us, however touched my life so meaningful way and definitely lead me to write this thesis. To my grand mom and my wonderful cousin Tahir Çanakcı...

List of Publications

- Gundogdu, D., Incel, O. D., Salah, A. A., & Lepri, B. (2016). Countrywide arrhythmia: emergency event detection using mobile phone data. EPJ Data Science, 5(1), 25.

- Gundogdu, D., Finnerty, A. N., Staiano, J., Teso, S., Passerini, A., Pianesi, F., & Lepri, B. (2017). Investigating the association between social interactions and personality states dynamics. Open Science, 4(9), 170194.

- Gundogdu, D., Panzarasa, P., Oliver, N., & Lepri, B. (Under Review). Social capital and mobile phone data: Evidence from a developing country.

- Gundogdu, D., Baron, B., Lepri, B., & Musolesi, M. (In Process). Segregation or integration, which is better for the well-being of a city?

Abstract

Today, more than half of the world population is living in cities, which has been doubled in the last 50 years. The reason for that attraction is not only economical, but also security, education, and health. While people migrate to cities to reach improved life conditions, several issues raised by the increasing population. Recent studies have shown the importance of ethnic and cultural diversity of urban population to encourage tolerance, and to foster creativity and economic growth. Facing the urban growth challenges, we search for the key formulas to obtain healthy societies under the light of new type of data sources, such as mobile phone usage datasets. To this end, first we build up a tool to identify security related incidents from a country, which unstable political conditions held. Then we trace the formulas of healthy societies with examples from both developing and developed countries. We check the individual interaction and communication pattern effects (bridging and bonding) for the existence of social capital. Then we analyze aggregated ethnic diversity, and associate segregation scores with census data, and different ethnic groups preferences to move in the city, existence of any pattern for specific nation. The current studies are mainly hypothetical, with the absence of large scale real life data sources. This thesis aims to provide an insight to policy makers for building healthy societies, for the benefit of urban well-being.

Keywords
[social capital, complex networks, segregation, economical well-being, mobile phone datasets, call detail records]

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Motivations

Understanding human nature and behavior has a long history that lies back to the ancient times. Philosophers, social scientists, sociologists, psychologists, economists whose focus is to learn human behavior for ages, base their theories mainly on subjective observations, surveys performed through limited number of people, or in lab experiments. Today, almost everyone carries a mobile phone in his/her pocket, independent from his/her wealth. Yet with its basic capabilities, a mobile phone allows us to capture the human behavior in high precision in a very intimate way. To this end, the mobile phone usage is precious to understand human behaviour, even not including any kind of content of the conversations, but enabling us to model the social interaction with time and space.

The invention of mobile phones goes back to a couple of decades back, but in such a short period of time, the number of mobile phone users reached around 5 billion all around the world according to GSMA[1]. The power of this new technology can be better realized if we recall the total population of world is around 7 billion. What is surprising is the penetration of these new technology is so high even in the poorest countries. According to the World Bank report, the poorest household's most likely

[1]https://www.gsma.com/newsroom/press-release/number-of-global-mobile-subscribers-to-surpass-five-billion-this-year/

to have a mobile phone than access to clean water [55]. In developing countries, the reason for this phenomena is mostly the adequacy of the necessary infrastructure. In Africa, the mobile phone coverage, is wider than the general land lines, and it allows people to make business by mobile payment, since there is limited access to financial institutes in Sub-Saharan rural parts.

The power of mobile phone data lying beneath how vastly it has been generated. Each call, texting, connection to the Internet leaves a record in the Mobile Network Operators (MNOs), which is necessary for issuing an invoice for the provided service to connect people. This research, is about quantifying the value of aggregated digital breadcrumbs —mobile phone usage data— for the well-being of human beings.

In the recent years there has been an extensive research effort for analyzing mobile phone usage data to infer, model and/or predict human traits and behaviors. Thanks to some main MNOs' *Data for Development Challenges (D4D)*[2] initiatives on sharing subset of their aggregated and anonymized Call Detail Records (CDRs) to research communities in several challenges. In particular, Orange Telecom's challenges in 2013 and 2015 aims at exploiting sustainable development goals for developing countries in several categories (i) health, (ii) agriculture, (iii) transport/urban planning, (iv) energy and (v) national statistics. These challenges demonstrate success in various applications such as: energy planning, transportation system planning, detecting clean water dwells, poverty analysis, disease dissemination etc.[3]. The insights are attracting governments, policy makers, organizations like United Nations since it allows fast decision making, either by complimenting current solutions or as standalone to assess social

[2]http://www.d4d.orange.com/en/Accueil

[3]https://www.gsma.com/mobilefordevelopment/programme/digital-identity/discussing-data-for-development

problems almost in real time.

Due to privacy it is not possible to leverage CDR even for the purpose of improving living conditions, especially in developing countries. On the other hand, benefits of using mobile phone usage data are so promising that, the academic institutes are looking for solutions to overcome privacy issues. One of them is OPAL project[4], a platform to run algorithms on without accessing the meta data.

The power of aggregated data for the well-being of the societies has to be evaluated. A British statistician, Sir Francis Galton, mentioned about *wisdom of crowds* when he figured, how precise the average estimates of the weight of an ox, from a country fair in West England [76]. The group was composed of around 700 people from different backgrounds, few were experts, but most of them were common people. On the contrary to the expectations, group's estimate was only 1% from the real value. This phenomena has been later investigated in different domains, including politics and economy [194]. In that thesis the presence of cumulative intelligence through the mobile phone usage data is questioned for different aspects.

In that context, this thesis aim to shed light on several questions present in social science, with the tools of computer science and data obtained from real life. To this end, we have first evaluated the common sense of the society against unexpected threats, with some evidence from a developing country. The hourly aggregated number of calls given for each cell tower, allows us to learn the normal expected call volume, thus to predict the presence of anomaly in the region. During the data collection period, in addition to some several social events, such as African Football Cup, New Year, there were also some emergency events caused by the political instability and conflicts. Under these circumstances, by just looking at the

[4]http://www.opalproject.org

number of calls, we could identify with some precision the type, time, and location of the event. This is an expected behavior as society tends to share the new information/threat through their connections. When we carry this hypothesis one step further, the presence of the communication network urges us to question how the structure of the communication network shape up the society as each connection bring valuable information. The value generated through networks, is called *social capital*, which has been debated for a long time, for existence and also methods for measurement [36, 49]. In particular, we correlate social capital with the structure of the network to make inferences over four indicators: (i) economical well-being, (ii) civic engagement and democratic participation (iii) public health, and (iv) public safety and security. Contrary to the current literature [64, 37, 49], we observe different behaviour for economic well-being, and conclude that it could be the outcome of a developing country, since similar results are observed in another developing country, Colombia [128]. For civic engagement and democratic participation, our results demonstrate bridging structures, bring more inclusion to the society. Similar to that, again bridging concludes safer places for society. We could not find a significant correlation between public health neither bridging nor bonding network structure. As we find positive relation between diversity and society well-being, next we carry out our analysis to diversity of cultures in urban life. For this setting, we used data from a developed country for analyzing segregation through the city, we compare static census base residential segregation versus dynamic segregation through mobility of the phone users.

This thesis is structured as follows: in Chapter 2 we describe the mobile phone data then report the state of the art in the context of social good. We present our contributions in modeling aggregated human behavior in that context, in the following three chapters (3,4,5), which can be evaluated as stand-alone, complementary studies. Specifically, in Chapter 3 we exploit mobile phone usage data to detect countrywide anomalies and validate our approach by compiling an event dataset from various sources (shared in Appendix B), this chapter is the partial work from our publication [96]; in Chapter 4 we enlarge the scope of our analyses from spatial points (cell towers) in time series to network properties by constructing communication and mobility networks. The bridging and bonding structure of networks and the relation to socio-economic well-being of the region, have been analyzed. This chapter is composed of our journal submission, which is currently under review [97]. In Chapter 5, we move towards a bit geography domain, and extend network properties to spatial polygons, and in each uniform polygon we analyze how different cultures live, explore, and utilize the city; either in homogeneous or segregated way. Finally in Chapter 6, this thesis is concluded and future research directions have been elaborated. Addition to the main text of the thesis, supplementary supportive documents are given in appendices. In Appendix A detailed definitions of datasets are provided. The ground truth information which has been gathered from various sources are listed in Appendix B and Appendix C.

List of Publications

- Gundogdu, D., Incel, O. D., Salah, A. A., & Lepri, B. (2016). Countrywide arrhythmia: emergency event detection using mobile phone data. EPJ Data Science, 5(1), 25.
- Gundogdu, D., Panzarasa, P., Oliver, N., & Lepri, B. (Under Review). Measuring social capital through mobile phone usage: Evidence from a developing country.

The following paper has been published during the course of the PhD but is not included in this thesis:

- Gundogdu, D., Finnerty, A. N., Staiano, J., Teso, S., Passerini, A., Pianesi, F., & Lepri, B. (2017). Investigating the association between social interactions and personality states dynamics. Open Science, 4(9), 170194.

Conference talks and poster presentations

- Gundogdu, D., Moretti, M., & Lepri, B. (2017). Explorative analyse of two Italian cities: Turin and Venice for diversity, NETMOB 2017, Poster presentation.

- Gundogdu, D., Finnerty, A. N., Staiano, J., Teso, S., Passerini, A., Pianesi, F., & Lepri, B. (2016). Investigating the Role of Social Interactions on Personality States Dynamics, International Conference on Computational Social Science (IC2S2), Chicago, USA, Invited talk.

- Gundogdu, D., Emergency Event Detection Using Mobile Phone Data (2016), Oxford Big Data Symposium, Invited talk.

- Gundogdu, D., Incel, O. D., Salah, A. A., & Lepri, B., (2015) Social Event Detection in Aggregated Mobile Phone Data Using Markov Modulated Poisson Process, International Conference on Computational Social Science (IC2S2), Helsinki, Finland, Poster presentation.

In the rest of the thesis, mobile phone usage data and CDR are used interchangeably.

Chapter 2

Background

Mobile phone operators can analyze the behavior of a large number of people from their aggregated mobile phone usage [22, 209, 24]. In particular, the almost universal adoption of mobile phones is generating an enormous amount of data about human behavior with a breadth and depth that were previously inconceivable. In 2013, there were 6.8 billion mobile phone subscriber accounts worldwide, with millions of new subscribers every day, corresponding to a penetration of 128% in the developed world and 90% in developing countries [7].

In this chapter, we first explain the nature of data, and then list the usage of CDRs in the literature, grouped under various topics in social well-being.

2.1 Call detail records (CDR)

CDRs are mainly used for billing purposes by telecommunication operators. Every time a mobile phone communicates, a cell tower delivers the communication through the network, and a new CDR is created. This record reports the time of the communication, and the cell tower that handled it. Various types of CDRs exist, recording various types of communication (e.g., incoming calls, outgoing SMSs, etc.).

Table 2.1: An example from a call detail record.

Caller Id	Callee Id	Timestamp	Duration
123	125	20121004112000	0.056 sec
522	1154	20121004112000	0.125 sec

Access to these data is strictly controlled, because of data privacy issues. Occasionally research institutes and mobile network operators initiate challenges, to use data for good. To retain the user's anonymity, each phone number is coded such that the original phone number can not be obtained and no additional information related to the customer can be inferred. Another common approach is to aggregate the number of calls and the total duration, either in the level of cell towers or geographic region. The details of the two datasets, used in this theses are given in Appendix A.

In this thesis two different CDR datasets have been used.

- Data for Development (D4D) Côte d'Ivoire
- Telecom Italia Big Data challenge

These two datasets are dissimilar, apart from their country of origin, but also for the level of aggregation. Specifically, D4D data set is richer with four different level of aggregation. Subset-1 allows us to model information flow, as the number of calls are given as initiating cell tower to destination cell-tower. In Subset-2 the initiator cell tower data is available for a small subset of user, which allows us to model the mobility of the mobile phone users. Thus, we obtain the location data in cell tower precision for both subsets of Côte d'Ivoire, beside Telecom Italia dataset contains the aggregated data in grid level, where the dimensions are set by the challenge organizers. However the given data is aggregated in *country code* level, which lets us to be able to work on segregation, with the assumption that the caller/callee is from that country. As one can see that, neither of the datasets include the information as given in Table 2.1 for the privacy. Hence, the most fine granular data is from D4D dataset Subset-2, we have *caller id*, *cell tower id* and the *timestamp*. The more detailed information can be found for both of these datasets in Appendix A.1.1.

2.2 Mobile phone data and society

The CDRs can be exploited for various purposes, such as: to extract mobility patterns [77, 86, 116], to model social interactions [146, 183], to analyze the dynamics of a city [129, 54], to understand epidemics [215, 199], to estimate population densities [57], to predict energy consumption patterns [26], and socio-economic indicators and outcomes of territorial disputes [65, 28, 202].

In the following sections, we elaborate the related work that has leveraged mobile usage data: (i) to understand public safety and security, (ii) economic development, (iii) public health, (iv) democratic participation, and (v) segregation and diversity as these are the indicators that we focus on in this thesis with respect to their relationship to the social well-being.

2.2.1 Public security and safety

Large scale social events can happen anytime, anywhere, and without warning. Examples are clashes between ethnic communities, violence among supporters of political groups or sports clubs, demonstrations and celebrations. Some of these events cause migration phenomena (e.g., refugees and internally displaced people) [150, 152], higher crime rates and spreading of infectious diseases [20, 211], and result in economic crises [32]. Subsequently, researchers have recently started to automatically identify emergency events by using new sources of data, such as geo-referenced social media and mobile phone data [174, 16, 203].

Mobile phone usage data have been used as a proxy to understand the public safety. In this thesis we elaborate the use of this dataset in public safety and security from two different perspectives (i) detecting emergency events in Chapter 3 and (ii) the effect of diverse interaction patterns within and between different communities for safety and security in Chapter 4.

Detecting social and emergency events

In the last few years, several studies have shown that natural and man-made emergency events (e.g. earthquakes, floods, bombings, riots) can be reflected by dramatic increases in calling and mobility behaviors [203, 108, 13, 17, 81, 131, 77, 224, 10].

The assumption behind these work is that significant changes in behavior are captured by mobile phone data and indicate the occurrence of extreme events. Indeed, people tend to share and inform each other about an emergency event typically right after it is realized [83]. However, planned non-emergency events also occur and may provoke significant changes in mobile phone behavior. Bagrow *et al.* found easily detected changes in call frequency during festivals, concerts, sport events; however these changes may also be a result of other events such as holidays [13, 59].

Social interactions and safety

What kind of communities perform better in safety, close or open? The debate on that question has not been resolved yet. Coleman [49], Forrest and Kearns [72] focus on the importance of small, and core communities/networks for safety. Although it is in a way intuitive, the data driven approach points to the direction [28], in fact this theory has been favored by the well known urban theorist Jane Jacob's, as diversity brings safer neighborhoods in [105]. Our results supports the latter, which bridging is more informative metric to assess the safety of that community.

2.2.2 Economic development

Understanding economic growth has been an important question to answer, especially in developing economies with inadequate resources to enable systematic measurements. Recent studies focus on alternative methodologies to replace long run household surveys to infer economic development in a cheaper and more scalable way, such as using satellite images [220, 66, 69], correlating mobile phone usage with different data sources such as geopositioning data from search engine data [60], or credit card usage [98, 175]. Regarding previous work that has studied the relationship between phone data usage and economic factors, we must distinguish between individual and aggregate analyses. In terms of individual modeling, Blumenstock *et al.* recently worked on a large dataset from Rwanda with prominent results [24]. Through the analysis of the data from 856 subscribers of a main mobile network operator in Rwanda, they conclude that household possession and mobile phone usage is strongly correlated. Mao *et al.* propose the *CallRank* model, which analyzes in- and out-going calls to infer the socioeconomic conditions of a region [136]. San Pedro *et al.* developed the MobiScore system to automatically infer credit scores from mobile phone data [175] and evaluated it with ground truth from a Latin American country. Regarding aggregate modeling, Soto *et al.* [191] built machine learning models to automatically infer the socio-economic status of a region from mobile phone usage variables. Finally, Eagle *et al.* focused on a network perspective aimed at identifying social ties and structural holes and their relation to the economic growth in the United Kingdom [64]. Their findings are consistent with Burts' theory of structural holes and their benefits to communities [37].

2.2.3 Public health

Throughout the history, human beings have been threatened by pandemics. Maybe the most famous one is *black death*, the plague disease ends up the lives of 75 to 200 million people in Eurasia between 1346 and 1353 [88]. Currently the world is more connected than in those times with transportation backbones and new technologies. This high integration however also brings possible global risks, to disseminate diseases very fast and vast coverage. Such as recent Ebola outbreak has spread from west Africa to Spain, United Kingdom and United States with some fatal incidents. Understanding mobility patterns is crucial to take preventive initiatives against disease dissemination for the public safety [156, 14]. Thus, the availability of vast coverage of mobile phone datasets, allowed scientist to undercover the mobility patterns of human beings [113, 142] to model the infectious diseases like malaria [145, 215, 195], ebola [214], H1N1 etc.

2.2.4 Democratic participation

Democratic participation is a kind of civic engagement like, being involved into volunteer organizations, reading local newspapers, supporting a local football team. Apart from the other features we observed, democratic participation targets the collective benefits, rather than individual. There are not much work on using mobile phone usage for assessing the influence effect of civic engagement. On the other hand, in the literature there are quite flourishing discussions on how to use mobile phone infrastructure as a framework for elections [133, 43]. In particular, the participating through SMS to Uganda elections [100], in the context of Government 2.0 [48], or for the smart cities of our future [15].

2.2.5 Segregation and diversity

Mobility in the context of urban planning has been analyzed by many scholars with the availability of mobile phone usage datasets [39, 44, 116]. Russian and Estonian segregation has been analyzed through language preferences by the mobile phone operators in [106]. In that work, the findings show that the minority group Russian, presents a limited mobility, comparing to Estonians, and the presence of segregation is more distinguishable when the observation period extends. Various studies use complementary data to amplify the use of CDR dataset, such as Flickr data [84] or taxi and bus routes [38] to assess the tourists' movements in the city.

Chapter 3

Emergency Event Detection

Large scale social events that involve violence may have dramatic political, economic and social consequences. Such chaotic life conditions, creates forms of suffering, destruction for the well-being of the communities. To sustain the social fabric of the societies, the way we handle the disasters is vital [159]. Hence, researchers have started using mobile phone data for developing tools to identify such emergency events in real time.

In the case of an emergency, defining the casualties and taking action aftermath is *time critical*. In the age of information, both noise and actual signals disseminate very fast, thus reliability and *accuracy of the information* has to be confirmed. In a scenario of an earthquake disaster, which effects a large geographical region, after defining the effected locations, *prioritizing* is needed as the sources are limited. Under these circumstances, not being able to evaluate information evenly, even worse if the facts are contaminated with misinformation, may conclude in a serious consequences. In the case of a conflict among ethnic or political groups the authentication of the information is a vital challenge. Thus, it is necessary to build such applications from multiple sources [52]. Detecting emergency events from CDR can be implemented as a complementary tool, to overcome the misinformation challenges.

We apply a stochastic model, namely a Markov modulated Poisson process (MMPP), for spatio-temporal detection of hourly and daily behavioral anomalies. We use the call volumes collected from an entire geographic region. Our work is based on the assumption that people tend to make calls when extraordinary events take place. We validate our methodology using a dataset of mobile phone records and events (emergency and non-emergency) from the Republic of Côte d'Ivoire (see Appendix A). Our results show that we can successfully capture anomalous calling patterns associated with violent events, riots, as well as social non-emergency events such as holidays, sports events. Moreover, call volume changes also show significant temporal and spatial differences depending on the type of an event. Our results provide insights for the long-term goal of developing a real-time event detection system based on mobile phone data.

3.1 Previous approaches in event detection

The call data can be analyzed at the individual level, to detect changes from the expected call and mobility behavior [203, 9, 61]. This approach requires some restrictions for data privacy, and incurs high computational cost. These problems can be overcome by processing on the aggregated call volume of each cell tower [224].

Supervised approaches, which have high accuracies, start with the time and the location of an already known event and then look for anomalous calling behavior at that time and location by defining a baseline. In this thesis we focus on unsupervised approaches that can detect anomalies in streaming data thus can be used in the time of the event, which is more applicable to real life.

Recently, Dobra *et al.* proposed an unsupervised behavioral anomaly detection system that identifies days and locations with unusual calling or mobility behavior without knowing the event [59]. The authors used mobile phone records from Rwanda in order to connect the identified anomalous days and locations with extensive records of violent and political events (e.g., protests, violence against civilians) and natural disasters (e.g., earthquakes). Specifically, they computed the aggregated daily mobility and calling patterns of each site, starting from the individual behavior of all assigned subscribers for the corresponding site. This methodology suffers from high computational cost and works only on a daily basis. Our approach instead, takes as input the calling volume of a cell tower without accessing individual data. Hence, we are also able to detect hourly variations and the behavioral responses occurring within hours of an event.

There are two main approaches for classification of emergency and non-emergency events from mobile phone data. The first approach is the analysis of mobility patterns to model individuals and crowds, which has been used to identify concerts and sport events [203, 40]. The second approach is the analysis of temporal patterns, which can be used to differentiate between event types according to the duration of behaviors [224]. We propose a third approach by observing spatio-temporal characteristics of anomalous behaviors.

The D4D Côte d'Ivoire dataset [23] is used by numerous studies to validate event detection approaches. Paraskevopoulos *et al.* [158] used numerical analysis to detect anomalies. Their model is based on call duration statistics, which is calculated by dividing the cumulative call duration by the number of calls, for each day and for each cell tower. During an emergency, people may indeed tend to make more calls, but these can be with shorter duration. Hence, looking at the total duration may not be sufficiently discriminative. Dong *et al.* used the D4D dataset for modeling the movements of flocks of people [61]. Their assumption is that people move in groups when an extraordinary event happens. This assumption can provide valuable insights to detect and characterize protests. However, this mobility pattern is not observed during attacks against civilians that take place in urban areas. In Section 3.4 we will provide comparative results with the two recent methods described briefly here.

In the literature, anomalous event detection, which corresponds to the task of detecting anomalies from time series, is strongly related to *outlier detection* [226], and *change point detection* [46, 168]. Previously, Zhang *et al.* used change point detection approach to detect anomalies in mobile phone data [225]. However, the problem of this method is that it does not preserve the periodicity of data. Our data is representing quite similar behavior for each day; low from midnight until morning and slightly decrease at noon and high during the day, as shown in Figure 3.1.

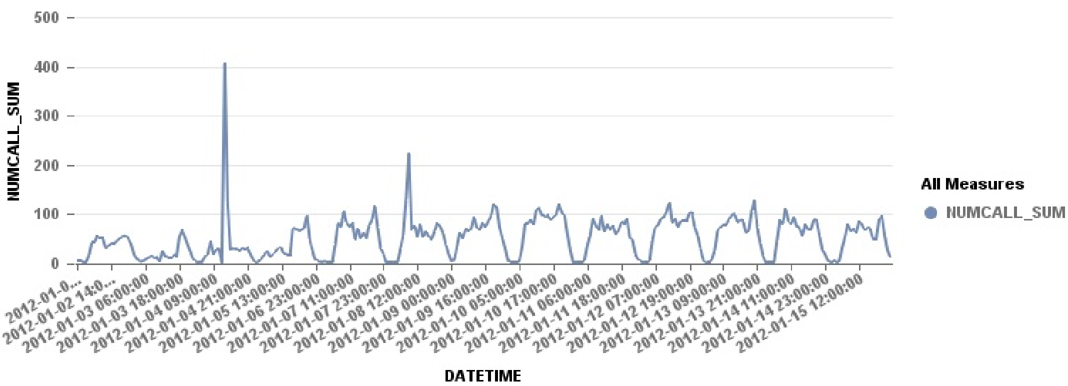

Figure 3.1: Number of calls from a cell tower number 524, from 1st of January to 15th of January 2012.

In our application scenario, the normal behavior is observed by the temporal (hourly) changes of the aggregated call volumes, collected from different cell towers. Similar discrete, count-based analysis problems have been tackled in statistics, econometrics, psychology, and ecology [78, 173, 41, 216].

Other researchers proposed the usage of Hidden Markov Models (HMM)

[132, 218] for the anomaly detection task. As highlighted by [18], the Markov modulated Poisson process (MMPP) approach that we propose to use in our work is a special case of HMM and Markov Chain Monte Carlo (MCMC) approaches. MMPP has also similarities with the change point methodology [47] but it has the advantage of preserving periodicity.

3.2 Stochastic model for anomaly detection in periodic data

Stochastic models can be used to make accurate predictions even with incomplete data. CDRs, even they generally cover very large geographic regions with quite representative number of people, has its own limitations. For example the Mobile Network Operator (MNO) (the data provider) may not represent the whole population. Other limitations are detailed in Section 6.2.

The nature of CDRs presents a periodicity in the number of calls for each unique cell tower. A similar behaviour in data has been found a work from Ihler *et al.* [104], in particular to understand an event (e.g. concert, football match) by counting number of cars passing by a stadium or (e.g. meeting, conference) by counting the number of people entering the university. Therefore we implement the methodology Markov modulated Poisson process, to understand anomalous events through the number of calls, taken place by the cell tower. Implementation of MMPP can be seen as extending the work of Dobra *et al.* which has mentioned in the previous section [59]. The periodicity which is repetition of similar behaviour in each hour of the day, is modeled through graphical models, the details of the model is given in the following section.

3.2.1 Markov Modulated Poisson Process

Our goal is to investigate the usage of mobile phone data to detect and characterize anomalous events. Poisson processes are commonly used to model count data (data in which the observations can take only the non-negative integer values) [104, 184] in many domains, such as modeling rare incidents in psychiatric hospitals [78], and traffic analysis [185, 222].

In our setting, the count data are represented by the number of total calls in 1-hour time intervals, denoted as $N(t)$, for each cell tower, denoted as k.

The model is represented as a two state Markov chain as shown in Figure 3.2, with a normal call behavior state $z(t) = 0$ and an abnormal call behavior state $z(t) = 1$ (see Equation 3.1). The transitions between the states are defined with a transition probability matrix M_z in Equation 3.2.

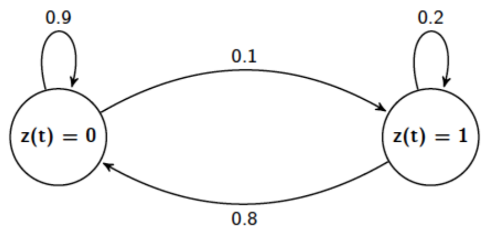

Figure 3.2: Event transition state diagram, showing $z(t) = 0$ normal and $z(t) = 1$ abnormal states.

$$z(t) = \begin{cases} 1, & \text{if there is an event in time t.} \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

$$M_z = \begin{pmatrix} 1 - z_0 & z_1 \\ z_0 & 1 - z_1 \end{pmatrix} \tag{3.2}$$

Our observations are denoted as $N(t)$, and hidden variables are the amount of calls in a normal call pattern $N_0(t)$, and the amount of calls initiated because of an anomalous event $N_E(t)$. The definition of $N^k(t)$ for a cell tower k is given by the summation of hidden variables $N_0^k(t)$ and $N_E^k(t)$, as shown in Equation 3.3.

$$N^k(t) = N_0^k(t) + N_E^k(t) \tag{3.3}$$

Hence, we assume that the number of calls are generated through a heterogeneous Poisson distribution $\text{Pois}(N, \lambda(t))$, with the rate value that is a function of time $\lambda(t)$, as shown in Equation 3.4. Changes in the count data are modeled taking into account the periodicity depending on the day and on the hour of the day. In Section 3.2.2, this property is used to calculate the posterior distributions.

$$Pois(N; \lambda(t)) = e^{-\lambda(t)}(\lambda(t)^N / N!) \tag{3.4}$$

The normal call pattern of a cell tower k is dependent on the day (i) of the week (δ_i^k) and the hour (j) of the day $\eta_{j,i}^k$, and λ_0^k represents the average rate of cell towers in one week. We can formulate the rate function of the Poisson distribution $\lambda^k(t)$ as the product of the initial rate λ_0^k, the daily effect δ_i^k, and the hourly effects $\eta_{j,i}^k$, Equation 3.5. Plate notation is used to depict the repeating form of the data, as shown in Figure 3.3.

$$\lambda^k(t) = \lambda_0^k \delta_{d(t)}^k \eta_{d(t),h(t)}^k \tag{3.5}$$

Conjugate prior ensures that the posterior distribution is coming from the same family of the prior distribution. Hence, our model becomes tractable and it can be written in closed form. Random variables δ, η, denoting the day of the week and the hour of the day, are coming from a multinomial distribution as shown in Equation 3.6 and the conjugate prior is selected as Dirichlet distribution in Equation 3.7.

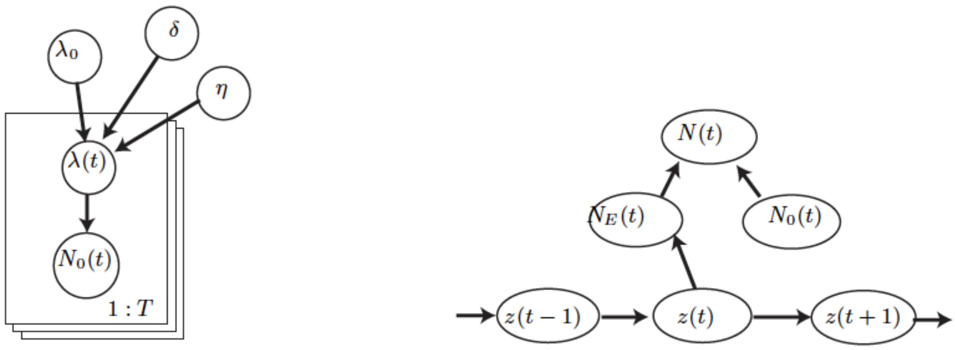

Figure 3.3: **Left**: The graphical model to understand hour of day and day of week effect. **Right**: Observation $N(t)$ and hidden variables $N_0(t)$ and $N_E(t)$, taken from [104].

$$\begin{aligned}\sum_{i=1}^{7}\delta_i &= 7\\ \sum_{i=1}^{D}\nu_{j,i} &= D,\ \forall j.\end{aligned} \tag{3.6}$$

$$\begin{aligned}\lambda_0 &\sim \Gamma(\lambda; a^L, b^L)\\ \frac{1}{7}[\delta_1, ..\delta_7] &\sim Dir(\alpha_1^d ... \alpha_7^d)\\ \frac{1}{D}[\eta_{j,i}, ...\eta_{j,D}] &\sim Dir(\alpha_1^h ... \alpha_D^h)\end{aligned} \tag{3.7}$$

In the left side of the Figure 3.3, the day (δ) and hour (η) effects are shown in graphical form. Plate notation is used to depict the repeating form of the data. On the right side of Figure 3.3, $z_{(t-1)}$, $z_{(t)}$ and $z_{(t+1)}$ show the time series properties of the event transitions. $N_0(t)$ and $N_E(t)$ are hidden variables, and $N(t)$ is our observation.

$$z_0 \sim \beta(z; a_0^Z, 0^Z) \tag{3.8}$$

$$z_1 \sim \beta(z; a_1^Z, b_1^Z) \tag{3.9}$$

$$N_E(t) = \begin{cases} 0, & z(t) = 0 \\ P(N; \lambda(t)), & z(t) = 1. \end{cases} \tag{3.10}$$

$$\lambda(t) \sim \Gamma(\lambda; a^E, b^E) \tag{3.11}$$

Event probabilities can be sampled through the densities of a normal event $P(N; \gamma)$ times the probability of having an anomalous event $\gamma(t)$ as shown in Equation 3.12.

$$\int P(N; \gamma)\Gamma(\gamma; a^E, b^E) = Nbin(N; a^E, \frac{b^E}{1 + b^E}) \tag{3.12}$$

3.2.2 Inference and Parameter Estimation

In our model, we need to estimate the parameters for the Gamma distribution Γ $(\gamma; a^E, b^E)$ in order to model the probability of an anomalous event $p(z(t)^k|N(t)^k)$.

The hyper parameters of Gamma Distribution Γ, a^E and b^E in Equation 3.11 set the distribution's sharpness or smoothness between the transition states Z_0 and Z_1. In traumatic events, the transition generates a sharp peak.

Let us assume $\{N_0(t), N_E(t), z(t)\}$ are given. Then, we may estimate the parameters of the distribution by computing the maximum likelihood [46], given that all other variables are conditionally independent. However, as in Figure 3.3, only $N(t)$ is observed, without separation into $N_0(t)$ and $N_E(t)$. The rest of the parameters, including $z(t)$, are estimated from the observations. Since we do not know when a social or emergency event takes place, we estimate the parameters of the model with the information we obtained. In that respect, we take one cell tower, which is known for having an associated event, and calculate the deviation from the mean value of the hour of the day $\eta^k_{j,i}$, and day of the week δ^k_i from the observation, as in Equation 3.3. After tuning these parameters for a single cell tower, we apply them to all $N_E(t)$ in the dataset. We exclude the cell tower that is used for calculating the hyper parameters from the evaluation set. Since the model may be sensitive to the selection of hyper parameters, this selection is important for the evaluation of the model. We recommend the selection of a cell tower that represents the average rate function of the all cell towers. Furthermore, the selected antenna should have a number of observed events, so that the rate change can be quantified. The duration of the experiment is not important if the data can represent the normal behavior of the cell tower, which corresponds to the mean value of the call volume in our case.

3.3 Evaluation of the model

To evaluate event detection models, it is necessary to annotate existing data by cross-referencing it with news sources. We provide a list of important events used in our analysis in Appendix B. For each important event, we grouped the cell towers close to the location provided in the ground truth, and associated events to cell towers with some certainty. Each cell tower's probability distribution is normalized before calculating the average probability of the region. If the event probability is higher than a defined threshold (τ) (e.g. 0.15), we classified the output as an event has been detected in the region.

To assess the MMPP's performance, we have implemented a baseline model, which we describe in Section 3.3.1. In addition to that, the results are compared with the work of Dong *et al.* [61] and Young *et al.* [224] detailed in Section 3.4.

3.3.1 Baseline Model

For each cell tower, the hourly average call volume is calculated as shown in Equation 3.13. In our notation, each cell tower is denoted with k, days are indexed with i, hours with j, $N(t)$ is the observation for time t, and D is the total number of days for the data collection period. Once the averages are calculated, we subtract the average value, Φ, from the observed call volume for time t. If the obtained value is higher than a defined threshold (τ), the event is labeled as anomalous.

$$\begin{gathered}\Phi^k_{i,j} = \frac{\sum_{i=1,j}^{D}\sum_{i,j=1}^{24} N(i,j)^k}{D}, \forall k \\ \left|(N(i,j)^k - \Phi^k_{i,j})\right| > \tau\end{gathered} \tag{3.13}$$

3.3.2 Event Records

Our data on violent and political events, major holidays and major events (e.g., elections, the African Football Cup) come from a various structured and unstructured public data sources such as Armed Conflict Location and Event Data Project (ACLED) [2], United Nations Council and International Crisis Group reports, local and international news. ACLED and UN Security reports include extensive data on conflict-related events including riots, protests, killings, and battles. The information is obtained by local or international newspapers (e.g. Notre Voie, Le Patriote, France24, BBC) and radio sources, and it includes details such as the date and location of each event, the type of event, the groups involved, and the fatalities. Côte d'Ivoire has been suffering unstable political conditions during the data collection period. According to the United Nations Human Rights Department, more than 600,000 Ivorians were displaced in the country and around 200,000 Ivorians migrated to neighboring countries in order to be in secure living conditions [5].

In total, we have gathered 19 emergency events (e.g., confrontations between groups, protests) and 11 non-emergency events (e.g., national and regional holidays, African Football Cup games). The complete list of events (emergency and social non-emergency events) is shown in Table B.1 and Table B.2.

3.4 Results

Our approach identifies many hours and days with unusual increase in calling volume. As in [59], these anomalies are found sometimes in a specific site and sometimes across multiple sites. Specifically, our approach provides the probability of having an anomalous event in a specific location.

Our goal is to identify emergency events. However, non-emergency events might also produce behavioral changes in calling volume. Hence, it is relevant to highlight and discuss the specific behavioral signatures of different events in order to detect emergencies effectively. In this regard, from the 19 annotated emergency events, our method automatically detected 15, while for the social non-emergency events it was able to identify 8 out of 11 events. Similar to [59] and [104], we find some hourly and daily anomalies that we were not able to match to any of the recorded events (i.e. possible false positives).

We compare our MMPP approach with a baseline model, which has been explained in Section 3.3.1, and show that our approach outperforms the baseline. Indeed, from the 19 annotated emergency events, the baseline approach is able to identify 8, while for the social non-emergency events, it is able to identify 7 out of 11. While the nature of the problem dictates a very large dataset with very sparse true positives, the results are quite promising. Before describing our results in detail, we compare our method with the event identification approach by Dong *et al.* [61], as well as with the event type classification approach described by Young *et al.* [224] in the next subsection.

Dong *et al.* proposed a methodology in 2015 for event identification based on modeling individuals' mobility as flocks through the city, and they tested this methodology on the D4D dataset [61]. They reported a *precision* value of 0.0676 and a *recall* value of 0.9200. In order to provide a fair comparison, we tested the MMPP approach on the same experimental setting proposed in their paper, with 140 days of data and 25 ground truth events. The results show that MMPP outperforms this approach in all the measures (see Table 3.1).

Table 3.1: Comparative results for the MMPP model and for Dong *et al.*'s approach [61].

	Precision	**Recall**	**F-Score**
MMPP	0.31786	0.9219	0.4727
Dong et al.	0.0676	0.92	0.125

We also compare our approach with the one proposed by Young *et al.* in 2014 to classify social non-emergency events [224]. This approach posits that social events have longer temporal duration than emergency events. Since the duration of deviation is a major parameter, we compared the methods for observation periods of two, three, and four consecutive hours. Setting this to two hours means that an event is detected if the observed behavior deviates from the expected for two hours. For each of the three settings, Figure 3.4 shows the number of social events detected by a given number of cell towers. The best performance is obtained by using two consecutive hours, where the average number of detected social events is equal to 1.34. Using MMPP, on the other hand, we are able to detect 8 social events.

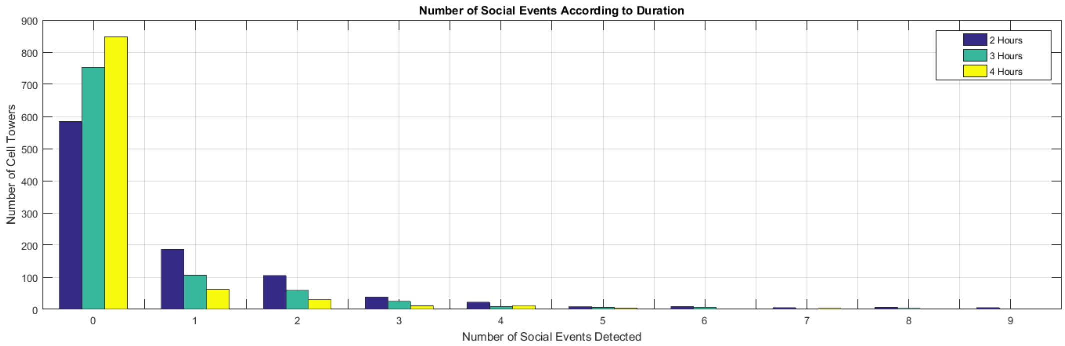

Figure 3.4: The distribution of detected social events with the approach of [224]. Selecting a longer duration threshold for unexpected behavior reduces the number of detected events.

3.5 Discussion

In this section, we match some of the anomalies identified along with key events that occurred in Côte d'Ivoire and we discuss what we have learned from matched and unmatched events.

In 2010 and 2011, Côte d'Ivoire suffered from a *post election crisis*. The election results initiated a civil war, and the effects of the war were still present during the data collection period. Therefore, our emergency events are mainly due to these unstable political conditions.

Violence against civilians - February 3, February 19, and March 8, 2012. On February the 3rd, the proposed system did not detect any significant changes in call volume near the location of the event, close to Bouaké. Instead, we see a clear anomaly in this area on February the 4th (see Figure 3.5).

Similarly, no anomalies were detected on February the 19th, while on March the 8th we detect an anomaly close to Bouaké (cell tower 964) around 11am. Interestingly, this finding is not detected by the baseline model.

Figure 3.5: Left: The cell tower locations on the Bouaké-Katiola road. Right: The corresponding probabilities of having an event on 4th of February.

Violence against civilians - February 11 and February 13, 2012. In Arrah, it was reported that the confrontation of two groups ended with three people killed and at least 19 injured. The visualization of our model's output is shown in Figure 3.6 for cell tower 113. In the top plot, the red line denotes the hourly and daily averages and the dark line denotes the observed calling volume. The plot in the middle shows the predictions of an event occurring, with the annotated events marked below.

Protests against the government - February 18, 2012. MMPP detects calling volume increases from a single cell tower at 3pm and 5pm on February 18, 2012. The cell tower is located in the city center of Abidjan, the economic capital of the Côte d'Ivoire. A non-violent protest against the government

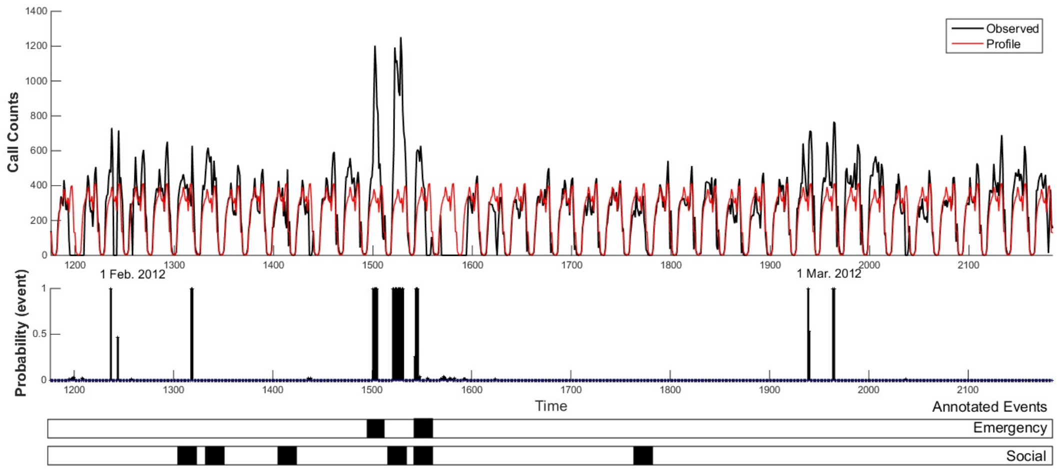

Figure 3.6: MMPP results for cell tower 113 (Arrah). In the top plot, the red line denotes the hourly and daily averages and the black line denotes the observed calling volume. The plot in the middle shows the predictions of having an event, with the annotated events marked below.

was reported by the United Nations on the same day in Abidjan, in front of the "Congres National de la Resistance Pour la Democratie" headquarters located in the city center. We assume those protests took place in the western part of the map, shown in Figure 3.7. Each orange dot represents cell towers with an anomalous call volume. Unfortunately, the reports of United Nations do not specify the duration of the protest.

Elections and violent clashes in Bonon and Facobly - February 26, 2012. After 5pm, we observe anomalous calling patterns in two regions of the mid-west, Bonon and Facobly, for the cell towers 239, 1154, 374, 181. On the same day, a couple of events were recorded for these two regions, political elections and consequent violent clashes between political opponents, as shown in Figure 3.8. The violent events ended with the death of five people. Interestingly, we observe that these violent events seem to have effects not only in the regions involved, but also in Abidjan. Indeed, we detect an

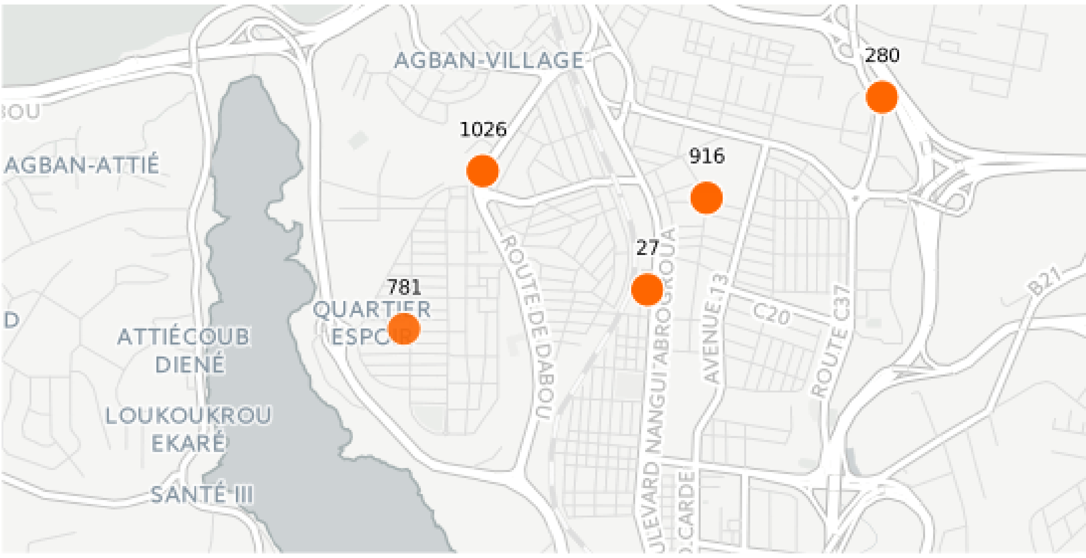

Figure 3.7: Cell tower locations of Abidjan with high call volume anomalies during the "Congres National de la Resistance Pour la Democratie" meeting on the 18th of February.

anomaly in Abidjan after 5pm. Although Abidjan is not the administrative capital of the Côte d'Ivoire, the headquarters of the two opposition parties are located in this city.

Violence against civilians - February 29, 2012. The event records report that FRCI (Forces Republicaines de Côte dIvoire) shot civilians in Séguéla, and this violent repression resulted in the death of two people. On the same day, we observe an unusual calling activity in the northern part of Séguéla, close to the border with Guinea.

As previously mentioned, we also collected 11 non-emergency events (social events and holidays). The main characteristic of those types of events is that they affect almost the entire country. In the following, we discuss a couple of relevant non-emergency events.

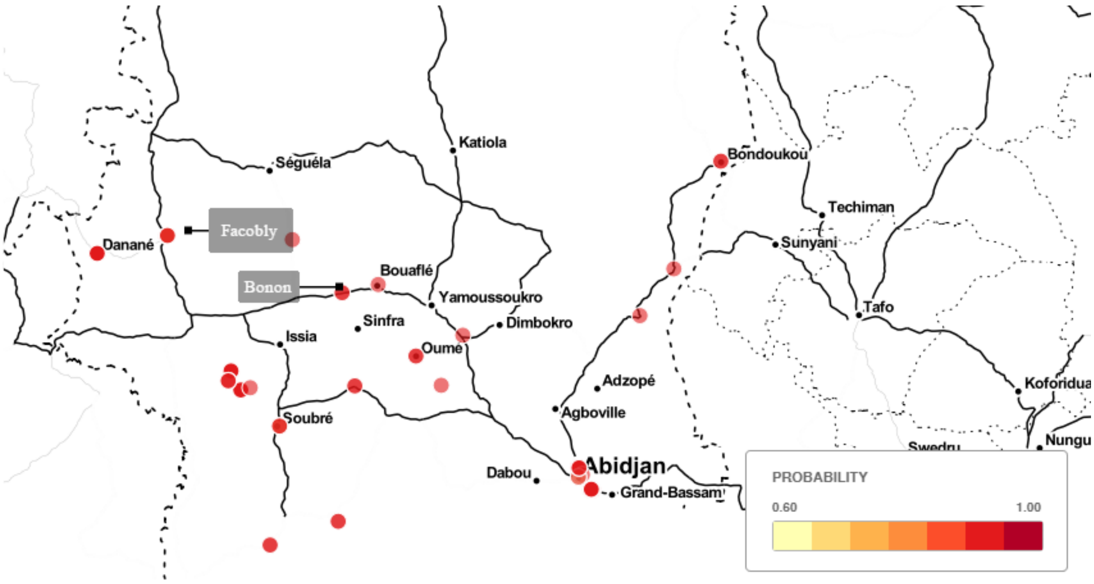

Figure 3.8: Overall country anomaly probabilities on February 26, during renewal of elections in Bonon and Facobly.

African Football Cup, January 21 February 12, 2012. In 2012, the African Football Cup was held in Equatorial Guinea and Gabon. Côte d'Ivoire played in the final match and dramatically lost the championship on penalties. It is worth noting that football is one of the most important social activities in Côte d'Ivoire and that the day after the finals was proclaimed a national holiday. This enthusiasm is spread across the entire country. Interestingly, African football cup matches have a highly specific call signature, causing a very high call volume for one or two hours after the match on cell towers across the country. Figure 3.9 contrasts the anomalies detected before and after the match. Specifically, the left side of Figure 3.9 shows the anomalies detected before 8pm and the right side shows the ones detected after 8pm.

Mavlid an Nabi (celebration of the birth of the prophet), February 4 - February 5, 2012. The Côte d'Ivoire is composed of different ethnic groups and religions. The north is mainly Muslim and the south is mainly Christian. During these two days we found calling anomalies in the northern part of the country.

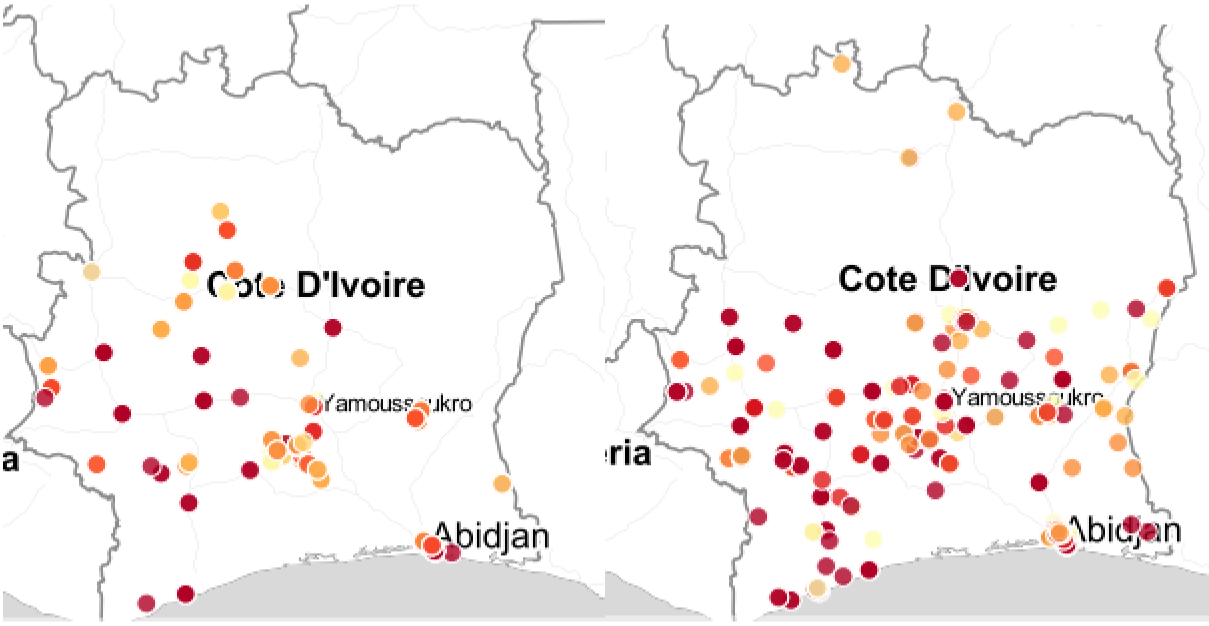

Figure 3.9: The detected anomalies on the 4th of February. On the left before 8pm, on the right after 8pm. Darker dots correspond to higher probabilities.

Ash Wednesday Christian festival, February 22, 2012. We do not observe any significant difference in the calling patterns of the Christian south and the Muslim north. Only in the capital city, Yamoussoukro, during the whole day and night we register some unusual calling activity. Interestingly, Yamoussoukro has the biggest church in the world, the Basilica of Our Lady of Peace. This monument usually sees large number of Christians gathering during the festival.

Our results do not confirm the patterns observed by [187] and [161] that found a positive correlation between the mobile phone coverage and the probability of having a conflict. In our setting, the densest population and the highest coverage is in the Abidjan area, while the violent events tend to happen in the mid-west of Côte d'Ivoire, where many conflicts between different ethnic groups exist.

Another finding is an irregularly high call volume on the first day of each month. One possible reason is the high penetration of mobile payments in Côte d'Ivoire. People typically pay their utilities, rents, etc. on this day. However, we do not have any formal way of confirming this, and we assume this behavior is normal. Note that our work suffers from a number of limitations. First of all, we focused our attention only on anomalous increases in calling patterns. However, emergency and non-emergency events may also cause changes in mobility patterns, as shown by [59]. Second, the annotated event records may be incomplete given the scarcity of information and the chaotic socio-political situation during the data collection period. Moreover, the exact event locations are often not provided and the dates can fluctuate a bit in the reports. For example, UN records report events in Arrah from February 11 to February 13, while *Le Figaro* newspaper [4] reports the same events from February 12 to February 13 (our results seem to confirm the dates provided by *Le Figaro*). Third, we observed daily data losses in several cell towers, e.g. *cell tower 49* on the 5th of December. This data loss tends to generate a high number of false positives when the average call volume of a cell tower is high, e.g. greater than 1500. Finally, the MMPP performance is sensitive to prior parameters, even though we have empirically shown that cell tower call volumes can be easily and robustly estimated from a few days worth data [104]. Yet, in our case, the models performance may be sensitive to the cell tower selected for tuning the parameters.

3.6 Conclusion

In this work, we have proposed an approach based on Markov modulated Poisson processes for spatio-temporal detection of hourly and daily behavioral anomalies in call volume. It is based on the assumption that people tend to make calls when extraordinary events take place. We validate our methodology using a dataset of mobile phone records, together with emergency and non-emergency events from Côte d'Ivoire. Our results show that we can capture anomalous calling patterns associated with violent events, protests, holidays and major sport events (e.g. African Football Cup games). One of our findings is that the impact area is a better feature than the temporal duration for identifying social non-emergency events from mobile phone data, which runs counter to the general opinion held in the literature. Another valuable finding is that for event detection, analysing mobile phone activity as a time series gives better performance compared to tracing movements of the masses. It is worth noticing that our methodology uses only aggregated call volumes of cell towers and no data can be traced back to individuals. Hence, there are minimal - if any - privacy concerns.

In sum, we believe that our work contributes to the process of creating an effective emergency detection system that may be used by governments, policy makers, and international organizations to significantly increase human well-being and the wealth and security of countries.

Chapter 4

Social Capital: The Value Generated Through Connections

Sociologists and economists have been debated on the intrinsic value of social relations in the last decades. Namely, social capital has long been associated with the opportunities of accessing the valuable resources that actors can tap from the social structure underlying the interactions among them. Despite the overall consensus on the structural signature of social capital, there is still arguments over the relative benefits associated with different types of social structure, and their impact upon individuals and society. In this chapter, we advocate a two-faceted perspective on social capital, as value originating from both closed bonding structures, rich in third-party relationships, and open bridging structures, rich in brokerage opportunities. We aim to shed light on how measures of social capital, computed using data on mobile phone usage, correlate with key social indicators. To this end, we draw on aggregated mobile phone usage data, and investigate social interactions uncovered from the communication and mobility networks in the city of Abidjan in Côte d'Ivoire. We begin by defining network metrics that characterize the social capital of each of the 10 districts (communes) in Abidjan. We then compare such metrics across districts and find significant differences between them. Finally, we compute correlations between the defined social capital metrics and a number of district-level socio-economic indicators broadly related to socio-economic well-being, health, and democratic participation.

Using mobile phone data to measure social capital has important advantages with respect to the previous approaches, namely: it collects data on everyday life rather than experimental data collected in the lab; it is generally large-scale considering the large levels of adoption of mobile phones both in developed and developing economies; and it provides a methodology that could be easily extended to other locations and countries.

While mobile data could offer significant advantages as a relevant source of information related to the social capital of a region, there has been little research carried out to explore how measures of bridging and bonding social capital, computed using data on mobile phone usage, correlate with key social indicators, in particular in developing countries [192]. This work aims to fill such a gap, validating hypotheses and theories on bridging and bonding social capital in Abidjan, the economic capital of a developing African country.

Our main findings suggest that (1) mobility and communication networks act as sources of social capital through different mechanisms; (2) social cohesion is beneficial especially within wealthy communities; (3) node degree and density do not accurately represent economic well-being, especially in mobility networks; (4) there is no universal explanation for developing and developed countries, which have different characteristics. By providing empirical evidence of the existence of different structural sources and generative mechanisms of social capital across a large city's different districts; (5) different network structures contribute different aspects of the social capital; in a developing country for economic well-being bonding is preferred, however for civic engagement and security bridging is better. Our work contributes to a better understanding of the intricate relationship between communication and mobility networks in a developing country, and how these networks can enhance and sustain socio-economic growth and prosperity.

4.1 What is social capital?

In this study, we define social capital as the value generated from the "investment in social relations, bonding similar people together and bridging diverse people, with norms of reciprocity" [45, 56]. Hence, social capital consists of building trust networks in a society and measuring the value derived from the existence of such networks.

However, social capital does not have a unique undisputed definition. Network-based definitions of social capital differentiate between two types of social capital, namely *bonding* and *bridging*:

- *Bonding social capital* is based on building trust between individuals who share the *same social identity*, through a close and dense network. Thus, this form of social capital derives from the mutual trust present in the group. Traditional within-group favoritism is an example of this type of social capital [74].

- *Bridging social capital* is the value generated through the connection of individuals with *different characteristics* (*e.g.* socio-economic levels, ethnicity, or cultural background). The positive effect of this form of social capital is to support the flow of ideas and the access to different information sources. Bridging social capital is quantified through the analysis of *spanning structural holes*, which allow different communities to interact and connect through weak ties [36].

In addition, we can consider two levels of aggregation for social capital: *individual* and *collective*. *Individual social capital* refers to the benefits that an individual enjoys thanks to his/her belonging to trusted social relations. *Collective social capital* captures the aggregate benefits of a certain community derived from the existence of bonding and bridging social connections that follow accepted norms of reciprocity [119].

Thus, even if the importance of social capital has been established in the literature [126, 92, 91], it is not easy to empirically measure it. In the sociology and economics literature, we find three methodologies to measure social capital: (i) surveys [75], (ii) measures that collect data on non-profit associations density and on selfless activities such as donating blood or voting for a referendum [162], and (iii) experimental in-lab measures [19].

In the last decades, social capital has been studied in sociology [49, 164, 31, 37], economics [85], and political science [75] as one of the key elements to promote the development of modern societies. Several works have highlighted the role played by social capital for economic growth [163], providing evidence that it reduces the transaction costs associated with formal coordination mechanisms [114], and predicts strong economic performance and financial development [93]. Moreover, additional direct and indirect benefits have been attributed to social capital, such as helping people find new job opportunities [90, 42], create new businesses [208, 178], enable safe and secure residential environments [72], improve public health [206, 109, 82], increase democratic participation [164, 207] and reduce corruption [68, 153].

4.2 The importance of information diffusion to understand social capital

Information diffusion is a key variable that contributes to the creation of social capital. The underlying assumption of social capital is that social networks carry valuable information for the well-being of the peers. Burt's theory of structural holes states that new information flowing through distant social connections could generate prosperity. However, the absence of these types of connections creates social cohesion, which results in the same information circulating through the network [37]. In parallel to this theory, Onnela *et al.* show the importance of weak ties or structural holes to enable information flow [156].

Information diffusion has recently been analyzed by Smith and Capra through the analysis of mobile phone data [190]. Mobile phone usage data allows us to analyze not only the communication patterns of a very large population, but also their mobility patterns to infer possible physical connections between different places [58]. Diffusion of information is generally modelled through epidemic models, such as the Susceptible-Infectious-Recovered (SIR) model [111]. Miritello *et al.* used a SIR model to explain information diffusion with temporal and topological features of social ties by modelling human interactions via mobile phone usage data [147].

Here, in Figure 4.1, we visualized the aggregated calls taken place in Côte d'Ivoire in week 50, to show how information disseminate among the country.

4.3 Quantifying social capital through communication network

In this chapter, we propose using mobile phone data to construct network variables as a proxy to measure social capital. We focus on *collective social capital* at a district (commune) level in a large city in a developing economy (*i.e.* Abidjan, the economic capital of Côte d'Ivoire). In addition, we analyze the relationship between the proposed mobile phone data-based metrics and several socio-economic indicators.

Using mobile phone data to measure social capital has important ad-

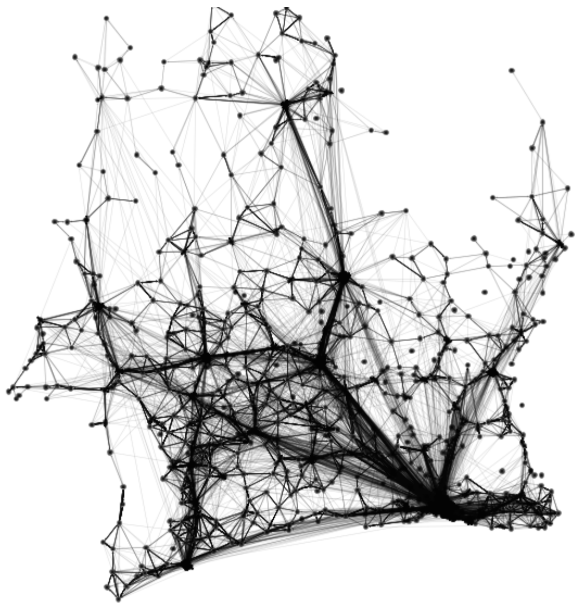

Figure 4.1: Aggregated calls from Week 50, showing the information diffusion.

vantages with respect to previous approaches, namely: it collects data on everyday life rather than experimental data collected in the lab; it is generally large-scale considering the large levels of adoption of mobile phones both in developed and developing economies; and it provides a methodology that could be easily extended to other locations and countries.

While mobile data could offer significant advantages as a relevant source of information related to the social capital of a region, there has been little research carried out to explore how measures of bridging and bonding social capital, computed using data on mobile phone usage, correlate with key social indicators, in particular in developing countries [192]. This work aims to fill such a gap, validating hypotheses and theories on bridging and bonding social capital in Abidjan, the economic capital of a developing African country.

As the structure of the network impacts the outcome, network properties have an important role to play in the characterization of social capital. The capital can be utilized depending on the bonding or closure of the network, according to [30, 49, 162, 37]. This may help to share resources among the individuals in that group. Moreover, the bridging structures with access to different kind of resources, with more utility, may increase the benefits and hence the social capital [90, 127].

Contrary to the general intuition, close communities where bonding social capital high, do not perform well for public safety and security. Which is in fact parallel to the Bogomolov *et al.* and Forrest and Kearns findings; the more diverse the area the more secure [28, 72]. A similar study to understand empirically social capital and security has been performed by Buonanno *et al.*, in their work they found a negative correlation between blood donation and crime related features [33]. If we take blood donation as a civic engagement, which can be treated as a similar indicator for democratic participation. We could see that bridging social capital brings more democratic participation, which is in parallel with the current literature [99].

4.4 Socio-Economic indicators

In this section, first we give some background information regarding the city, then we describe in detail the socio-economic indicators we correlate with the communication and mobility network characteristics of the geographical areas under investigation. We focus not only on indicators of economic well-being, but also on proxies for civic engagement and democratic participation, and public safety and security.

4.4.1 Abidjan: Economic capital of Cote d'Ivoire

Abidjan is the biggest city in Côte d'Ivoire, with 20% of the country's total population. It is composed of 10 communes, as shown in Fig 4.2 and in

Table 4.1. Given its population and economic power, it is the region in the country with the densest levels of mobile cell towers (antennae) with 378 out of 1,238.

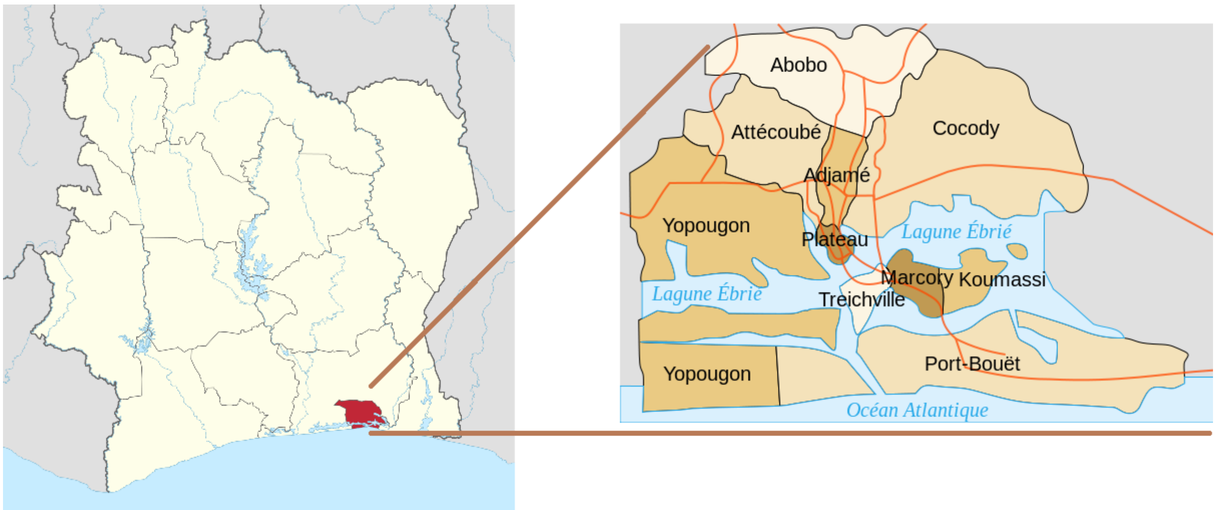

Figure 4.2: Left: Côte d'Ivoire's economic capital Abidjan shown in red. Right: 10 communes of Abidjan.

Even though Abidjan is the economic capital of Côte d'Ivoire, it has the highest levels of income inequality compared to the rest of the country. Adjame, Marcory and Treichville are the city's poorest communes; Cocody is its wealthiest commune; Plateau is the administrative and commercial commune with skyscrapers; and Yopougon and Abobo are the most populous communes of the city, including some slum housing and communal areas [136].

To sum up, the city of Abidjan has several characteristics that make it a suitable environment to address our research questions. In particular, (1) it is the wealthiest and most populated city in Côte d'Ivoire; (2) it has high levels of mobile phone adoption; (3) there are relevant demographic indicators available for each of its 10 communes; and (4) the disparity in socio-economic indicators across districts is large [136] (in fact, its slums have developed under the shadows of its skyscrapers [155]).

As the first step in our analysis, we compute six network variables that are related to social capital both for the communication and mobility networks and for each of the 10 communes in the city of Abidjan, as described in Section 4.6.

Table 4.1: Abidjan's communes population, area coverage in Km^2 and number of cell towers in each commune, from 2014 census.

	Commune	Population	Area in (Km^2)	Number of cell towers
1	Abobo	1.030.658	112.7	45
2	Adjame	372.978	12.1	23
3	Attecoube	260.911	38.6	14
4	Cocody	447.055	76.1	91
5	Koumassi	433.139	4	17
6	Marcory	249.858	12.6	31
7	Plateau	7.488	8.9	35
8	Port-Bouet	419.033	60.5	25
9	Treichville	102.580	11.4	20
10	Yopougon	1.071.543	117	77

4.4.2 Economic well-being

Traditionally, different metrics have been used to capture the economic development of a country, such as the Gross Domestic Product (GDP) and the Gross National Income (GNI). However, when we restrict the analysis to a district (commune) level, such indicators are neither well defined nor accessible for the 10 communes of the city of Abidjan. Therefore, we rely on two alternative datasets: (i) the land use master plan of Abidjan [107] and (ii) the budgets for the 10 communes from 1990 [62], as shown in

C.1 [62]. We use this data to obtain suitable indicators for economic development, and we study the association between these indicators with the social capital metrics described in the following sections.

Previous qualitative studies have suggested that: (1) Adjame, Marcory, and Treichville can be regarded as poor areas [135, 136]; (2) Cocody can be classified as a rich, residential commune, as reflected by its large percentage of expatriates and consulates located within its boundaries [186]; (3) Abobo is a residential low-class area [186]; (4) Adjame is a commercial place with markets and a bus station [186]; and (5) Plateau is the business district.

Moreover, data on land usage provides relevant information regarding the economic status of communes, particularly given that some of the Abidjan's communes have slums. We obtained land use data from the urban master plan of Abidjan published in September 2014 [107]. The plan was constructed through the collaboration between the Japan International Cooperation Agency (JICA) and the Ministry of construction, housing, sanitation and urban development (MCLAU) [1]. Data on land usage include several categories, including *informal settlement*, *commercial/office land*, *educational land*, and *security offices*. In particular, we used the *informal settlement percentage* of each commune to assess its economic condition, and the *security offices* land coverage to assess its security. An additional data source for land use was obtained from the thesis of Quonan Christian Yao-Kouassi on the spatial distribution of housing settlements in Abidjan, shown in C.2 [221]. The type of settlements are changing from prospectus unshared houses, to slums without covered floors.

[1]http://open_jicareport.jica.go.jp/pdf/12230603$_0$1.pdf

4.4.3 Civic engagement and democratic participation

The norms and networks of civic engagement play a crucial role in performance of the government, according to Putnam [164, 165]. He associates democratic participation with dense networks, which corresponds to social cohesion, or less structural holes. There are various studies to measure democratic participation through social networks, but not much for the mobile phone usage. The one, which analyzed bridging and bonding effect on democratic participation through surveys of land and mobile phone users [99]. In that work, Hampton argued that the bridging or highly connected nodes, have a significant impact on the democratic participation. Indeed, this is in parallel with our findings.

People's democratic participation is significantly influenced by the underlying network of social interactions connecting them [99]. We used the number of voters and the percentage of voters out of all registered citizens to investigate associations between democratic participation and various types of the underlying social structures (e.g., bonding and bridging structures). Specifically, we used data obtained from the 2011 presidential election, as shown in C.3. The percentage of participation varies significantly across communes, ranging from 15.7% in Cocody to 36.8% in Treichville.

4.4.4 Public health

Abidjan has a public health service in the city which provides access to health in its communes. However, malaria and HIV are still threatening the city, particularly areas with poor sanitation. With respect to social capital and its relation to public health, we obtain estimated HIV prevalence, and malaria cases for each commune in the city.

We obtained the number of people living with HIV (PLHIV) from a report by The U.S. Presidents Emergency Plan for AIDS Relief (PEP-

FAR) [200], in the first column of Table C.4 and malaria cases in the second. Wang *et al.* report the number of confirmed malaria cases from 2001 data obtained by *Programme National de Lutte Contre le Paludisme. CHU: Centre Hospitalier Universitaire* [210], shown in Table C.4.

In the literature, social capital is related with well-being of the society from different perspectives, such as life expectancy and mortality rates [115], obesity [110, 112], mental health [148]. In these researches social capital is modeled with "trust". Our work is different as we model social capital in means of network structure. So the hypothesis we introduce is the importance of information flow rather than building trust to evaluate social capital. Such as World Health Organization (WHO) organizes campaign against sexually transmitted diseases in Africa, in order to prevent the people from HIV. The bridging social capital lets the communities to access information.

Countries like United States, with high incomes have lower life expectancy rate, comparing to developing countries like Sri Lanka, Jamaica [115]. Kawachi *et al.* compared body mass index (BMI) and neighborhood to empirically analyze social capital [110, 112]. Thus, this approach may be arguable which can be the outcome of another phenomena: *homophily* [144]. The society may take care about your health, however social capital definition underlies the importance of structure of neighbors. So the hypothesis we introduce is the importance of information flow rather, such as World Health Organization (WHO) organizes campaign against sexually transmitted diseases in Africa, in order to prevent the people from HIV. High social capital lets the communities to access this type of information, the risk would be less. Apart from that, unexpected positive association between mental distress and bonding capital has been observed in [148].

4.4.5 Public safety and security

To measure public safety and security, we used the number of stolen cars from 2009 and 2010, obtained from the Statistical Department of Côte d'Ivoire [12]. The data is shown in Table C.5. In addition to these records, the land usage of security related buildings in the commune is also analyzed, as mentioned before.

4.5 Characterizing social capital from mobility and communication networks

From the mobile dataset, we constructed two different networks: (i) a *mobility* network, denoted by G_m, and constructed by analyzing consecutive calls by the same user from Subset-2 in our dataset; and (ii) a *communication* network, denoted by G_c, obtained from traffic from cell tower to cell tower available in Subset-1 of our dataset. In both networks, A is the set of all cell towers in the country $A = \{a_1, a_2, ..., a_n\}$, and U is the set of all unique users in our dataset, $U = \{u_1, u_2, ..., u_n\}$.

4.5.1 The mobility networks

The mobility network is constructed based on the location of the cell towers in consecutive calls by the same user. If user u makes a call from cell tower a_i and a subsequent call from cell tower a_j, where $i, j \in A$, and $i \neq j$, the path of user u is denoted as a pair $P_u(a_i, a_j)$, which corresponds to a movement from cell tower a_i to cell tower a_j. T represents the entire time period in days, which is 140 days in our case.

Thus, we constructed the mobility graph $G_m(A, P, W)$, where cell towers A are the nodes, user paths $P_u(a_i, a_j)$ are the edges connecting nodes a_i and a_j, and $\mathbf{W}$ are the weights of the edges, which correspond to the number of transitions between the two nodes connected by the edge. We computed $\mathbf{W}$ for all users $u \in U$, and then we aggregated the paths between the same nodes as shown in Equation 4.1.

$$W_{i,j} = W_{i,j} + 1, \text{if } P_u(a_i, a_j) > 0. \tag{4.1}$$

$\mathbf{W}$ is an asymmetric matrix whose order is the number of cell towers (A) in the dataset.

We constructed the mobility network using the data in Subset-2 with the geographic granularity given by the locations of cell towers. In this network, information is assumed to flow between connected pairs of nodes, and therefore between the locations to which the nodes are associated. Thus, as people move from one location to the another, they act as vehicles of information across space.

4.5.2 The communication networks

Since we do not have data on which users made or received calls, we modeled the information flow through the hourly aggregate records of call volumes in each initiating and receiving cell tower, which is available in Subset-1, defined in Appendix A.1.1.

We thus used the total number of phone calls that took place hourly between each pair of cell towers a_i and a_j in A. From this data, we constructed the call graph $G_c(A, P, W)$ where the weight of the link between each pair of nodes corresponds to the total number of phone calls that took place between them within an interval of one hour. Like the mobility graph, the communication graph is also directed.

4.5.3 The Abidjan networks

We constructed the Abidjan mobility and communication networks following the methodology described above. From Subsets-1 and -2, we selected the data corresponding to cell towers which are within the city boundaries. In order to capture communication and mobility patterns involving places outside the city, an arbitrary node 0 was created to represent phone calls taking place outside the city's boundaries.

The total number of nodes (cell towers) in the Abidjan networks is 378 (in addition to the arbitrary external node). The total number of links for communication and mobility networks are $689,244$ and $142,080$, respectively.

We computed network statistics for each cell tower, obtained six metrics characterizing the social capital originating from the mobility and communication networks, and then aggregated values at the Abidjan's commune level. In the next section we shall describe in detail these metrics, namely: (i) the degree centrality, (ii) the betweenness centrality, (iii) the effective size, (iv) the efficiency, (v) the node density, and (vi) the local clustering coefficient of each commune.

4.5.4 Temporal and spatial aggregation

Before computing the network metrics, we performed a temporal and spatial aggregation. More specifically, we first aggregated the data for the entire observation period (i.e., 140 days) and for each network. Next, we computed the six metrics for each cell tower and then we aggregated them spatially at the commune level, because the demographic and socio-economic indicators publicly available are only provided at this level of granularity. Notice that the number of cell towers in each commune is not constant, as shown in Table 4.1. In particular, we computed the mean, the standart deviation, the median, the mode, the min., and the max of each distribution for each metric. Descriptive statistics of the data and variables of interest are presented in Appendix C.1.

4.6 Network metrics to assess social capital

In this section we describe the six network metrics that we computed to investigate social capital. These six metrics are: (i) degree centrality, (iii) betweenness centrality, (iii) effective size, (iv) efficiency, (v) network density, and (vi) local clustering coefficient [29, 162, 137]. Notice that both the communication (G_c) and the mobility networks (G_m) are directed and weighted graphs. However, not all of our metrics leverage these characteristics. Indeed, some of our metrics (i.e., effective size and efficiency) use an undirected graph and one of our metrics (i.e., clustering coefficient) assumes an undirected and unweighted graph. In these latter cases, we transformed the directed graph into an undirected one by computing the average weight of each edge.

In what follows, we shall organize our discussion of the metrics into two main building blocks. We begin by introducing the simplest centrality measure (degree centrality) that we shall use to detect the value a node that can be extracted solely from the direct connections to its local neighbors. We shall then broaden our perspective also to account for connections between the first neighbors of the nodes or even for connections beyond first neighbors and local networks. To this end, we shall group the network metrics into two groups, depending on which type of social structure they are aimed to capture [122, 118].

The metrics included in the first group – betweenness, effective size, and efficiency – share the idea that a node can draw influence, power, importance from the absence of edges in the network [37, 90]. In this sense, they capture the degree to which nodes span structural cleavages, holes, discontinuities in the network, and can therefore act as potential intermediaries for others to interact and communicate with one another. This is indeed the idea of social capital as the bridging potential that nodes can tap from open structures rich in brokerage opportunities. The metrics in the second group, on the other hand, share the idea that nodes can draw value from the presence of edges in their local or global network [49]. Specifically, they capture the degree to which a node's contacts are also connected with each other, and therefore the triads centred on the node are closed into triangles. These metrics can thus be used to test hypotheses on social capital rooted in the value of social cohesion and in the bonding potential that nodes can tap from closed structures rich in third-party relationships.

4.6.1 Degree centrality

In graph theory, the degree centrality of a node refers to the number of connections incident upon the node [71, 154]. In the context of the communication network, a connection is established whenever a call takes place from one cell tower to another. In the mobility network, a connection between any two locations is established whenever the same mobile user has made consecutive phone calls connected to the cell towers of the two locations. In our current work, we established a connection between locations if either an inbound or an outbound link between the corresponding cell towers was created. Thus reciprocity is not accounted for.

It has been widely suggested that degree centrality quantifies the potential amount of information a node may have access to [71]. Thus, centrality can be used to measure the amount of resource an actor can mobilize through her or his direct contacts, and in this sense it can be seen as a measure of the value that can be extracted from one's local network. From this perspective, for instance, in Borgatti, Jones and Everett's work [29] degree centrality has been positively related with social capital. However, degree in itself is not sufficient for capturing the potential benefit a node can derive from the underlying network. Indeed, degree alone does not reflect the value an actor can extract beyond her or his direct connections, nor the bridging or bonding potential of the node's ego-centred network.

4.6.2 Network metrics for open structures

This section will outline three network metrics that aim to capture brokerage opportunities and thus the value that actors can extract from the bridging potential of the underlying network. The first of these metrics – betweenness centrality – is a global centrality measure that accounts for edges lying across the whole network. The other two measures – effective size and efficiency – are local measures that only reflect the topology of the node's ego-centred network, that is the node's direct connections to its first neighbors and the connections between these neighbors.

Betweenness centrality

Betweenness centrality measures the extent to which a node lies on the paths that connect other nodes [73, 154, 71]. Nodes with large betweenness centrality may gain considerable influence within a network by virtue of their control over information flowing to and from other nodes. These nodes are also the ones whose removal from the network will most disrupt communication between other nodes because they lie on the largest number of paths along which messages travel [29]. From this perspective, it can be argued that betweenness centrality quantifies the bridging potential of a node that leverages on the lack of edges connecting other nodes. Thus, betweenness can be used to test hypotheses on the value that can be extracted from open social structures providing nodes with brokerage and bridging opportunities.

Effective size

The effective size of node is ego-centred network aims to capture the degree to which the node's first neighbors are redundant. Formally, it can be expressed as a function of the following two matrices: the transition matrix **P**, and the marginal strength matrix **M**. The element p_{il} of matrix **P** measure the proportion of is network time and energy invested in the relationship with node l, as defined in [37, 122]:

$$p_{il} = \frac{w_{il} + w_{li}}{\sum_m (w_{im} + w_{mi})}, \tag{4.2}$$

where $w_{il} + w_{li}$ is the sum of the weights of the two links connecting i to l, while $\sum_m (w_{im} + w_{mi})$ is the total strength of node i. This is the sum of node i's out-strength, $s_i^{out} = \sum_m w_{im}$, and in-strength, $s_i^{in} = \sum_m w_{mi}$, i.e., the sum of the weights of all the incoming and outgoing links incident upon node i. Notice that by definition $0 \leq p_{il} \leq 1$ for all i, l with $p_{il} = 0$ if there is neither a link from node i to node j, nor a link from node j to node i. The transition matrix **P** is row-stochastic: $\sum_j P_{il} = 1$.

The entry m_{jl} of the marginal strength matrix **M** is defined as:

$$m_{jl} = \frac{w_{jl} + w_{lj}}{\max_m (w_{jm} + w_{mj})}. \tag{4.3}$$

Notice that $0 \leq m_{jl} \leq 1$ for all j, l with $m_{jl} = 0$ if there is neither a link from node j to node l nor from l j. Finally, the two matrices **P** and **M** defined above are asymmetric.

Thus, the effective size S_i of node i's ego-centred network can be formalized as follows:

$$S_i = \sum_{j \in \mathcal{N}_i} \left[1 - \sum_l p_{il} m_{jl} \right]. \tag{4.4}$$

where $\mathcal{N}_i$ is the set of node i's neighbors.

Effective size was originally introduced by Burt to quantify the degree to which node i's ego-centred network is rich in structural holes, and thus provides the node with bridging potential and brokerage opportunities [36].

Efficiency

Efficiency can be defined as the normalized value of effective size, namely the ratio between node i's effective size and degree. Formally,

$$\text{Efficiency}_i = \frac{S_i}{\text{Degree}_i}. \tag{4.5}$$

Like effective size, efficiency quantifies the bridging potential of a node. As it is normalized by the degree, the measures takes value within 0 and 1. The closer it is to 1, the greater the bridging potential of i's ego-centred network. That is, a unitary value means that each of node i's neighbors is non-redundant, whereas a value close to *zero* would imply that all of i's first neighbor are redundant, and are thus unable to provide unique and non-overlapping information or resources.

4.6.3 Network metrics for closed structures

In this section, we shall outline two metrics for capturing the social capital that originates from closed structures. The first metric – density – is a global network-based volume measure that aims to detect the number of edges in the whole network. It directly captures the degree of closure of the whole network, and indirectly is sensitive to the closure of each node's ego-centred network. The second measure – the local clustering coefficient – is a node-level local measure of social cohesion that captures how closed the node's ego-centred network is.

Network density

Network density is a volume measure, traditionally defined as the ratio between the number of actual links in the network and the maximum number of possible links [71]. Formally, density is defined by Equation 4.6. Typically, given a fixed number of nodes (n) in the network, higher density contributes towards enhancing social cohesion, as it fosters the creation of tightly knit and closed structures, rich in third-party relationships [122]. A large body of literature on social capital has regarded dense networks as wellsprings of social capital, as a result of the benefits that interacting actors can derive from the enforcement of social norms and control, reduced risks of free-riding, and better opportunities for the exchange of tacit, complex and proprietary information.

$$\begin{aligned} \text{All possible connections} &= \frac{n(n-1)}{2} \\ \text{Network density} &= \frac{\text{Actual connections}}{\text{All possible connections}} \end{aligned} \tag{4.6}$$

Local clustering coefficient

In order to capture the social cohesion of node i's ego network, we now focus on i's local clustering coefficient, (C_i), which measures the probability that a pair of node i's friends (first neighbors) also befriend each other [154, 212]. Formally, the local clustering coefficient can be formalized as

$$C_i = \begin{cases} \frac{E[G_i]}{\frac{k_i(k_i-1)}{2}} & \text{for } k_i \geq 2 \\ 0 & \text{for } k_i = 0, 1 \end{cases} \tag{4.7}$$

where $E[G_i]$ is the number of edges included in the subgraph G_i induced by the set $\mathcal{N}_i$ of node i's first neighbors (i.e., $E[G_i]$ is the number of edges linking i's neighbors), and k_i is the degree of node i. Notice that C_i is normalized such that it varies between 0 and 1, $0 \leq C_i \leq 1$. In particular, it takes the value of 0 when none of i's neighbors are connected with each other, and takes the value of 1 when all pairs of i's neighbors are connected. So conceived of, the local clustering coefficient thus captures the degree to which the triads centred on i are closed into triangles, and thus can suitably measure the degree to which node i's ego-centred network is closed and rich in third-party relationships. In this sense, local clustering has been widely used in studies of social capital to investigate the benefits of closed structures [122].

Furthermore, the local clustering coefficient is formally related to Burt's redundancy measure, which in turn is closely related to ego network density. To formalize this relationship, we begin by showing how effective size can be expressed as the difference between the number of node i's first neighbors and i's ego-centred network redundancy. Formally,

$$S_i = \mathcal{N}_i - \frac{2E[G_i]}{\mathcal{N}_i}, \tag{4.8}$$

where $\frac{2E[G_i]}{\mathcal{N}_i}$ is the redundancy of node i's ego-centred network, measured as the average degree of i's neighbors. In turn, $\frac{2E[G_i]}{\mathcal{N}_i}$ is simply the ego-centred network density, scaled by a factor of $\mathcal{N}_i - 1$.

Moreover, it has been demonstrated that effective size is formally related to the local clustering coefficient as follows [122]:

$$S_i = k_i - (k_i - 1)C_i. \tag{4.9}$$

where $k_i = \mathcal{N}_i$. If we now plug Equation 4.8 into Equation 4.9, we obtain:

$$C_i = \frac{\frac{2E[G_i]}{\mathcal{N}_i}}{k_i - 1} = \frac{\text{Redundancy}_i}{k_i - 1} \tag{4.10}$$

Thus, the local clustering coefficient can be expressed in terms of local network redundancy. The more redundant a node's contacts are, the larger the proportion of triads centred on the node that close into triangles, and thus the higher the local clustering. Most importantly, local clustering and effective size vary with redundancy in opposing ways (see Equation 4.8 and Equation 4.10). In this sense, it has been argued that the former can be seen as a mirror image of the latter [122]. Indeed, while effective size is sensitive to the *absence* of edges between node i's alters, and thus captures the non-redundancy of edges in i's ego-centred network and its bridging potential, the local clustering coefficient is sensitive to the *presence* of edges in i's ego-centred network, and therefore captures the redundancy of edges in the network as well as its social cohesion and bonding potential. In this sense, the two measures, and thus the two local network topologies they aim to quantify, can be regarded as the two faces of the same coin [122].

4.7 Results

In this section we present the results obtained by computing bivariate correlations between social capital network metrics and a number of district-level socio-economic indicators broadly related to (i) economic well-being, (ii) civic engagement and democratic participation, and (iii) public safety and security. These results are obtained by computing Pearson's r, Spearman's ρ, and Kendall's τ correlation coefficients. More precisely, we have computed the mean, the mode, and the median of each node's network metrics in order to have single values for each district to correlate with the socio-economic indicators of the region. While in this aggregation process we loose some information, we can still see different behaviors among poor and wealthy regions.

4.7.1 Economic well-being

Interestingly, we have found negative correlations between open bridging network metrics (i.e. effective size and efficiency) and indicators of economic wealth, while positive correlations were found with indicators of economic deprivation. For example, the *effective size*, computed from the communication network, is negatively associated with the revenue budget per person (Pearson's $r = -0.9145$, $p = 0.00021$) and the capital budget per person (Pearson's $r = -0.90812$, $p = 0.00027$). Moreover, the communication network-based efficiency is negatively associated with the revenue budget per person (Pearson's $r = -0.8079$, $p = 0.00468$) and with the capital budget per person (Pearson's $r = -0.8133$, $p = 0.0042$), and the mobility network-based efficiency is negatively associated with the revenue budget per person (Pearson's $r = -0.7973$, $p = 0.0057$) and with the capital budget per person (Pearson's $r = -0.8148$, $p = 0.00408$). Thus, it seems that the found associations tend to be valid both for metrics obtained from the communication network and for the ones computed from the mobility network. In Table 4.4 we summarize the significant results ($p < 0.05$).

Table 4.2: Communication networks' average degree, density, effective size, efficiency, local clustering coefficient, and betweenness centrality. The communication network measures have been calculated for each of the ten communes of Abidjan.

		Open	Structure		Close	Structure
	Degree	Between. Centrality	Eff. size	Efficiency	Network density	Local clust.coeff.
Abobo	1200.200	1128.866	624.555	0.521	2.96E-05	0.946
Adjame	1195.565	1123.036	609.005	0.509	7.74E-06	0.946
Attecoube	1202.538	1147.459	619.986	0.516	2.47E-06	0.945
Cocody	1141.789	1054.691	560.872	0.476	1.14E-04	0.948
Koumassi	1188.118	1125.058	654.206	0.551	4.23E-06	0.947
Marcory	1025.103	841.657	573.949	0.552	1.17E-05	0.958
Plateau	1001.588	789.809	427.779	0.394	1.61E-05	0.961
Port-Bouet	1067.750	896.124	607.666	0.568	8.42E-06	0.956
Treichville	1113.850	950.254	612.798	0.550	5.85E-06	0.954
Yopougon	1175.697	1143.794	599.460	0.502	8.22E-05	0.946

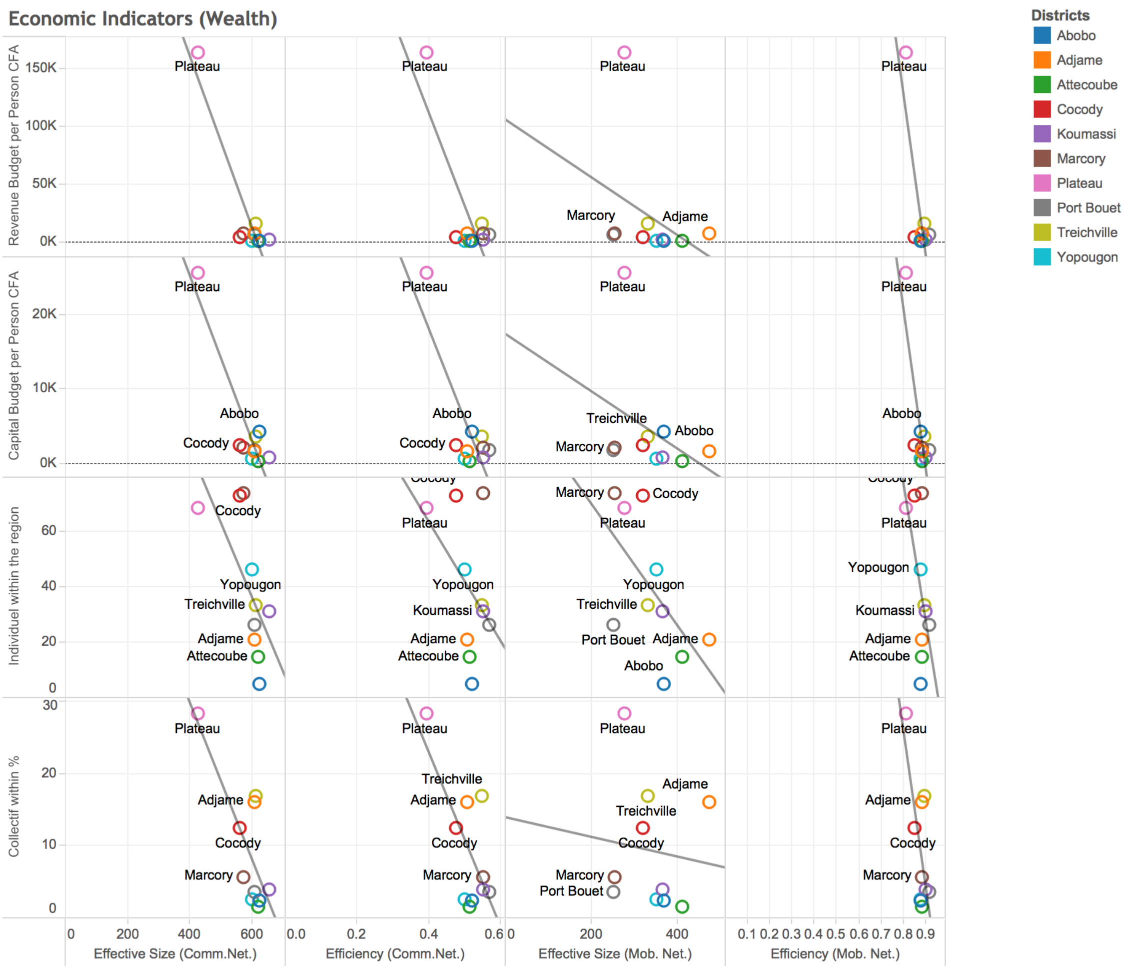

Figure 4.3: Effective size and efficiency for economical indicators showing wealth. Effective size and efficiency for three economic indicators: budget per person, individual residents, collective residents percentage.

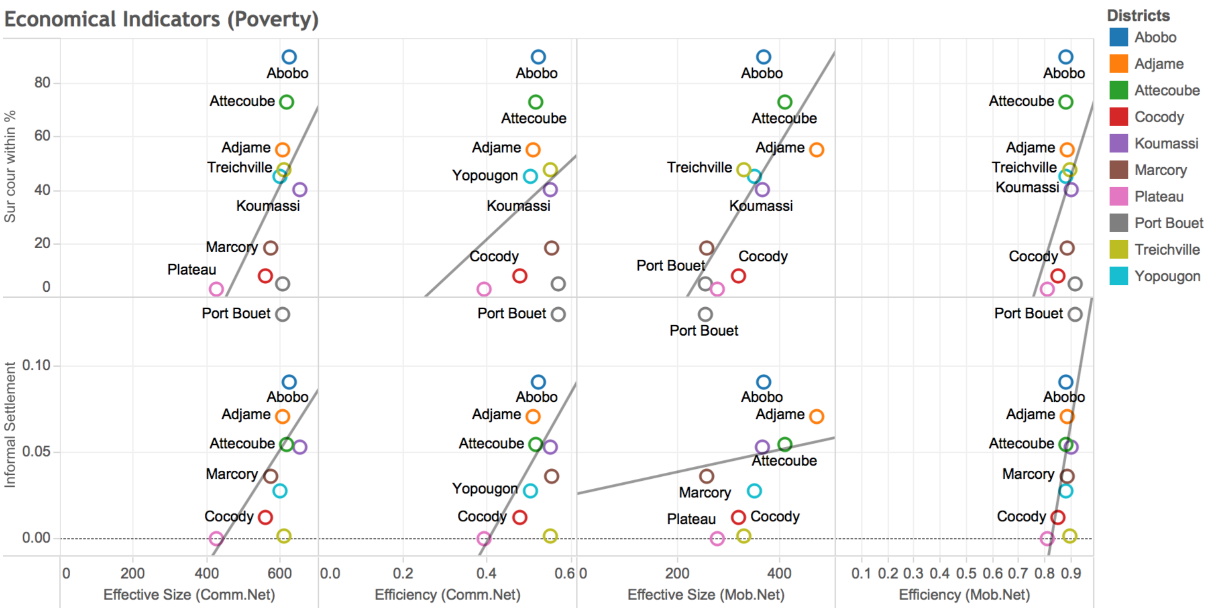

Figure 4.4: Effective size and efficiency for economical indicators showing poverty. Effective size and efficiency for three economic indicators: Informal settlement, sur cor percentage in the district.

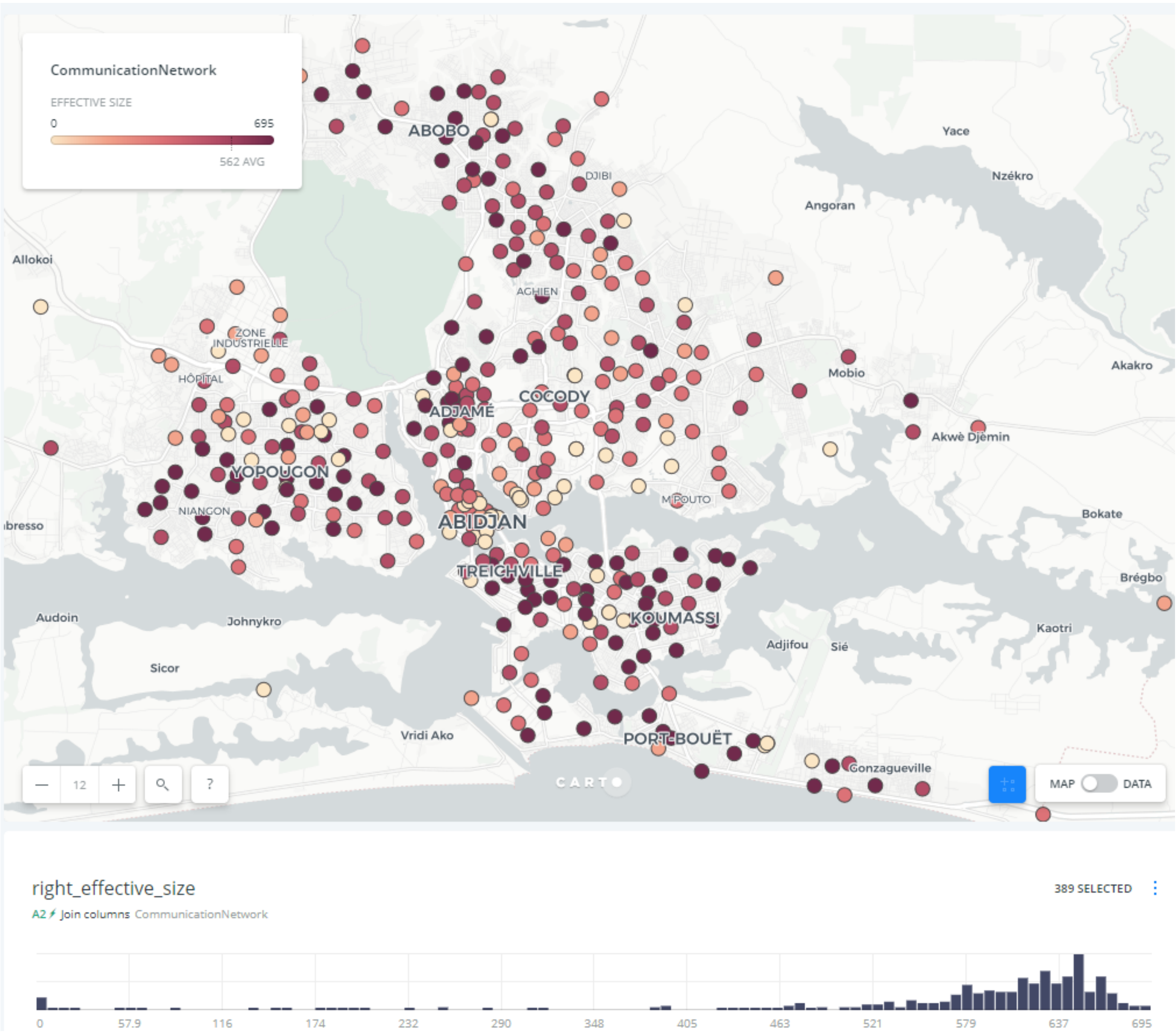

Figure 4.5: Abidjan effective size from communication network.

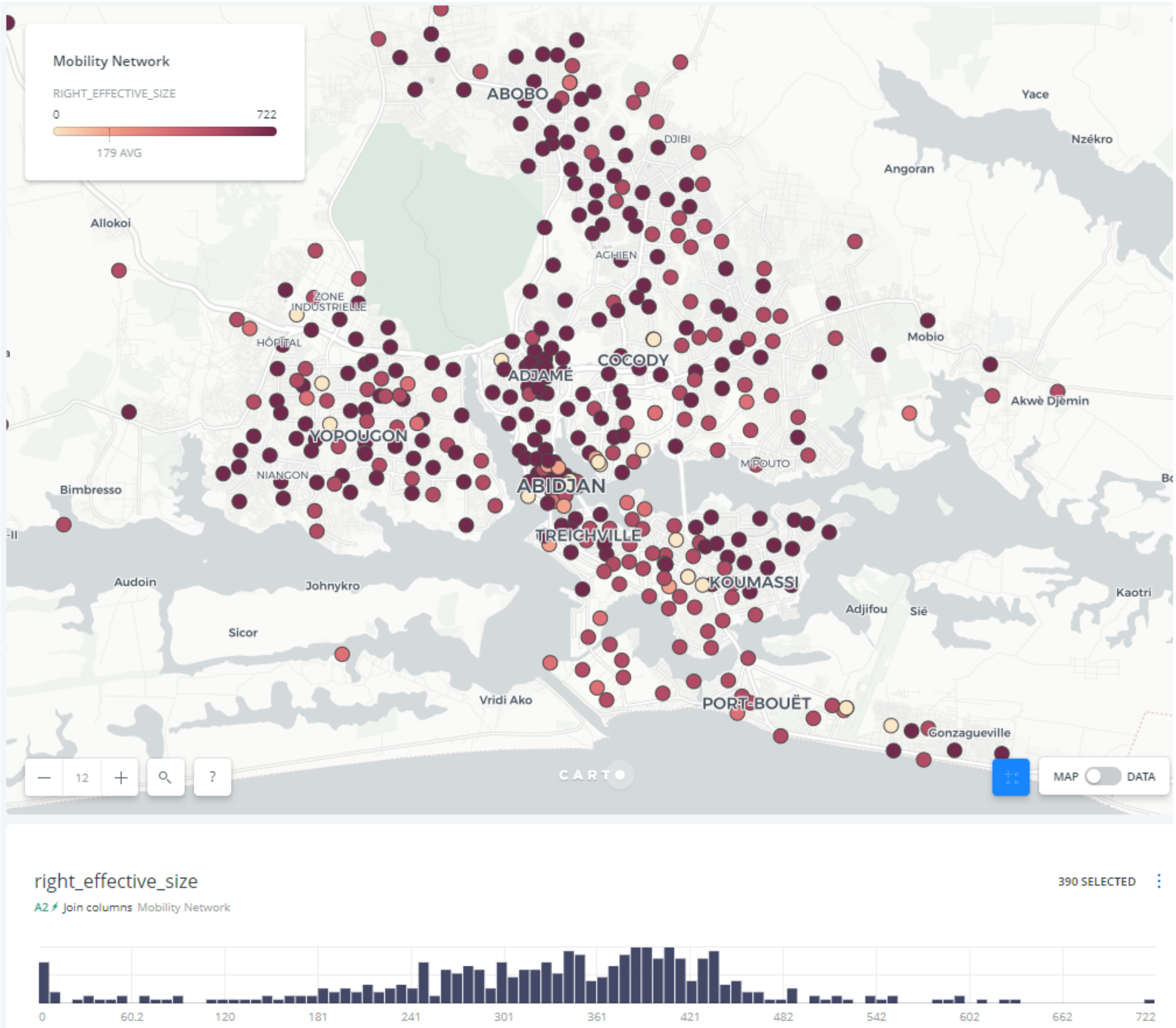

Figure 4.6: Abidjan effective size from mobility network.

Table 4.3: Mobility networks' average degree, density, effective size, efficiency, local clustering coefficient, and betweenness centrality. The mobility network measures have been calculated for each of the ten communes of Abidjan.

		Open Structure			Closed Structure	
	Degree	Between. centrality	Eff. size	Efficiency	Network density	Local clust.coeff.
Abobo	417.489	2642.071	369.086	0.881	4.84E-04	0.666
Adjame	533.174	7693.997	474.576	0.884	1.25E-04	0.555
Attecoube	466.308	4076.686	412.843	0.882	3.86E-05	0.626
Cocody	370.258	904.283	319.693	0.850	1.75E-03	0.756
Koumassi	406.353	1431.543	366.306	0.901	6.73E-05	0.690
Marcory	285.690	497.009	256.203	0.885	1.76E-04	0.805
Plateau	324.176	845.653	276.827	0.811	2.04E-04	0.776
Port-Bouet	276.000	861.109	253.292	0.916	1.19E-04	0.772
Treichville	369.250	1067.158	331.506	0.894	8.98E-05	0.737
Yopougon	394.880	2021.052	350.846	0.881	1.27E-03	0.688

Table 4.4: Summary results on economic well-being indicators based on both communication and mobility networks. Only statistically significant results are reported with *pvalue* < 0.05. Note: '*': $p < 0.05$; '**': $p < 0.01$; '***': $p < 0.001$.

	Revenue budget per person		Capital budget per person		Percentage of land covered by unshared houses		Percentage of land covered by shared houses		Percentage of land covered by slums		Informal settlement land use	
	Comm.	Mob.	Comm.	Mob.	Comm.	Mob.	Comm.	Mob.	Comm.	Mob.	Comm.	Mob.
Degree	-0.64*				-0.68*				0.78**	0.72**		
Open structures (bridging soc.cap.)												
Betweenness centrality	-0.63*								0.73*			
Effective size	-0.91***		-0.91***		-0.66*		-0.76**			0.73**		
Efficiency	-0.81**	-0.79**	-0.81**	-0.81**			-0.69*	-0.71*				0.63*
Closed structures (bonding soc.cap.)												
Local clustering coeff	0.64*		0.62*			0.7*			-0.72*	-0.75**		

4.7.2 Civic engagement and democratic participation

The political scientist Putnam associates high levels of democratic participation and civic engagement with dense networks, which corresponds to high level of social cohesion and few structural holes [164, 165]. Instead, our findings show an opposite trend: in particular, we have found a positive correlation between *effective size* and *democratic participation*, both in the communication (Pearson's $r = 0.66$, $p = 0.00021$) and mobility (Spearman's $\rho = 0.66$, $p = 0.0375$) networks.

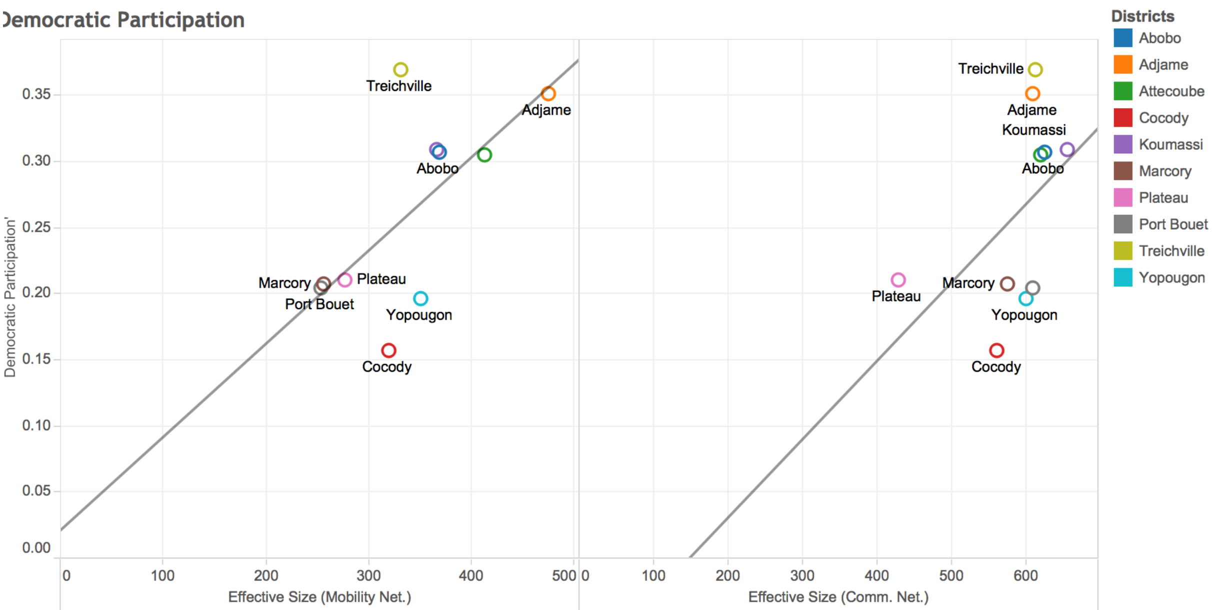

Figure 4.7: Effective size versus democratic participation for mobile and communication networks.

4.7.3 Public safety and security

Our findings show that more connected (i.e. high degree centrality), and more diverse and bridging districts (high betweeness centrality, effective size, and efficiency) have a lacking presence of police stations and other

Table 4.5: Summary of results for civic engagement and democratic participation, based on both the communication and mobility networks. Only significant results are reported, with ($p < 0.05$). '*' is shown for ($p < 0.05$), '**' is shown for ($p < 0.01$), '***' is shown for ($p < 0.001$). By default Pearson's correlation coefficients are shown, (S) stands for Spearman's ρ.

Democratic participation			
		Communication	Mobility
Bridging soc. cap.	Effective size	0.66* (S)	0.66*
Bonding soc. cap.	Network density	-0.6327*	-0.61*
	Local clustering coefficient		-0.64*

public safety and security offices. For example, degree centrality is negatively associated with the proportion of land covered by security offices (Pearson's $r = -0.71055$, $p = 0.02127$). Similarly, effective size and betweeness centrality are inversely correlated with the proportion of land covered by security offices (Pearson's $r = -0.60287$, $p = 0.06505$ and Pearson's $r = -0.6760$, $p = 0.0318$, respectively). Instead, districts showing higher social cohesion (i.e. high levels of local clustering coefficient) have more security offices (Pearson's $r = 0.63182$, $p = 0.050$). For a summary of the obtained results see Table 4.6.

Interestingly, our results also show that bridging districts (i.e. high effective size) experience an lower numbers of stolen cars (Kendall's $\tau = -0.50128041$, $p = 0.0436314$). However, we observe a statistically significant result only for the communication network. Indeed, the mean values of the bridging and bonding characteristics of the mobility network do not correlate, in a statistically significant way, with the number of stolen cars.

Focusing our attention on the the mode values of the efficiency (an example of bridging network characteristics), we find a negative association with the number of stolen cars in 2009 (Pearson's $r = -0.62994$, $p = 0.0.05093$) and in 2010 (Pearson's $r = -0.71570$, $p = 0.0199$) for mobility

Table 4.6: Summary of results for public safety and security based on both the communication and mobility networks. Only significant results are reported, with ($p < 0.05$). '*' is shown for ($p < 0.05$), '**' is shown for ($p < 0.01$), '***' is shown for ($p < 0.001$). By default Pearson correlation coefficients are shown; (K) stands for Kendall τ.

Security				
	Security Land coverage		Stolen cars 2009	
	Comm.	Mob.	Comm.	Mob.
Degree	-0.71*	-0.63*		
Open Structures (bridging soc. cap.)				
Betweenness centrality	-0.67*			
Effective Size	-0.60*	-0.66*	-0.5* (K)	
Closed structures (bonding soc. cap.)				
Network density			0.722*	0.727*
Local clust. coeff.	0.647*	0.7*		

networks.

4.8 Discussion

The measurement of social capital is a challenging task, as both social interactions and the value that such interactions bring to the society are hard to model and quantify. In general, there is a lack of ground truth data about such interactions. In this paper, we have proposed to use aggregated mobile phone usage data to overcome this lack of information on social interactions.

Moreover, the availability of high-quality, high-granularity socio-economic data is also a challenge, particularly in developing countries. For example, the communication and mobility networks that we have constructed from the mobile phone usage data have 1238 nodes, while the publicly available socio-economic indicators are only for 10 districts of the city of Abidjan.

Nonetheless, our findings shed a novel light on the relationship between bridging and bonding social network metrics and a diverse set of socio-economic indicators of a developing country, providing some surprising results.

Our findings shed light on the relationship between social capital and socio-economic indicators with some surprising results.

Economic well-being We find a negative correlation between bridging social capital metrics and wealth indicators (*e.g.*, wealthy residential areas) in both the communication and mobility networks, and a positive correlation between bridging social capital metrics and poverty indicators. Thus, our results do not seem to corroborate Burt's theories on the role played by the structural holes [37]. In his seminal work, the sociologist Burt showed that the individual's economic remuneration within a company increase with the number of structural holes surrounding the node (*i.e.* the subject) [37]. Indeed, an individual who acts as a bridge between two or more closely connected groups of people could gain significant comparative advantages; for example his/her position allow him/her to transfer valuable information from one group to another, and to combine a more diverse set of ideas [37]. Along this line, Eagle *et al.* have shown that the diversity of individuals' social relationships is also strongly correlated with the economic development of their communities. In particular, their results have found that socio-economic opportunities in a given community in United Kingdom (UK) increase with the number of structural holes in the ego networks of the members.

Our results, indeed, seem more in line with the observations done by the sociologist Coleman on the role played for economic well-being by the *community closure* [49]. In particular, Coleman used the term 'closure' to refer to mutual knowledge and social ties between community members who support each other and sanction deviant behaviors. In a recent work, Lotero *et al.* have conducted an extensive survey among $56,513$ citizens of Medellin, Colombia, about their mobility behaviors and their socio-economic characteristics, such as age, gender, job, etc. [128]. Interestingly, they have found that mobility routines and travels of rich people are highly redundant and localized, while the opposite trend characterizes the mobility routines of poorer citizens [128]. This means that the social interactions and mobility routines of the economically disadvantaged people are more diverse than the ones of rich people.

A possible explanation may be related to the *Maslow's hierarchy of needs* [138]. Since the more basic needs, such as the biological and physiological needs, have not been covered yet, the behaviors may be ruled by fulfilling these needs. Once these needs have been met, the search for *safety* is the next ruling stimuli. In developing countries, most of the communities are at the very bottom of the Maslov's pyramid; hence a larger number of structural holes could be indicative of the efforts, done by their members, in order to achieve more prosperity. On the opposite, the members of richer communities in developing countries are driven by the search for safety.

Taken together, our findings and Eagle *et al.* findings [64] provide evidence that there may not be one universal pattern regarding the relationship between social capital and economic well-being but at least two patterns: one for developed countries and one for developing ones.

Civic engagement and democratic participation According to Putnam's theories [163, 164, 162], closed and cohesive communities, namely communities with high levels of bonding network metrics, are expected to show an higher democratic participation and civic engagement. However, in our analyses we find evidence of the opposite phenomenon as the denser the community *i.e.*, the lower is the democratic participation. In addition, our result show that communities with high values of bridging social capital have higher levels of democratic participation, a finding in line with some previous works in social science [134, 198, 99]. Indeed, Magee [134] found a significant correlation between diversity and voluntary participation and neighborhood social support, and Tindall and Cormier [198] found that network diversity predicted participation in voluntary organizations as well as extent of political ties.

Interestingly, the areas with higher levels of democratic participation tend to be economically deprived districts of Abidjan. One potential explanation for this finding is that poor communities may have a larger incentive to participate in the democratic process with the aspiration to instill changes that would lead them to improve their quality of life [50]. However, we cannot conclude, from our correlational analysis, if being able to access different types of information (i.e., having high levels of bridging capital), their socio-economic status or some other hidden factors is what motivates that civic involvement.

Public safety and security Our results provide evidence that the more diverse is the network of interactions of a district, the more secure the district is (i.e. less stolen cars). Thus, our findings support Jacobss theory of *natural surveillance* that high diversity of functions in an area and high diversity of people (gender diversity and age diversity) act as 'eyes on the street' decreasing the number of crimes [105, 72]. Similar results were found recently by several crime predictions studies [28, 204, 27].

Another interesting finding is that security offices are more likely to be built in communities with high level of bonding social capital, which are the ones were less effective is the role played by the *natural surveillance*. Note that we did not carry out a causality analysis. Hence, we cannot make any inferences regarding the causes behind higher safety in neighborhoods with higher network diversity.

Public Health Positive effect of raised consciousness among the societies against health related problems are investigated through number of people living with HIV and malaria cases declared to health units. On the contrary to the prior expectations, we could not find any significant relation between health and information flow for these two networks. Again for the heterogeneity of the districts, in means of number of non-Ivorian, does not reveal any significance either bridging nor bonding features.

4.9 Conclusion

We investigate the value that large scale aggregated mobile phone usage data has to infer variables related to bridging and bonding social capital in the case of Abidjan, the biggest city of Côte d'Ivoire. From the mobile data, we build two networks that characterize the interactions and information flows in the city. Next, we investigate the relationship between the inferred social capital metrics and socio-economic indicators. Some of our findings support existing theories and previous work [28, 99] whereas other findings do not corroborate the literature [164].

Given our findings and the fact that we have only analyzed one city in a developing economy, we believe that further research is needed to better assess the generalization power of our findings and shed light on the relationship between bonding and bridging social capital and socio-economic indicators. We plan to continue our research using large scale mobile data to characterize social capital to other regions.

Our work complements previous work on this topic by analyzing the impact of bridging network structures –inferred from mobile phone usage data– in a large city within a developing economy. We build two different networks from this data: a communication network and a mobility network. Next and following Burt's work on social capital [37], we compute six metrics for each of these networks, including weak and strong ties, described in Section 4.6. Finally, we correlate the six network variables with several demographic indicators, including socio-economic status.

Starting from the complexity as those authors mentioned, our work focused on how social ties between regions relate to the economical status. Following the work of Burt on social capital, the weak and strong ties can be calculated also on mobile phone dataset to exploit the social brokerage [35]. There are different metrics in the literature to uncover complex networks, Latora *et al.* in their study compare those metrics and proposed effective size (efficiency) [122].

Chapter 5

Residential Segregation or Segregation in Mobility

Everything is related to everything else, but near things are more related than distant things.
Geographer Waldo R. Tobler

Segregation, in the context of sociology, is the exclusion of some of the identities from the rest of the society for being different. The difference may be coming from age, religion, race, physical conditions etc. The majority may be represented as students in a school, employees in an office, or as well as the people in the cities [181]. In urban life in general, segregation corresponds to residential segregation, which includes separation of two or more social groups within a specified geographic area, such as a municipality, a county, or a metropolitan area [197].

The social distances between majority and minority groups, associated residential segregation with spatial inequalities [139], however Schelling describes the causes of segregation as a complex mechanism, where all the causes need to be diminished for complete integration, neither of them drive this phenomena alone. He grouped the causes as: (i) individually motivated, (ii) collectively enforced, and (iii) economically induced [181]. Another important contribution of Schelling is showing segregation as a function of preference among the neighborhood, which change over time. Granovetter generalized the work of Schelling for collective behavior, which covers not only for segregation, but all collective decisions given by society, such as voting, information diffusion etc. The *equilibrium state*, which corresponds to the benefit/cost threshold value, has to be reached for collective behavior in order to be effective [89]. Maybe the solution for integration lies beneath promoting benefit rather than the perception of the cost among society, which in general being raised by seeing the other as a threat.

The scientists analyze the residential segregation phenomena and the roots mainly in the census data [103]. Censuses are very powerful to have an opinion of the society, as American historian Frederic Jackson Turner's famous statement on his thesis: census with the combination with other source of data becomes "*lenses through which we form images of our society*" [205]. However as we mention in Section 5.2.1, some drawbacks resides in it, the most important for our work is being static, not including the daily life of interactions of the immigrants. Thus, to solve the urban challenges, we need further tools rather than static census data [124]. Since the availability of new forms of data, like CDRs we could observe things that we could not before.

In this perspective we use the country code of the calls as an indicator for the foreigners' location in the city. Here, we use data collected from the Telecom Italia Mobile (TIM) phone network in order to analyze two months of spatio-temporal activity patterns of migrants in six major Italian cities: Bari, Milan, Naples, Palermo, Rome and Turin. We exclude Venice because of its touristic dynamic. The details of the dataset is given in Appendix A.1.2.

In this work, we would like to answer the answers the below questions:

Static versus dynamic segregation

- Q1. The level of segregation from two different data sources resembles each other or not?

- Q2. Does greater urban diversity reduce segregation, do bigger cities perform better for having less segregation [103]? Two contradictory claims exist one from Schelling's work having smaller segregation for that question. In his work from 1969 *"the minority clusters become absolutely larger as the minority of itself becomes smaller"* [181]. The second is more recent work from Lamanna *et al.* claims, as big global metropolitan cities have been threatened by social spatial segregation [117].

Different nationalities' neighborhood preference motives within the city We elaborate Turin in the first phase of our study and leave the detailed analysis (Q3 and Q4) as future work.

- Q3. Can we find some clues if there are some nations, independent from their economical status, if exist prefers being segregated?

- Q4. Is there a pattern between different cities?

A possible application from our work can help urban planners, municipalities and policy makers to create spaces in the city such as Trafalgar Square in London [223]. Thus, the methodological formulation can be achieved if we can describe the cases, perform better.

5.1 History of ethnic and racial segregation

From the early 1900s until 1968 (Fair Housing Act of 1968), racial segregation or discrimination was legal and wide-spread in United States of America (USA) [141]. Even the majority and minority populations existed in many societies in the time being in different countries, the line between black and white people in America, was devastating. It corresponds to the period of historical African-American civil rights movement [143, 53]. Indeed, the pioneer articles for segregation mainly focused on two-groups — white and black — [63, 103, 181, 182, 34]. Several decades pass but the problem of segregation persist even on the digital platforms [67, 79, 120].

Ethnic and racial discrimination have been prohibited in USA for half a century ago, however in practice it is still an issue, not only for African-Americans but also for other ethnics groups and not only in United States, but also for Europe. Migrants expect to find improved life conditions in cities and therefore they move from poor to rich countries, from rural to urban places, and between cities [179, 180]. The massive influx of refugees are migrating to Europe, even putting their life in danger, almost every day. As a result of these migration waves, the urban segregation is inevitable in all metropolises.

Apart from integration related problems, this transition has been favored by several scholars [36, 94, 164], as cultural diversity brings prosperity to society, with bridging network structures we mentioned in Chapter 4. Thus without integration, some issues like (*seclusion, discrimination, xenophobia, etc.*) may be raised, from the increasing ethnic and cultural diversity of the population.

5.2 Evidence of segregation through census

The origin of the word *"census"* is coming from Latin, means "taxing". As well as the etymology of census, the written historical documents show that it arises from the need of taxation since ancient ages [213]. Census has become an important tool, in the following years, not only for Kings' treasure calculation tool, but for quantifying demographic, socio-economic status of the society.

Relatively pioneer work from 1928 by Burgess [34], states that the central of the cities, Zone 1 and 2 are generally the entry point for immigrants, and they prefer to be in the business district. His statements are quite relevant still. In his work Louf and Barthelemy claims that, larger cities are less socially polarized as higher income households prefer large cities, but low income do not[130].

In addition to the census based studies, there is a survey based study from 8 Italian cities [25]. Only Milan is in common in this study and the TIM data we use.

5.2.1 How to count a nation and assumptions

National institutes or census bureaus generally organize censuses in each 10 years, which is conducted by enumerators through a questionnaire [1]. Complementary or alternative approach is *sampling*, instead of counting each individuals, surveys are conducted through a subset of population then the results are projected to larger units. In general, in order to reduce the errors, in particular: *coverage, sampling, content* biases. Therefore censuses are controlled from several sources, such as administrative records.

The smallest unit is called as census tract (block) [213]. It is defined as homogeneous, compact, territorial units, which administrative boundaries

[1] https://unstats.un.org/unsd/demographic/sources/census/docs/P&R_Rev2.pdf

form through them. The boundaries of census tract has to be align with the administrative boundaries, which in turn allows building demographics from census tracts level up to the nation. The census results are aggregated in the level of census tract. From definition, the census tracts aim to have uniform number of people inside, however in year 2000 USA census, tract population is set to 4000 as optimum, with 1500 and 12000 as the minimum and the maximum thresholds. As you may notice, the variance of upper and lower bound is quite far. In addition to that, spatial information is also not being taken into account for the census tract construction, the only measure for setting the boundaries of tracts is population size [124]. Thus, according to the arguments of Lee *et al.* the conventional researches for segregation which mainly based on census data, may encounter measurement errors because of these assumptions.

5.2.2 Census areas and administrative structure in Italy

Italian National Institute of Statistics (ISTAT) is organizing the population census in every ten years. The last census has been done in 2011. The smallest unit of census area (tract) is called *Sezioni di censimento* in Italian[2]. However in addition to this, the data can be accessible also in different administrative levels, as the *sezioni di censimento* must be aligned with the administrative borders. Such as, the smallest administrative unit is called *commune*, followed by *province* and the first sub-nation level *regione*. In addition to that there is a categorization of metropolitan areas, called *circoscrizioni*.

Here, the census data, regarding each nationality resident in a given city, are obtained from ISTAT as shown in Table 5.1. The seven cities under consideration have different characteristics in terms of population hetero-

[2] http://www.istat.it/it/files/2016/02/AttiCIS_Fascicolo_5.pdf

geneity or the number of foreign residents. Interestingly, while Milan's foreign population is more heterogeneous, Rome and Turin are significantly dominated by the Romanian community (see Figure 5.1 and Figure 5.2).

City	Italian	Foreigners	% of Foreigners	# of Grid
Milan	2.769.201	439.308	0.1586	1.419
Rome	3.816.517	523.957	0.13	928
Turin	2.059.453	222.744	0.108	571
Naples	3.005.147	108.751	0.036	535
Venice	773.914	81.782	0.105	496
Bari	1.223.947	39.873	0.032	144
Palermo	1.235.797	35.609	0.028	165

Table 5.1: Italian and foreign population in the seven Italian cities. The list is ordered by the percentage of foreign population. Source ISTAT 2011.

In Figure 5.1 and Figure 5.2, the top presence of immigrant nationalities are shown from the cities Turin, Milan, Rome and Naples, respectively.

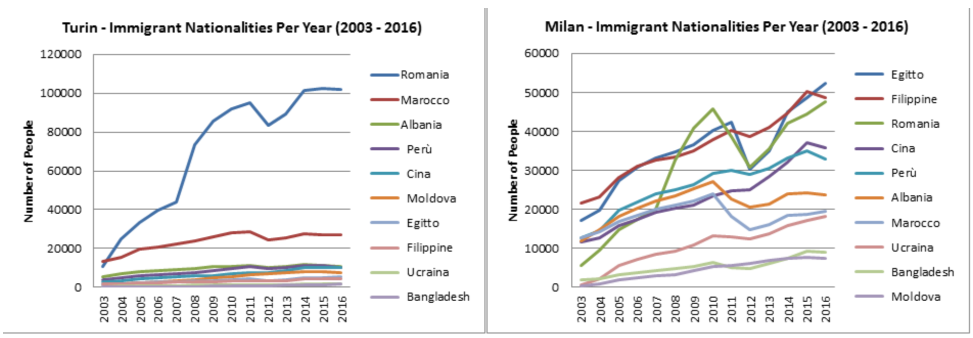

Figure 5.1: The ten top foreign nationalities present in Turin and Milan from 2003 to 2016 from ISTAT.

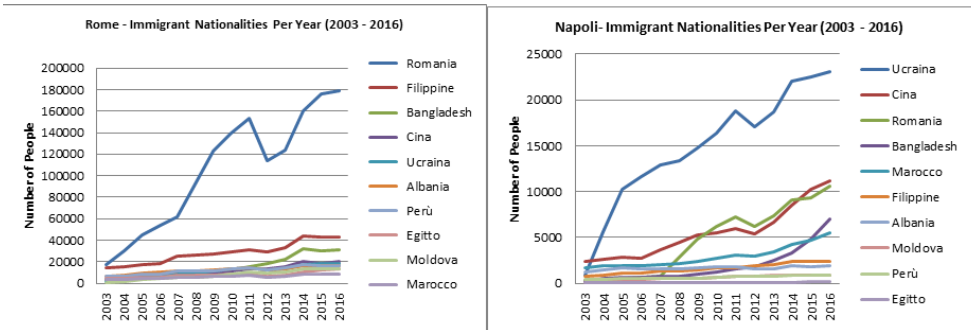

Figure 5.2: The ten top foreign nationalities present in Rome and Naples from 2003 to 2016 from ISTAT.

5.3 Social aspects of segregation

Cities are the places where pragmatic solutions need to be found daily to promote integration and to address the issues raised by the increasing ethnic and cultural diversity of the population. Several studies in urban planning and urban sociology have underpinned the importance of ethnic and cultural diversity of urban population [223, 189]. The heterogeneity of the communities may foster creativity [157], encourage tolerance [70], and boost economical growth [166]. For example, the urban planning theorist Leonie Sandercock terms her ideal city *Cosmopolis*, describing a metropolis that allows people from a variety of ethnic and racial backgrounds to have equal rights in the city environment [176, 177]. The political theorist and urban planner Susan Fainstein has shown that some cities are more successful to create diversity and social integration than others [70].

For most cities, the efforts to manage diversity and decrease segregation and discrimination rely on social policies, e.g., income support, public education, etc. that are usually under the responsibility of national governments. However, the social inclusion of newcomers also depends on the quality of the daily interactions and exchanges that occur among the individuals, communities, and social groups, and on how these communities live spatially in the different urban areas. Yet in our work, we would like to discover the segregated hot spots to improve integrity, beside being able to characterize the diverse hot spots, later allow us to use it on prior.

5.3.1 Urban diversity and economical development

In the last decades cities become the center of attention to people from various background, for the economical potential. Especially when we look at the whirlpool of big, cosmopolitan cities all around the world. The effect of segregation in the urban life has been questioned for the last decades [141, 167]. For the economic development the effect of diversity and segregation is the two side of the same coin. Segregation means social cohesion, if we recall from previous chapter, it corresponds to bounding structure of the network. If the segregated community is prosperous, wealth will stay in the same society, however if the majority is poor, they would not benefit from the prosperity. In sum, rich become richer and poor stays poorer. On the other hand, diversity means, structural holes and open structure in the community. The values/ideas can be interchangeable among the society, which fosters new ideas and prosperity [37, 170].

Adam Smith correlated the economic growth with the diversity and complexity of the labor [188]. Hidalgo and Hausmann empirically proved Adam's ideas in their work [101]. As mentioned in the well-known book of Acemoglu and Robinson "Why Nations fail?", only the nations which allows the society to able to reach the sources, with equal rights can perform better [8]. If and only the equal rights given to migrants may allow to integrate and contribute to the urban life. Lazer and Friedman like their scholars, stated that being connected is good, only if your network is composed of people different than you [123]. So diversity is not enough, unless it is integrated.

5.3.2 Urban diversity and public safety and security

In particular, some works have recently provided evidence supporting Jane Jacobs' natural surveillance theory [28, 204, 27]. In *The Death and Life of Great American Cities* [105], one of the most influential book in city planning, the urban activist Jane Jacobs stated that population diversity in a neighborhood contribute to natural surveillance and, thus, discourage crime. Indeed, as people are moving around an area, they effectively become "eyes on the street" able to observe what is going on around them.

Along a similar line, De Nadai *et al.* [54] used mobile phone data, jointly with Open Street Map and census data, to test the four diversity conditions (mixed land uses, small blocks, a mix of buildings of different ages and forms, and a dense concentration of people and buildings) proposed by Jane Jacobs to promote life in a city. Although the analyzed cities were different from the places for which Jacobs's conditions were spelled out (i.e., great American cities), De Nadai *et al.* [54] found those conditions to be indeed associated with urban life.

5.4 Quantifying urban spatial diversity and segregation

Measuring spatial structure is crucial for understanding how people utilize the urban space in time. It was first Duncan and Duncan, who take attention to the importance and challenges of quantifying segregation for the sake of understanding social well-being [63]. However it was Massey and Denton, who formalized the segregation indexes into five dimensions and their properties. Specifically they are:

- (i)*Evenness* refers to the differential distribution of two groups in urban space, such as over represented in some part of the city, not existing the rest [21].

- (ii) *Exposure* refers to quantify the potential contact, or the possibility of interaction, between minority and majority group members within the city [140].

- (iii) *Concentration* refers to the relative amount of physical space occupied by a minority group in the urban space. Groups that occupy a small share of the total area in a city are said to be residentially concentrated.

- (iv) *Centralization* is the degree to which a group is spatially located near the center of the city.

- (v) *Clustering* is the degree of spatial units inhabited by minority group, discarding the majority or become majority within that cluster.

The followers Reardon and Sullivan evolve the Massey and Denton approach by including two-group segregation to multi-group segregation [171].

5.4.1 Multi-group segregation metrics

We focus on how different ethnicity's utilize the city, therefore we implemented the multi-group segregation metrics from Reardon and Firebaugh [171]. In the following sections, we summarized the six measures for segregation.

Notation In the following segregation metrics equations, we will use the following notations: t denotes size and π denotes proportions; subscripts i and j index census tract and subscripts m and n are indexes for minority groups.
t_j = Population in unit j

T = Total population
π_m = Proportion of group m
π_{jm} = Proportion of group m in unit j.

Apart from these group proportions (π_m and π_{jm}), Simpson Interaction Index and Theil Entropy are used in the calculations of the measurements.

Simpson Interaction Index: Simpson Interaction Index, is denoted as I, [125, 217]:

$$I = \sum_{m=1}^{M} \pi_m (1 - \pi_m).$$

Theil Entropy: Theils Entropy Index, is denoted as E [196]:

$$E = \sum_{m=1}^{M} \pi_m \ln \left(\frac{1}{\pi_m}\right).$$

Both I and E can be seen as measures of the diversity of a population when both are equal to zero if and only if all individuals are members of a single group no diversity and both are maximized if and only if individuals are evenly distributed among the M groups ($\pi_m = \frac{1}{M}$) for all m.

The indices which are shown below, represent from 0 (complete integration) to 1 (complete segregation). Multi-group dissimilarity, Gini coefficient, information theory are belong to *Evenness* dimension.

Multi-group dissimilarity

Dissimilarity (D) measures the evenness of the populations, the percentage of a groups population that would have to change residence for each neighborhood to have the same percentage of that group as the metropolitan area.

$$D = \sum_{m=1}^{M} \sum_{j=1}^{J} \frac{t_j}{2TI} |\pi_{jm} - \pi_m|.$$

Gini coefficient

The Gini coefficient or Gini index is the mean absolute difference between minority proportions weighted across all pairs of areal units, expressed as a proportion of the maximum weighted mean difference [140].

Here, Reardon *et al.* generalized the two-group case of Gini index (G) to M group, by weighting sum over the absolute difference of all pairs, normalized by the maximum possible value of this sum.

$$G = \sum_{m=1}^{M} \sum_{i=1}^{J} \sum_{j=1}^{J} \frac{t_i t_j}{2T^2 I} |\pi_{im} - \pi_{jm}|.$$

Information theory

The information index (H) measures the (weighted) average deviation of each areal unit from the metropolitan areas entropy or racial and ethnic diversity, which is greatest when each group is equally represented in the metropolitan area.

According to Reardon's work, Information index is the best metric to quantify the multigroup segregation [171].

$$H = \sum_{m=1}^{M} \sum_{j=1}^{J} \frac{t_j}{TE} \pi_{jm} \ln \frac{\pi_{jm}}{\pi_m}.$$

Squared coefficient variation

The squared coefficient of variation (C) function can be interpreted as a measure of the variance of the dissimilarity functions or as a normalized chi-squared measure of association between groups and units.

$$C = \sum_{m=1}^{M} \sum_{j=1}^{J} \frac{(\pi_{jm} - \pi_m)^2}{(M-1)\pi_m}$$

Relative diversity

Relative diversity (R), measures the relative diversity of the clusters. Which is modified from Goodman and Kruskal's [87] cross classification work. It is the level of within cluster segregation, between units. Larger, more diverse, and more segregated clusters contribute more to overall segregation than do smaller, more homogeneous, and less segregated clusters [172].

$$R = \sum_{m=1}^{M} \sum_{j=1}^{J} \frac{t_j}{TI} (\pi_{jm} - \pi_m)^2$$

Normalized exposure

Normalized exposure (P), measures the exposure of each groups, by taking a simple weighted average of the normalized exposures of each group to all other groups [171].

$$P = \sum_{m=1}^{M} \sum_{j=1}^{J} \frac{t_j}{T} \frac{(\pi_{jm} - \pi_m)^2}{(1 - \pi_m)}$$

5.5 Methodology for assessing neighborhood preferences of different ethnicity

In that section, we describe the methodology which we use in to answer **Q3** and **Q4**. By calculating *Local Moran I*, we reveal the significant locations' of each nation in the urban area. The presence of different nationalities in the same location (grid), assumed as they do not avoid interacting.

5.5.1 Local Moran I

Morans I is a correlation coefficient that measures the overall spatial auto correlation of spatial data [149]. In other words, it measures how one object is similar to others surrounding it. This is actually about first law of geography, *"Everything is related to everything else, but near things are more related than distant things"* [201]. Beside from statistics based on independence of the variables, in geography the spatial distance is important. Thus it measures how closer objects are similar. It gives the overall relation in the given geographic region, therefore it is also called as global Moran I.

Local Moran I index, *local indicator of spatial association (LISA)*, computes the local spatial autocorrelation statistics disaggregated to the level of the spatial units, allowing assessment of the dependency relationships across space [11].

The equation is as follows:

$$I_i = \frac{x_i - \hat{X}}{S_i^2} \sum_{j=1, j \neq i}^{n} w_{i,j}(x_i - \hat{X}),$$

where, x_i is an attribute for the spatial unit i, $\hat{X}$ is the mean of corresponding attribute, $w_{i,j}$ is the spatial weight between unit i and j, n is the total number of spatial units. S_i is the aggregated value of the attribute in unit i.

5.5.2 Hot-spots of communities from CDR

For each block we calculate the local Moran-I of each community. If the p-value is significant ($p < 0.01$), we label this grid as a hot-spot for that community. If the hot-spots are overlapped with the other communities, it is assumed to prefer being in the similar neighborhoods.

The details are given below:

- Intersect grid and the urban metropolitan area, as in Figure 5.3.
- Share the number of calls according to the intersection area: i.e. if whole grid is in the metropolitan area all calls are taken for computation, if half of the area is inside, take half of the calls. Assuming calls are uniform distributed to the grid.
- For each different nationalities (j), and for each grid (i) calculate the normalized value as: $z_{i,j} = \frac{Call_{i,j}}{\sum_i Call_j}$
- For each i and j, calculate Local Moran I, and detect the hotspots.
- Find the area percentage intersect two different nationalities' hot spots. If $Pvalue_{i,j} \cap Pvalue_{i,k}$ where $j \neq k$, means nationality j and nationality k present in that location significantly,
- build up **M**, NxN, asymmetric matrix. N is the total number of nationalities exist in the CDR of the city $N = n_1, n_2, .., N$.
- The grids are not uniform in means of the land coverage, the dimensions vary. That is why, in order to understand the preference of staying in the same location, we consider the land coverage of the grid. $M_{j,k} = \frac{\sum H_j \cap H_k}{\sum H_j}$ where H_j is the hot spots land coverage of nationality j. H_k is for the nationality k.

- The cosine similarity of each nation j, is calculated for each city. This comparison will be performed for different city similarity.

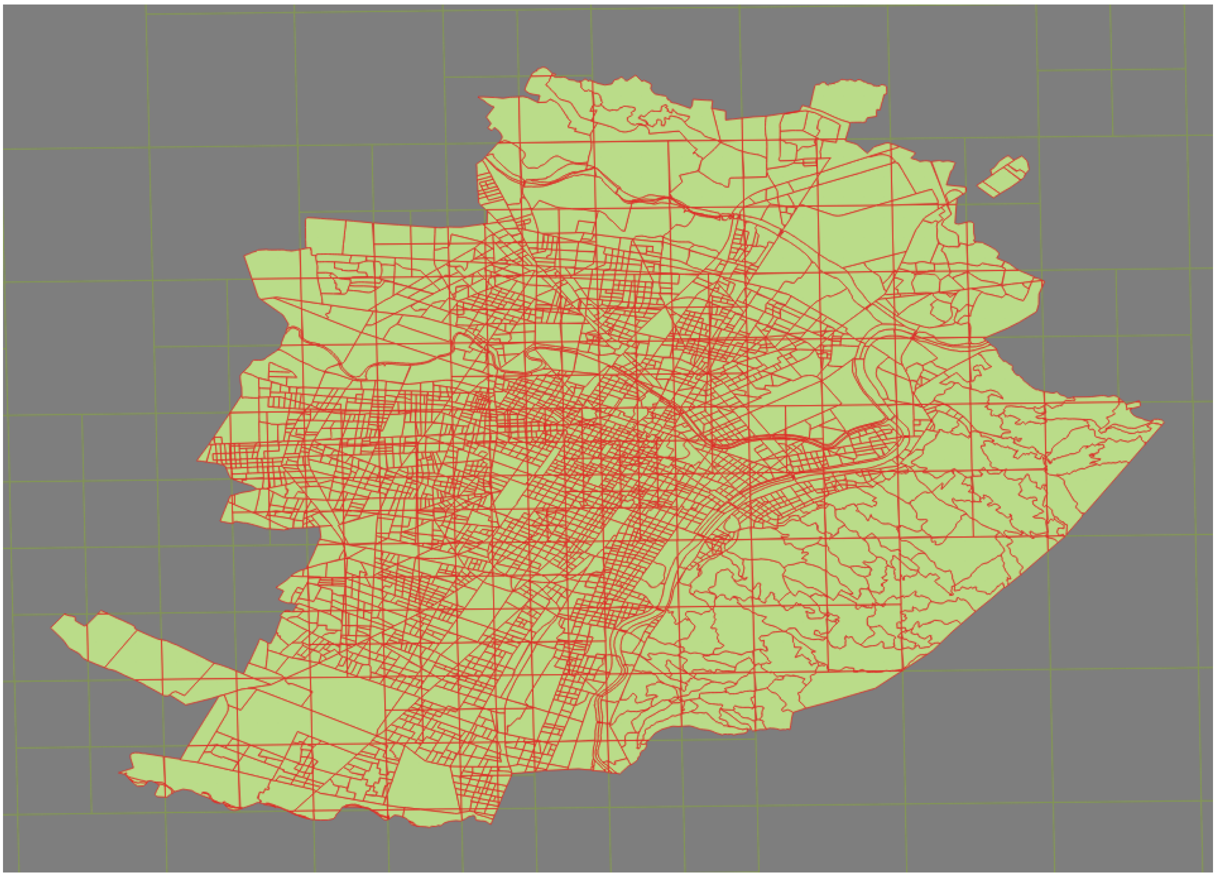

Figure 5.3: Turin call grids and census areas.

5.6 Results

In that section, we grouped our results in two separate sub-sections, (i) the segregation residing in census and the dynamic segregation metrics through CDR, (ii) the neighborhood preference of the immigrants.

In multigroup segregation metrics, three out of six metrics have very small values, very close to 0, especially from the CDR data. In particular, the metrics are: squared coefficient, relative diversity and normalized exposure. Both of them evaluates the within unit diversity, comparing to the urban area.

5.6.1 Static versus dynamic segregation

In this experiment, we compare static residential segregation metrics from census, and dynamic daily interactions through CDR, to exploit if both static and dynamic segregation are in line. CDR includes the country code of the incoming/outgoing communication however census data include continent level information, in particular: Europe, Africa, America, Asia, Oceania, and apolitical. Therefore, we grouped our country codes in these five continents, excluding the apolitical.

Here, we also keep in mind that the CDR data, is issued for billing purpose, therefore we actually do not know the nationality of the person, who make or receive a call (or send/receive a SMS). The call can either be business related, or tourists' call. However, the internet usage data have some difference, if the country code is different from the local operator's code, it means that the user is roaming.

In line with the studies conducted by Susan Fainstein [70], our results show that some cities experience more population heterogeneity than others. Indeed, even if Milan and Rome have a similar percentage of immigrants, as shown in Table 5.1, they have different segregation characteristics both in census, and CDR. According to residential segregation metrics from census, Milan is the most integrated for multigroup dissimilarity (0.1065), Gini coefficient (0.1404), and information theory (0.0103). Beside Rome has ($0.1879, 0.2521$ and 0.0319) respectively. When we look at the calls from CDR in Milan, the multigroup dissimilarity (0.4989), Gini coefficient (0.6454) and information theory is (0.1014), for Rome ($0.3782, 0.4818$ and 0.0409). It shows that Rome either immigrants or tourists have more diverse mobility patterns, therefore more integration comparing to Milan.

When we focus on Internet and SMS, Rome is the most segregated city among the six Italian cities for the three indices. Multigroup dissimilarity is for SMS and internet $(0.4120, 0.6130)$, Gini coefficient $(0.4912, 0.6909)$ and information theory $(0.0556, 0.1066)$, respectively. As we know from Internet usage, it shows the touristic activities, we can definitely say that tourists exploit only a portion of the city. Again from these figures 5.4 and 5.5, we can easily see that Milan is more business center, than Rome, as we do not expect to send/receive SMS for a developed country to make business. But it is more likely to use SMS for social connections. Thus, the share of tourists and businessmen in call data, can be evaluated within that perspective.

All segregation metrics results for six cities are given in Table 5.2.

Table 5.2: Multigroup segregation indexes for Bari, Milan, Napels, Palermo, Rome and Turin.

		Bari	**Milan**	**Napels**	**Palermo**	**Rome**	**Turin**
Multigroup Dissimilarity	**Census**	0.1963	0.1065	0.3683	0.2873	0.1879	0.1357
	CALL	0.3202	0.4989	0.5544	0.4741	0.3782	0.3801
	SMS	0.1720	0.2788	0.2948	0.3292	0.4120	0.1843
	INTERNET	0.1908	0.4926	0.3849	0.3137	0.6131	0.3766
Gini Coefficient	**Census**	0.2680	0.1404	0.4630	0.3875	0.2521	0.1847
	CALL	0.3857	0.6454	0.6642	0.5664	0.4818	0.4879
	SMS	0.2119	0.3684	0.3625	0.3976	0.4912	0.2353
	INTERNET	0.2438	0.6265	0.4813	0.3772	0.6909	0.4480
Information Theory	**Census**	0.0285	0.0103	0.0805	0.0680	0.0319	0.0193
	CALL	0.0386	0.1014	0.0904	0.0617	0.0409	0.0508
	SMS	0.0107	0.0319	0.0364	0.0351	0.0556	0.0137
	INTERNET	0.0105	0.0751	0.0409	0.0285	0.1066	0.0349
Squared Coefficient	**Census**	0.0013	0.0021	0.0059	0.0065	0.0060	0.0037
	CALL	0.0001	0.0003	0.0000	0.0000	0.0000	0.0000
	SMS	0.0000	0.0002	0.0000	0.0000	0.0005	0.0000
	INTERNET	0.0000	0.0002	0.0000	0.0000	0.0003	0.0000
Relative Diversity	**Census**	0.0036	0.0059	0.0181	0.0251	0.0143	0.0079
	CALL	0.0003	0.0008	0.0014	0.0011	0.0001	0.0001
	SMS	0.0001	0.0005	0.0010	0.0011	0.0015	0.0001
	INTERNET	0.0000	0.0006	0.0002	0.0002	0.0013	0.0000
Normalized Exposure	**Census**	0.0049	0.0077	0.0238	0.0329	0.0194	0.0103
	CALL	0.0003	0.0008	0.0024	0.0016	0.0001	0.0001
	SMS	0.0001	0.0007	0.0016	0.0018	0.0022	0.0001
	INTERNET	0.0000	0.0007	0.0004	0.0003	0.0017	0.0000

When we rank the cumulative evenness results (multigroup dissimilarity, Gini coefficient, and information theory) for each six cities, Napels is the most segregated, Bari is the least as shown in Figure 5.4. For the within unit segregation metrics, specifically relative diversity, normalized exposure and squared coefficient, Palermo is the most segregated. However keep in mind that the most of the shares are coming from residential segregation obtained from census, the mobility base segregation, is negligible close to zero.

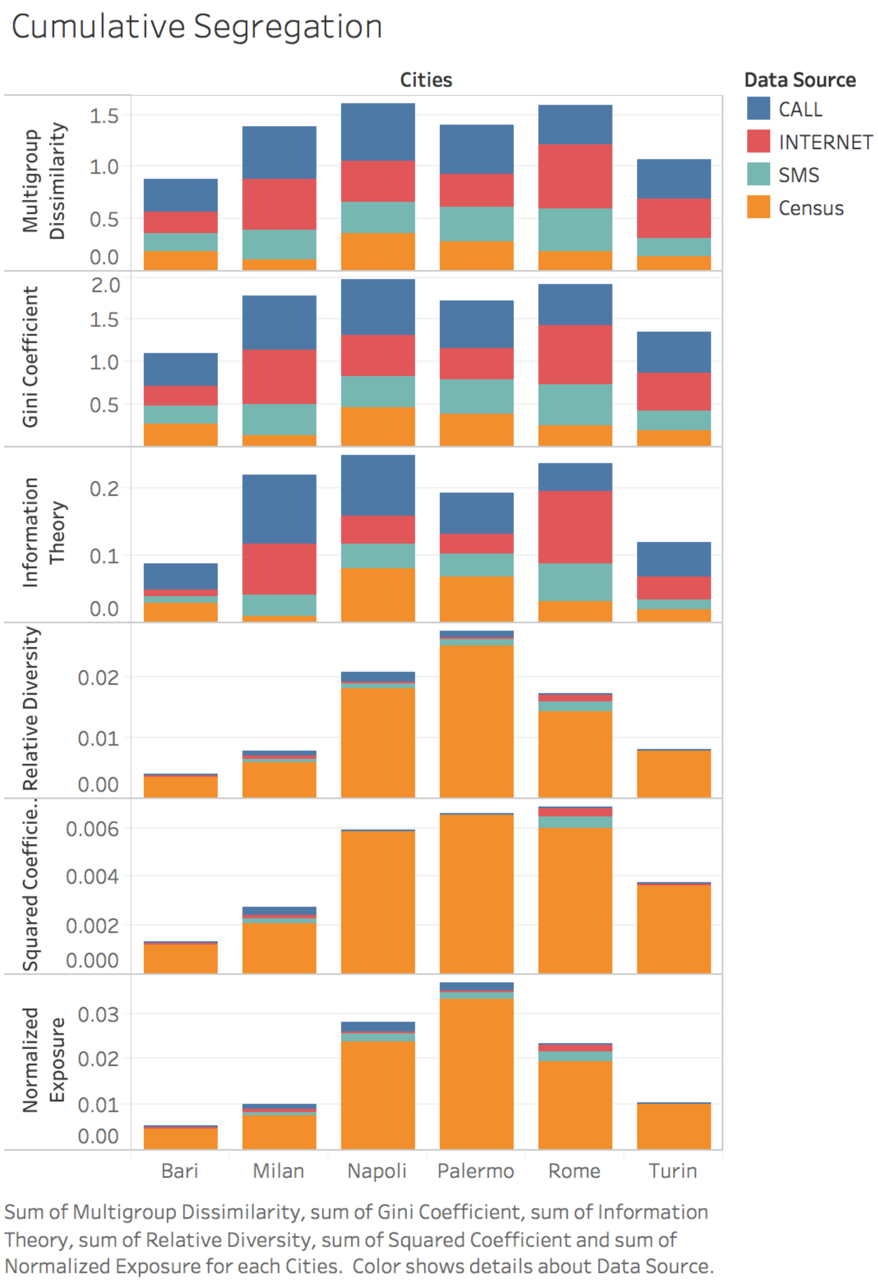

Figure 5.4: Cumulative multigroup segregation metrics for Bari, Milan, Napels, Palermo, Rome, and Turin.

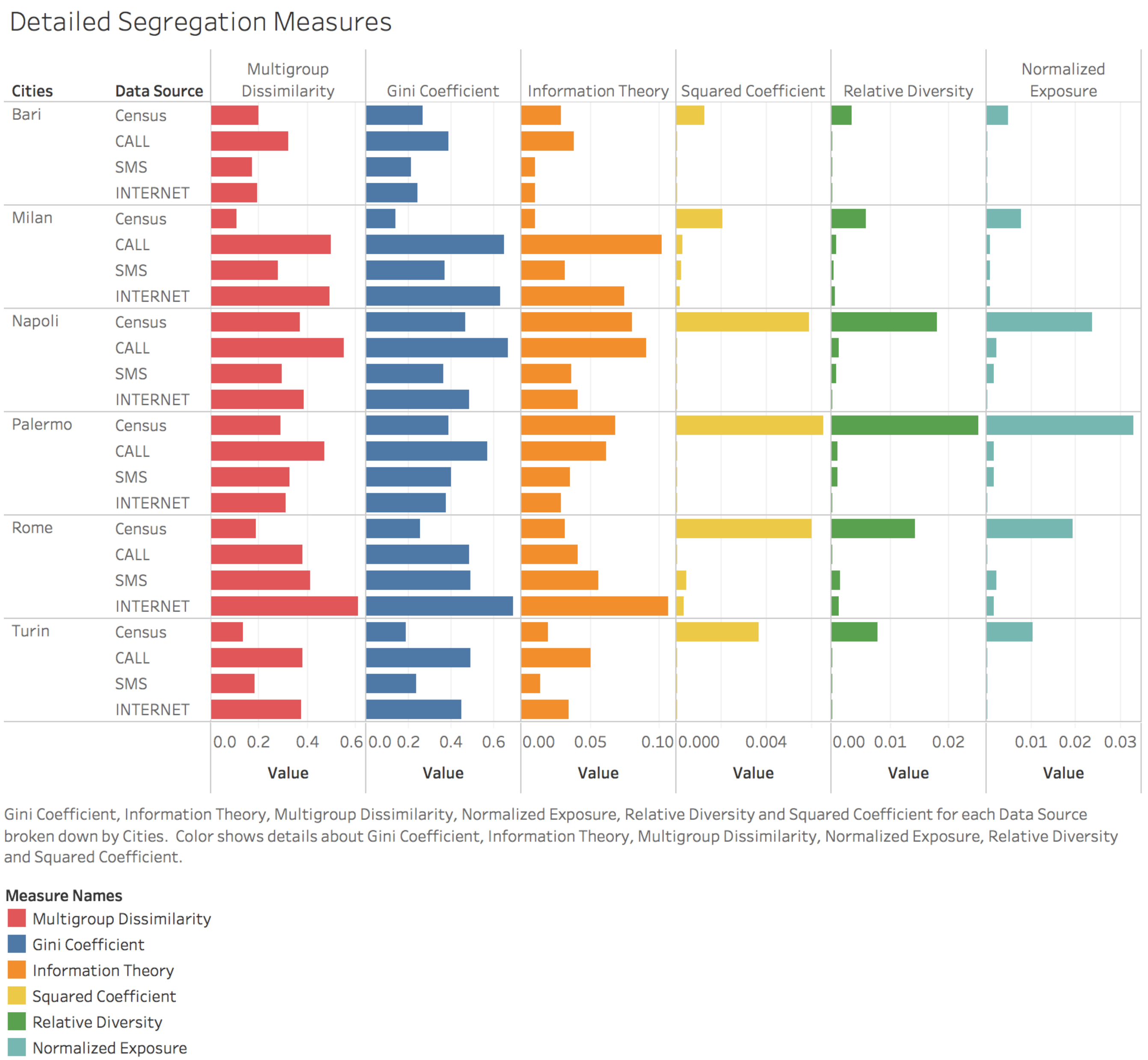

Figure 5.5: Detailed multigroup segregation metrics for Bari, Milan, Napels, Palermo, Rome, and Turin.

5.6.2 Neighborhood preferences for different nationalities

In that experiment, we exploit the favored locations of different nationalities in the city, not only considering where they live, but also where they work, travel and interact. Census data is not enough to assess such dynamic behaviors, therefore CDR is a better approach to understand these characteristics, if any.

Experiment setup 2 Here, we measure how different nationalities prefer to be present in different neighborhood within the urban space. The most representative ten nationalities have been selected from ISTAT data (see Table 5.3), which are; Albania, Bangladesh, China, Egypt, Morocco, Moldova, Peru, Philippines, Romania, and Ukraine. We are going to extend the cities to six, however up to now, we run the neighborhood preference study only for Turin.

	Turin	**Milan**	**Rome**	**Naples**
Top				
1	Albania	Albania*	Morocco	Moldova
2	India	Morocco	Peru	Brazil
3	Ukraine	Romania	Ecuador	Ukraine
4	Romania	Moldova	Albania	Algeria
5	Morocco	Brazil*	Moldova	Romania
Bottom				
16	Poland	Algeria	China	Senegal
17	Pakistan	China	Ukraine	Bangladesh
18	Nigeria	France*	Bangladesh	Pakistan
19	Bangladesh*	Senegal	Sri Lanka	Philippines
20	Brazil*	Poland*	Egypt	China

Table 5.3: The top five and bottom five communities, from ISTAT.

As we have seen in Table 5.4, the nationalities' preferences are not reciprocal. It is not row stochastic either as the same grid is the hot spot for other nationalities as well.

The results shows us that, Albanian's are more likely to be in the same neighborhoods, with Romanian and Ukrainian with the value (0.2421), and less likely with Bangladeshi. China and Bangladeshi are reciprocal for having same neighborhood to visit. Egypt and Moroccans have similar in means of preference. Peruvian prefers Ukrainian with (0.7448), and least Moroccans (0.0552). Philippines is similar to Peruvian with (0.7045) and (0.0455). Romanian are less likely to be in the Bangladeshi (0.0612) and Moroccans (0.0612), places. Peru, Ukraine and Albania's preference are more suitable for Romanians. Ukrainians are like Romanians they have less likely to interact with Bangladeshi and Moroccans. From these numbers we can see that, Romania, Ukraine and Albania some sort of clusters, and present same characteristics. Not only the way they interact, but also the interaction pairs of countries either in favor or not, is quite similar.

Table 5.4: Turin, hotspots overlaping matrix for 10 communities. Y axis represents the nationalities detected with hotspots, X axis represents how likely it overlaps the corresponding nationalities hotspots in the urban space.

	Albania	**Bangladesh**	**China**	**Egypt**	**Moldova**	**Morocco**	**Peru**	**Philippines**	**Romania**	**Ukrania**
Albania		0.0893	0.1071	0.0179	0.1786	0.0179	0.1071	0.1786	0.2321	0.2321
Bangladesh	0.2062		0.6186	0.4536	0.2887	0.4639	0.2474	0.2474	0.1237	0.2474
China	0.1791	0.4478		0.2687	0.2985	0.2687	0.2687	0.3134	0.2090	0.2687
Egypt	0.0250	0.2750	0.2250		0.2500	0.3000	0.2250	0.2000	0.2000	0.2000
Moldova	0.1606	0.1124	0.1606	0.1606		0.0482	0.2570	0.2731	0.1606	0.3052
Morocco	0.0408	0.4592	0.3673	0.4898	0.1224		0.0816	0.0816	0.1224	0.0816
Peru	0.1655	0.1655	0.2483	0.2483	0.4414	0.0552		0.6897	0.3862	0.7448
Philippines	0.2273	0.1364	0.2386	0.1818	0.3864	0.0455	0.5682		0.2045	0.7045
Romania	0.2653	0.0612	0.1429	0.1633	0.2041	0.0612	0.2857	0.1837		0.2449
Ukrania	0.1857	0.0857	0.1286	0.1143	0.2714	0.0286	0.3857	0.4429	0.1714	

The network between these ten communities, disregarding the spatial domain is as shown in Figure 5.6.

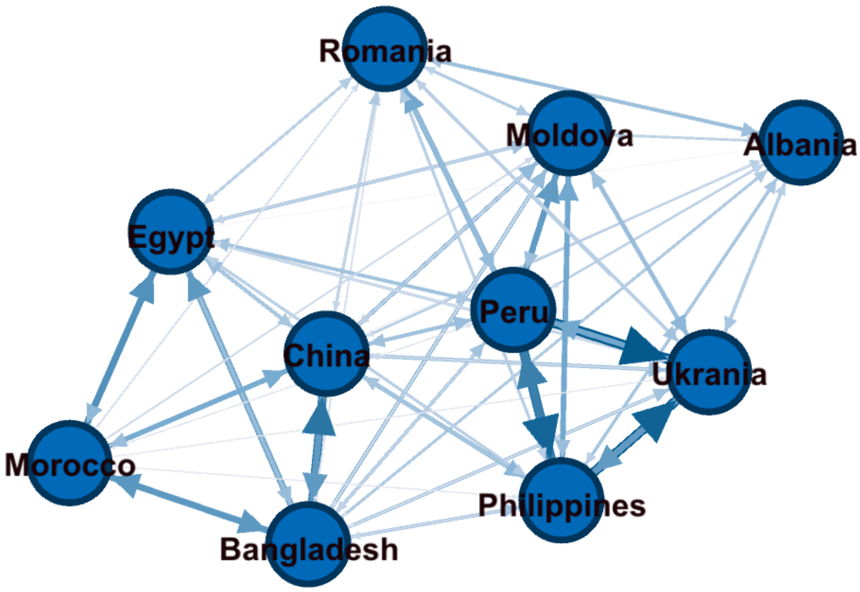

Figure 5.6: 10 minority prefered neighborhood from Turin.

5.7 Discussion

Segregation, is quite a challenging issue for urban scientists, municipalities and governments. Using static and mostly outdated census data has been the only tool to convey the analysis. Today, with the presence of different data sources, it is quite exciting to observe in high precision and almost real time, how different nations interact and use the urban space.

In the current work, we have explored, by using mobile phone data, how different ethnic and cultural communities live daily the urban space. We found out that high presence in census data does not necessarily imply a wide presence in the city. Some cultural effects seem also to shape how different ethnic communities live the urban space of a city, e.g., Bangladeshis and Chinese tend to stay in clustered communities in the city we analyzed.

However, there are quite limitations, and assumptions that we can say that using mobile phone data is quite a well proxy, for quantifying segregation problem.

As we mention in detail in Section 6.2, working with CDR has some challenges within itself. For this work we use Telecom Italia Mobile (TIM) Phone Dataset (see Appendix A.1.2). TIM has with 32.4% market share in Italy, it is fairly well representative to understand the call behavior. The incoming calls origin countries are assumed to be same as the destination of call receiver. For call and SMS, there is no discriminative properties to differentiate tourists from immigrants. While, for Internet usage data, it comes from roaming, if the country code is different from local.

The preference of one telecom operator is not uniform among the different ethnic groups. Such as in Turin, the immigrants represents the 15.4% of the citizenships (ISTAT 2015), in which the majority (39.5%) are from Romania. Their presence in the call dataset is proportional to their presence in the census data, this may be because they prefer using TIM as mobile phone operators, and use to communicate with their roots. In March 2013 TIM made a campaign to attract minorities, however this does not explain the rest of the other nationalities with negative correlation between census presence and number of calls. On the contrary, France, Poland, Swiss, Germany, Russia, Spain, UK show high presence in CDR dataset even with low presence in census. The possible explanations for that, there are more people living unregistered in Turin, or they can be tourists.

When we rank the dissimilarity for six cities, for call, SMS and internet usage, we can notice the most foreign attracted cities are Rome and Milan as shown in Figure 5.7. One can argue that Palermo and Napels are also famous as tourist attraction, however the time period corresponds to March and April, while the high season for south cities is summer.

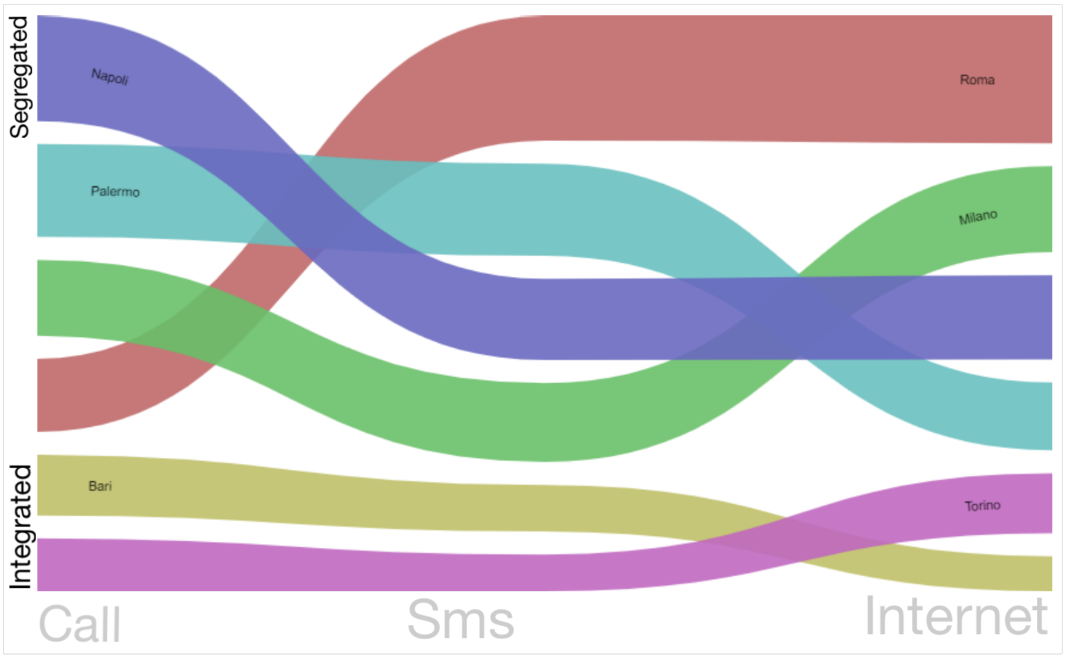

Figure 5.7: Dissimilarity of CDR data from six cities of Italy. Y axis from bottom to high, shows *Integrated* to *Segregated*. On X axis from the left part is Call data, in the middle SMS, and on the right Internet usage data are shown.

It is found out from survey based study in north Italy, that migrants who live in less segregated places are more likely to find a job [25]. In the same study, they found Bologna and Milan are the most homogeneous distributions among the immigrants. Our work based on census, gives the similar results as Milan is one of the integrated city, however the dynamic CDR data results do not reveal the same. Interestingly, Napels is the one of the most segregated city among the six cities we analyzed, however Bari, which is again in southern Italy but with less population, seems to be very integrated from CDR data.

We believe that the inclusion of immigrants to city life brings harmony and well-being to the society. There is also some scholars also work on the controversy, stating that the ethnic heterogeneity damages the harmony of the society [193]. Interestingly the results they have do not strongly prove their hypothesis.

The data we have, prevent us working on association between ethnic heterogeneity and socio-economic well-being, because of two reasons. One, the total number of cities, would provide us only six data point, which is not enough to say something significant. Second, the causality problem, more ethnic heterogeneity brings wealth or wealth of a region attracts different ethnicity, therefore, we avoid this association.

5.8 Conclusion

In that chapter we analyze ethnic groups' behaviour to utilize the urban space. To assess that we compare static segregation, from census data, and dynamic mobility among the city through CDR data.

In our first research question, we compare census and CDR segregation metrics, and see that they are quite different from each other. Which is in parallel with the study from Estonia [106]. In our second research question, for the segregation in bigger cities, we could not find a quite significant trace in either ways. Our last research questions about neighborhood preference, can partially performed, and leave the rest of that work as future work. But in the meantime, we conduct the analyze for Turin, to see how the different nationalities how leverage the city. We face similar clues about geographic closeness is an important factor for neighborhood preference, as we found in our previous work [95].

Chapter 6

Conclusion, contributions and insight for the future

...no matter how many instances of white swans we may have observed,
this does not justify the conclusion that all swans are white.
Sir Karl Raimund Popper

A data scientist may articulate the statement of Sir Karl Popper that swans are white, about the existence of black swans, one may argue that they are noises. We are not going to discuss about the swans, but the importance of data.

In the early ages, scientists have used first their intuition then their observations to understand the *nature*. The data they collect for tracing some evidence, was burdensome. Today we have piles of data, even more than we can process and be able to conclude meaningful outcomes as it arises.

Basically what we are doing in the piles of data, is trying to find the *Laplace's demon*. He stated that *"An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all* *items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes"* [121]. In the era of *Big Data*, we try to find the answers to our modern urban life problems, through the lens of mobile phone usage data. In the future, further policies, and privacy constraints may be overcome to allow these studies applicable for the well-being of social life.

6.1 Summary of the contributions

In this thesis, we investigate the usage of mobile phone data to solve several challenges in urban life. First, focus on detecting and characterizing emergency and non-emergency events. Specifically, we adopt the Markov modulated Poisson process framework [51, 104] (MMPP) for the spatio-temporal detection of hourly and daily behavioral anomalies in call volume, and we discuss the relationship between these anomalies and the actual emergency and non-emergency events that might have caused them. Compared to previous studies [108, 13, 17, 81, 203, 131, 77], we do not start with the location and the time of an already known event, but we use an unsupervised approach that spatio-temporally identifies unusual calling behavior. Moreover, our approach detects not only daily anomalies, but also hourly anomalies. Hence, we are able to capture behavioral responses occurring within hours of an event.

To validate our approach, we use mobile phone records from Côte d'Ivoire in Africa. The data, collected from December 1, 2011 to April 28, 2012 during the post-election crisis, were released for Orange's "Data for Development Challenge" (D4D) [3, 23] and contain calls between 5 million customers.

Our results show that we can successfully capture anomalous calling patterns associated with violent events, riots, as well as social non-emergency events such as holidays, sports events on an hourly and daily basis. Moreover, we illustrate that call volume changes also show significant temporal and spatial differences depending on the type of an event. Unlike previous work on classification of social events [224], we find that the coverage of the spatial impact is more significant than the duration.

Second, we advocate a two-faceted perspective on social capital; the value generated through social interactions originating from both closed bonding structures, rich in third-party relationships, and open bridging structures, rich in brokerage opportunities. We analyzed this well debated topic, in a data driven approach from two networks (communication and mobility) constructed through CDRs. The network metrics for representing the bridging/bonding characteristics have been associated with the several socio-economic indicators: economic well-being, security, civic engagement and public health. For economic well-being, in a developing country, the prospectus regions tend to communicate and interact in a bonding structure, beside the poorer regions behave more diverse and present more bridging activities in their communications. This finding is contradicting with several studies [64, 37], however there are some evidence from similar studies both from a developing country [128] and developed country [102] supporting our results. For safety our results shows that heterogeneity brings safer neighborhoods. Again heterogeneity and bridging structure is more driving force in the democratic participation and civic engagement. We could not find a significant relation between public health and communication and mobility patterns.

Third, we investigate the segregation in the urban metropolitan spaces, evidences from Italian cities. Apart from the common approach using census to understand segregation, we investigate the existence of dynamic, daily life segregation. After we compare different cities with each other, and showing the how it differs from census, we move in more detail. The neighborhood preference of different nationalities among the city is showing some similarities or not.

Although our aim is not to offer an explanation of the underlying causal factors, there may be some differences in social behavior for both developing and developed countries. Our study findings have highlighted several questions that could be investigated in future research, such as we need more data from several cities to make a more strong conclusions.

In summary, the main contributions of this thesis are:

- We construct a detailed database of emergency and non-emergency events in the Côte d'Ivoire using multiple sources of data and merge them with the geographical locations of the cell towers. This database is used as ground truth in our analysis;
- We adopt a Markov modulated Poisson process (MMPP) to spatio-temporally detect hourly and daily behavioral anomalies in call volume;
- We highlight and discuss the different spatial and temporal signatures of the discovered events;
- We shed light on the intrinsic value of social ties on socio-economic well-being of the society;
- We show that there is not one universal network structure which represents the value of social capital, thus it may be different for developing and developed countries;
- We show that census based segregation, and CDR based are not representative;

- We found some clues that some ethnic groups prefer same neighborhoods, especially if their country of origin is geographically close to each other.

6.2 Limitations of working with CDR data

Mobile phone usage data, or CDR is very powerful for urban analysis. However the limitations coming from the nature of the data source should be considered. According to Pestre *et al.* there are three points to tackle with during the CDR analysis [160]. Which are

Selection bias: The mobile phone operator, may not represent the underlying population. Such as some operators may more likely to be prefered by rich than poor or vice versa.

Compositional changes: The characteristic of the people may change over time, i.e. change the operator, or stop using phone.

Behavioral changes: Change in behavior for some reason, i.e. not using phone at the weekend etc.

In addition to these three main source of biases, we should add the fourth, the *socio-economic/geographic biases.* Especially in developing countries, the person who can not afford to buy a phone, can use other mobile phones, with charge. More general statement is, mobile phones are not a personal equipment, as in developed countries. Another bias, could be the one can have more than one mobile phone card, to use according to the promoted tariffs. The more prefered mobile phone subscriptions is prepaid (top-up) cards, rather than long term fixed subscriptions. Therefore, it is quite important to know the dynamics of the country, while analyzing CDR.

Another thing that Pestre *et al.* do not mention is, aggregation to avoid privacy issues. Mobile phone data is precious, to exploit sensible individual data [151], it is quite accepted to understand home and office location of a person through CDR data, even the precise trajectory can vary [169]. In order to avoid these issues, telecom operators share CDR data with researchers, aggregated through some predefined spatial areas. The geographic information, underlying this aggregation should be considered, for population distribution, or socio-economic well-being, most probably it is not uniform.

6.3 Future work and possible applications

There are several ways to enrich our works. Such as in Chapter 3 for anomalous event detection can be investigated with the data from a developed country to obtain a more general and strong. As mentioned in the original text it can be enriched with different source of data, such as social media, from Twitter and Facebook posts for more accurate online emergency alert system for public safety. It is crucial in the era of misinformation, to check different sources to have accurate interpretations.

The social capital related work mentioned in Chapter 4, can be extended with data both from developed and developing countries. It is quite necessary as, our findings are contradicting with the common intuition and some of the prior works. It is important for the policy makers to build up better societies to motivate interactions in a better way.

The urban segregation in Chapter 5, section on neighborhood preferences is a preliminary work, based on Turin. In that on going work, we will include the other five cities, how they are similar or different in utilization of the city by migrants. This patterns may reveal some clues such as gentrification. In current work we use LISA to assess spatial autocorrelation, however local Getis-Ord (Gi) [219] and local Geary [80] can be used to compare the performance of different methods.

Bibliography

[1] African mobile observatory 2011. http://www.gsma.com/spectrum/wp-content/uploads/2011/12/Africa-Mobile-Observatory-2011.pdf. [Online; accessed 5-Feb-2018].

[2] Armed conflict location and event data project. http://www.acleddata.com. [Online; accessed 5-Feb-2018].

[3] Data for development challenge. http://www.d4d.orange.com. [Online; accessed 5-Feb-2018].

[4] Le figaro newspaper. http://www.lefigaro.fr/flash-actu/2012/02/13/97001-20120213FILWWW00689-cote-d-ivoire-3-morts-dans-des-violences.php. [Online; accessed 5-Feb-2018].

[5] United nations refugee agency. http://www.unhcr.org/pages/4d831f586.html. [Online; accessed 5-Feb-2018].

[6] United nations security council reports. http://www.securitycouncilreport.org/un-documents/cote-divoire/. [Online; accessed 5-Feb-2018].

[7] The world in 2013, ict fact and figures. http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013-e.pdf. [Online; accessed 5-Feb-2018].

[8] Daron Acemoglu and James A Robinson. *Why nations fail: The origins of power, prosperity, and poverty.* Crown Business, 2013.

[9] Leman Akoglu and Christos Faloutsos. Event detection in time series of mobile communication graphs. In *Army Science Conference*, 2010.

[10] Yaniv Altshuler, Michael Fire, Erez Shmueli, Yuval Elovici, Alfred Bruckstein, Alex Sandy Pentland, and David Lazer. Detecting anomalous behaviors using structural properties of social networks. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 433–440. Springer Berlin Heidelberg, Berlin, 2013.

[11] Luc Anselin. Local indicators of spatial associationlisa. *Geographical analysis*, 27(2):93–115, 1995.

[12] Ibrahima Ba. Annuaire des statistiques demographiques et sociales. edition n7. *Socit dEtat*, page 135, 2012.

[13] James P. Bagrow, Dashun Wang, and Albert-Lszl Barabsi. Collective response of human populations to large-scale emergencies. *PLoS ONE*, 6(3):e17680, 03 2011.

[14] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, 2009.

[15] Michael Batty, Kay W Axhausen, Fosca Giannotti, Alexei Pozdnoukhov, Armando Bazzani, Monica Wachowicz, Georgios Ouzounis, and Yuval Portugali. Smart cities of the future. *The European Physical Journal Special Topics*, 214(1):481–518, 2012.

[16] Hila Becker, Mor Naaman, and Luis Gravano. Beyond trending topics: Real-world event identification on twitter. *ICWSM*, 11:438–441, 2011.

[17] Linus Bengtsson, Xin Lu, Anna Thorson, Richard Garfield, and Johan Von Schreeb. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in haiti. *PLoS medicine*, 8(8):1128, 2011.

[18] JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West. The markov modulated poisson process and markov poisson cascade with applications to web traffic modeling.

[19] Maria Bigoni, Stefania Bortolotti, Marco Casari, Diego Gambetta, and Francesca Pancotto. Amoral familism, social capital, or trust? the behavioural foundations of the italian north–south divide. *The Economic Journal*, 126(594):1318–1341, 2016.

[20] Richard A Bissell. Delayed-impact infectious disease after a natural disaster. *The Journal of emergency medicine*, 1(1):59–66, 1983.

[21] Peter Michael Blau. *Inequality and heterogeneity: A primitive theory of social structure*, volume 7. Free Press New York, 1977.

[22] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(10), 2015.

[23] Vincent D Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda,

and Cezary Ziemlicki. Data for development: the d4d challenge on mobile phone data. *arXiv preprint arXiv:1210.0137*, 2012.

[24] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

[25] Tito Boeri, Marta De Philippis, Eleonora Patacchini, and Michele Pellizzari. Moving to segregation: Evidence from 8 italian cities. 2012.

[26] Andrey Bogomolov, Bruno Lepri, Roberto Larcher, Fabrizio Antonelli, Fabio Pianesi, and Alex Pentland. Energy consumption prediction using people dynamics derived from cellular network data. *EPJ Data Science*, 5(1), 2016.

[27] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Emmanuel Letouzé, Nuria Oliver, Fabio Pianesi, and Alex Pentland. Moves on the street: classifying crime hotspots using aggregated anonymized data on people dynamics. *Big data*, 3(3):148–158, 2015.

[28] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, and Alex Pentland. Once upon a crime: towards crime prediction from demographics and mobile data. In *Proceedings of the 16th international conference on multimodal interaction*, pages 427–434. ACM, 2014.

[29] Stephen P Borgatti, Candace Jones, and Martin G Everett. Network measures of social capital. *Connections*, 21(2):27–36, 1998.

[30] Pierre Bourdieu. *The forms of capital.* Wiley Online Library, 1986.

[31] Samuel Bowles and Herbert Gintis. Social capital and community governance. *The Economic Journal*, 112(483), 2002.

[32] Carol Boyle, Gavin Mudd, James R Mihelcic, Paul Anastas, Terry Collins, Patricia Culligan, Marc Edwards, Jeremy Gabe, Patricia Gallagher, Susan Handy, et al. Delivering sustainable infrastructure that supports the urban built environment. *Environmental science & technology*, 44(13):4836–4840, 2010.

[33] Paolo Buonanno, Daniel Montolio, and Paolo Vanin. Does social capital reduce crime? *The Journal of Law and Economics*, 52(1):145–170, 2009.

[34] Ernest W Burgess. Residential segregation in american cities. *The Annals of the American Academy of Political and Social Science*, 140(1):105–115, 1928.

[35] Ronald S Burt. The gender of social capital. *Rationality and society*, 10(1):5–46, 1998.

[36] Ronald S Burt. The social capital of structural holes. *The new economic sociology: Developments in an emerging field*, pages 148–190, 2002.

[37] Ronald S Burt. *Structural holes: The social structure of competition.* Harvard university press, 2009.

[38] Francesco Calabrese, Massimo Colonna, Piero Lovisolo, Dario Parata, and Carlo Ratti. Real-time urban monitoring using cell phones: A case study in rome. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):141–151, 2011.

[39] Francesco Calabrese, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):0036–44, 2011.

[40] Francesco Calabrese, Francisco C Pereira, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. The geography of taste: analyzing cell-phone mobility and social events. In *Pervasive computing*, pages 22–37. Springer, Berlin, 2010.

[41] A Colin Cameron and Pravin K Trivedi. *Regression Analysis of Count Data*. Number 53. Cambridge University Press, Cambridge, UK, 2013.

[42] Karen E Campbell, Peter V Marsden, and Jeanne S Hurlbert. Social resources and socioeconomic status. *Social networks*, 8(1):97–117, 1986.

[43] Scott W Campbell and Nojin Kwak. Political involvement in mobilized society: The interactive relationships among mobile communication, network characteristics, and political participation. *Journal of Communication*, 61(6):1005–1024, 2011.

[44] Julián Candia, Marta C González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of physics A: mathematical and theoretical*, 41(22):224015, 2008.

[45] Dario Castiglione, Jan W Van Deth, and Guglielmo Wolleb. *The handbook of social capital*. Oxford University Press on Demand, 2008.

[46] Siddhartha Chib. Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221–241, 1998.

[47] Siddhartha Chib and Rainer Winkelmann. Markov chain monte carlo analysis of correlated count data. *Journal of Business & Economic Statistics*, 19(4), 2001.

[48] Soon Chun, Stuart Shulman, Rodrigo Sandoval, and Eduard Hovy. Government 2.0: Making connections between citizens, data and government. *Information Polity*, 15(1, 2):1–9, 2010.

[49] James S Coleman. Social capital in the creation of human capital. *American journal of sociology*, 94:S95–S120, 1988.

[50] James S Coleman. *Foundations of social theory*. Harvard University Press, 1990.

[51] David Roxbee Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 129–164, 1955.

[52] Diane Coyle and Patrick Meier. New technologies in emergencies and conflicts: the role of information and social networks. 2009.

[53] Donald R Davis, Jonathan I Dingel, Joan Monras, and Eduardo Morales. How segregated is urban consumption? Technical report, National Bureau of Economic Research, 2017.

[54] M De Nadai, J Staiano, R Larcher, N Sebe, D Quercia, and B Lepri. The death and life of great italian cities: A mobile phone data perspective. In *Proceedings of the 25th International Conference on World Wide Web*, pages 413–423, 2016.

[55] Mishra Deepak and Uwe Deichmann. *World Development Report 2016: Digital Dividends*. World Bank, Washington, DC. License: Creative Commons Attribution CC BY 3.0 IGO, 2016.

[56] Paul Dekker and Eric M Uslaner. *Social capital and participation in everyday life*. Routledge, 2001.

[57] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *PNAS*, 111(45):15888–15893, 2014.

[58] Pierre Deville, Chaoming Song, Nathan Eagle, Vincent D Blondel, Albert-László Barabási, and Dashun Wang. Scaling identity connects human mobility and social interactions. *Proceedings of the National Academy of Sciences*, page 201525443, 2016.

[59] Adrian Dobra, Nathalie E. Williams, and Nathan Eagle. Spatiotemporal detection of unusual human population behavior using mobile phone data. *PLoS ONE*, 10:e0120449, 03 2015.

[60] Lei Dong, Sicong Chen, Yunsheng Cheng, Zhengwei Wu, Chao Li, and Haishan Wu. Measuring economic activities of china with mobile big data. *arXiv preprint arXiv:1607.04451*, 2016.

[61] Yuxiao Dong, Fabio Pinelli, Yiannis Gkoufas, Zubair Nabi, Francesco Calabrese, and Nitesh V Chawla. Inferring unusual crowd events from mobile phone call detail records. In *Machine Learning and Knowledge Discovery in Databases*, pages 474–492. Springer, Portugal, 2015.

[62] Alain Dubresson. Abidjan: from the public making of a modern city to urban management of a metropolis. 1997.

[63] Otis Dudley Duncan and Beverly Duncan. A methodological analysis of segregation indexes. *American sociological review*, 20(2):210–217, 1955.

[64] Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.

[65] Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.

[66] Steeve Ebener, Christopher Murray, Ajay Tandon, and Christopher C Elvidge. From wealth to health: modelling the distribution of income per capita at the sub-national level using night-time light imagery. *international Journal of health geographics*, 4(1):1, 2005.

[67] Benjamin G Edelman and Michael Luca. Digital discrimination: The case of airbnb. com. 2014.

[68] Banfield Edward. The moral basis of a backward society. *Glencoe*, 111:85, 1958.

[69] Christopher D Elvidge, Paul C Sutton, Tilottama Ghosh, Benjamin T Tuttle, Kimberly E Baugh, Budhendra Bhaduri, and Edward Bright. A global poverty map derived from satellite data. *Computers & Geosciences*, 35(8):1652–1660, 2009.

[70] Susan S Fainstein. Cities and diversity should we want it? can we plan for it? *Urban affairs review*, 41(1):3–19, 2005.

[71] Katherine Faust and Stanley Wasserman. *Social Network Analysis: Methods and Applications*. Cambridge university press, 1994.

[72] Ray Forrest and Ade Kearns. Social cohesion, social capital and the neighbourhood. *Urban studies*, 38(12):2125–2143, 2001.

[73] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[74] Feng Fu, Corina E Tarnita, Nicholas A Christakis, Long Wang, David G Rand, and Martin A Nowak. Evolution of in-group favoritism. *Scientific reports*, 2:460, 2012.

[75] Francis Fukuyama. *Trust: The social virtues and the creation of prosperity*. Number D10 301 c. 1/c. 2. Free Press Paperbacks, 1995.

[76] Francis Galton. Vox populi (the wisdom of crowds). *Nature*, 75(7):450–451, 1907.

[77] Liang Gao, Chaoming Song, Ziyou Gao, Albert-László Barabási, James P Bagrow, and Dashun Wang. Quantifying information flow during emergencies. *Scientific reports*, 4, 2014.

[78] William Gardner, Edward P Mulvey, and Esther C Shaw. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological bulletin*, 118(3):392, 1995.

[79] Yanbo Ge, Christopher R Knittel, Don MacKenzie, and Stephen Zoepf. Racial and gender discrimination in transportation network companies. Technical report, National Bureau of Economic Research, 2016.

[80] Robert C Geary. The contiguity ratio and statistical mapping. *The incorporated statistician*, 5(3):115–146, 1954.

[81] Peter W. Gething and Andrew J. Tatem. Can mobile phone data improve emergency response to natural disasters? *PLoS Med*, 8(8):e1001085, 08 2011.

[82] Deborah E Gibbons. Interorganizational network structures and diffusion of information through a health system. *American Journal of Public Health*, 97(9):1684–1692, 2007.

[83] Marion Gibson. *Order from Chaos: Responding to Traumatic Events*. The Policy Press, Bristol, 2006.

[84] F Girardin, F Calabrese, F Dal Fiore, C Ratti, and J Blat. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive computing*, 7(4):36–43, 2008.

[85] Edward L Glaeser, David Laibson, and Bruce Sacerdote. An economic approach to social capital. *The Economic Journal*, 112(483), 2002.

[86] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[87] Leo A Goodman and William H Kruskal. Measures of association for cross classifications. *Journal of the American statistical association*, 49(268):732–764, 1954.

[88] Robert S Gottfried. *Black death*. Simon and Schuster, 2010.

[89] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.

[90] Mark S Granovetter. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380, 1973.

[91] Christiaan Grootaert. *Social capital and participation in everyday life*, chapter The missing link. Routledge, 2001.

[92] Christiaan Grootaert and Thierry Van Bastelaer. *The role of social capital in development: An empirical assessment*. Cambridge University Press, 2002.

[93] Luigi Guiso, Paola Sapienza, and Luigi Zingales. The role of social capital in financial development. *The American Economic Review*, 94(3):526–556, 2004.

[94] Luigi Guiso, Paola Sapienza, and Luigi Zingales. Civic capital as the missing link. Technical report, National Bureau of Economic Research, 2010.

[95] D Gundogdu, M Moretti, and B Lepri. Explorative analyse of two italian cities: Turin and venice for diversity. NetMob, 2017.

[96] Didem Gundogdu, Ozlem D Incel, Albert A Salah, and Bruno Lepri. Countrywide arrhythmia: emergency event detection using mobile phone data. *EPJ Data Science*, 5(1):25, 2016.

[97] Didem Gundogdu, Pietro Panzarasa, Nuria Oliver, and Bruno Lepri. Social capital and mobile phone communication: Evidence from a developing country. *Social Capital: An International Research Program*, 2018.

[98] Thoralf Gutierrez, Gautier Krings, and Vincent D Blondel. Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. *arXiv preprint arXiv:1309.4496*, 2013.

[99] Keith N Hampton. Comparing bonding and bridging ties for democratic engagement: Everyday use of communication technologies within social networks for civic and civil behaviors. *Information, Communication & Society*, 14(4):510–528, 2011.

[100] Johan Hellström and Anna Karefelt. Participation through mobile phones: A study of sms use during the ugandan general elections 2011. In *Proceedings of the fifth international conference on information and communication technologies and development*, pages 249–258. ACM, 2012.

[101] César A Hidalgo and Ricardo Hausmann. The building blocks of economic complexity. *proceedings of the national academy of sciences*, 106(26):10570–10575, 2009.

[102] Desislava Hristova, Matthew J Williams, Mirco Musolesi, Pietro Panzarasa, and Cecilia Mascolo. Measuring urban social diversity using interconnected geo-social networks. In *Proceedings of the 25th International Conference on World Wide Web*, pages 21–30. International World Wide Web Conferences Steering Committee, 2016.

[103] John Iceland. Beyond black and white: metropolitan residential segregation in multi-ethnic america. *Social Science Research*, 33(2):248–271, 2004.

[104] Alexander Ihler, Jon Hutchins, and Padhraic Smyth. Adaptive event detection with time-varying Poisson processes. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 207–216, 2006.

[105] Jane Jacobs. *Thedeath and Life of Great American Cities*. Jonathan Cape, 1961.

[106] Olle Järv, Kerli Müürisepp, Rein Ahas, Ben Derudder, and Frank Witlox. Ethnic differences in activity spaces as a characteristic of segregation: A study based on mobile phone usage in tallinn, estonia. *Urban Studies*, 52(14):2680–2698, 2015.

[107] Japan International Cooperation Agency (JICA). Republic of cote divoire the project for the development of the urban master plan in greater abidjan (sduga) final report. 2, 2015.

[108] Ashish Kapoor, Nathan Eagle, and Eric Horvitz. People, quakes, and communications: Inferences from call dynamics about a seismic

event and its influences on a population. In *AAAI Spring Symposium: Artificial Intelligence for Development*, 2010.

[109] Ichiro Kawachi, Lisa Berkman, et al. Social cohesion, social capital, and health. *Social epidemiology*, pages 174–190, 2000.

[110] Ichiro Kawachi, Sankaran Venkata Subramanian, and Daniel Kim. Social capital and health. *Social capital and health*, pages 1–26, 2008.

[111] William O Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences*, volume 115, pages 700–721. The Royal Society, 1927.

[112] Daniel Kim, SV Subramanian, Steven L Gortmaker, and Ichiro Kawachi. Us state-and county-level social capital in relation to obesity and physical inactivity: a multilevel, multivariable analysis. *Social science & medicine*, 63(4):1045–1059, 2006.

[113] Mikko Kivelä, Raj Kumar Pan, Kimmo Kaski, János Kertész, Jari Saramäki, and Márton Karsai. Multiscale analysis of spreading in a large communication network. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(03):P03005, 2012.

[114] Stephen Knack and Philip Keefer. Does social capital have an economic payoff? a cross-country investigation. *The Quarterly journal of economics*, 112(4):1251–1288, 1997.

[115] George Kritsotakis and Eva Gamarnikow. What is social capital and how does it relate to health? *International journal of nursing studies*, 41(1):43–50, 2004.

[116] Kevin S Kung, Kael Greco, Stanislav Sobolevsky, and Carlo Ratti. Exploring universal patterns in human home-work commuting from mobile phone data. *PLoS ONE*, 9(6), 2014.

[117] Fabio Lamanna, Maxime Lenormand, María Henar Salas-Olmedo, Gustavo Romanillos, Bruno Gonçalves, and José J Ramasco. Immigrant community integration in world cities. *arXiv preprint arXiv:1611.01056*, 2016.

[118] Renaud Lambiotte and Pietro Panzarasa. Communities, knowledge creation, and information diffusion. *Journal of Informetrics*, 3(3):180–190, 2009.

[119] Bram Lancee. *Immigrant performance in the labour market: Bonding and bridging social capital.* Amsterdam University Press, 2012.

[120] Morgane Laouénan and Roland Rathelot. Ethnic discrimination on an online marketplace of vacation rental. 2017.

[121] Pierre-Simon Laplace. *Pierre-Simon Laplace Philosophical Essay on Probabilities: Translated from the fifth French edition of 1825 With Notes by the Translator*, volume 13. Springer Science & Business Media, 2012.

[122] Vito Latora, Vincenzo Nicosia, and Pietro Panzarasa. Social cohesion, structural holes, and a tale of two measures. *Journal of Statistical Physics*, 151(3-4):745–764, 2013.

[123] David Lazer and Allan Friedman. The network structure of exploration and exploitation. *Administrative Science Quarterly*, 52(4):667–694, 2007.

[124] Barrett A Lee, Sean F Reardon, Glenn Firebaugh, Chad R Farrell, Stephen A Matthews, and David O'Sullivan. Beyond the census

tract: Patterns and determinants of racial segregation at multiple geographic scales. *American Sociological Review*, 73(5):766–791, 2008.

[125] Stanley Lieberson. Measuring population diversity. *American Sociological Review*, pages 850–862, 1969.

[126] Nan Lin. *Social capital: A theory of social structure and action*, volume 19. Cambridge university press, 2002.

[127] Nan Lin, Karen S Cook, and Ronald S Burt. *Social capital: Theory and research*. Transaction Publishers, 2001.

[128] Laura Lotero, Rafael G Hurtado, Luis Mario Floría, and Jesús Gómez-Gardeñes. Rich do not rise early: spatio-temporal patterns in the mobility networks of different socio-economic classes. *Royal Society open science*, 3(10):150654, 2016.

[129] Thomas Louail, Maxime Lenormand, Oliva García Cantú, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. *Scientific reports*, 4(5276), 2014.

[130] Rémi Louf and Marc Barthelemy. Patterns of residential segregation. *PloS one*, 11(6):e0157476, 2016.

[131] Xin Lu, Linus Bengtsson, and Petter Holme. Predictability of population displacement after the 2010 haiti earthquake. *PNAS*, 109(29):11576–11581, 2012.

[132] The Minh Luong, Vittorio Perduca, and Gregory Nuel. Hidden markov model applications in change-point analysis. *arXiv preprint arXiv:1212.1778*, 2012.

[133] Ann Macintosh. Characterizing e-participation in policy-making. In *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, pages 10–pp. IEEE, 2004.

[134] Marc Porter Magee. Civic participation and social capital: a social network analysis in two american counties. *Social capital: An international research program*, pages 308–331, 2008.

[135] Huina Mao, Xin Shuai, Yong-Yeol Ahn, and Johan Bollen. Mobile communications reveal the regional economy in côte divoire. *Proc. of NetMob*, 2013.

[136] Huina Mao, Xin Shuai, Yong-Yeol Ahn, and Johan 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.

[137] Peter V Marsden. Network data and measurement. *Annual review of sociology*, 16(1):435–463, 1990.

[138] Abraham H Maslow. A theory of human motivation. *Psychological review*, 50(4):370, 1943.

[139] Douglas S. Massey, Gretchen A. Condran, and Nancy A. Denton. The effect of residential segregation on black social and economic well-being. *Social Forces*, 66(1):29–56, 1987.

[140] Douglas S Massey and Nancy A Denton. The dimensions of residential segregation. *Social forces*, 67(2):281–315, 1988.

[141] Douglas S Massey and Nancy A Denton. *American apartheid: Segregation and the making of the underclass.* Harvard University Press, 1993.

[142] Naoki Masuda and Petter Holme. Predicting and controlling infectious disease epidemics using temporal networks. *F1000prime reports*, 5, 2013.

[143] Doug McAdam. Revisiting the us civil rights movement: Toward a more synthetic understanding of the origins of contention. *Rethinking social movements: Structure, meaning, and emotion*, pages 201–32, 2004.

[144] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[145] Sandro Meloni, Nicola Perra, Alex Arenas, Sergio Gómez, Yamir Moreno, and Alessandro Vespignani. Modeling human mobility responses to the large-scale spreading of infectious diseases. *Scientific reports*, 1, 2011.

[146] Giovanna Miritello, Rubén Lara, Manuel Cebrian, and Esteban Moro. Limited communication capacity unveils strategies for human interaction. *Scientific reports*, 3, 2013.

[147] Giovanna Miritello, Esteban Moro, and Rubén Lara. Dynamical strength of social ties in information spreading. *Physical Review E*, 83(4):045102, 2011.

[148] Carey Usher Mitchell and Mark LaGory. Social capital and mental distress in an impoverished community. *City & Community*, 1(2):199–222, 2002.

[149] Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[150] Hazel A Morrow-Jones and Charles R Morrow-Jones. Mobility due to natural disaster: Theoretical considerations and preliminary analyses. *Disasters*, 15(2):126–132, 1991.

[151] Mirco Musolesi. Big mobile data mining: Good or evil? *IEEE Internet Computing*, 18(1):78–81, 2014.

[152] Kevin Myers. Remembering refugees: Then and now. by tony kushner. *Cultural and Social History*, 5(3):379–382, 2008.

[153] Tommaso Nannicini, Andrea Stella, Guido Tabellini, and Ugo Troiano. Social capital and political accountability. *American Economic Journal: Economic Policy*, 5(2):222–250, 2013.

[154] Mark Newman. *Networks: an introduction*. Oxford university press, 2010.

[155] Brigit Obrist, Guéladio Cissé, Brama Koné, Kouassi Dongo, Stefanie Granado, and Marcel Tanner. Interconnected slums: water, sanitation and health in abidjan, côte d'ivoire. *The European Journal of Development Research*, 18(2):319–336, 2006.

[156] J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the national academy of sciences*, 104(18):7332–7336, 2007.

[157] Gianmarco IP Ottaviano and Giovanni Peri. The economic value of cultural diversity: evidence from us cities. *Journal of Economic geography*, 6(1):9–44, 2006.

[158] Pavlos Paraskevopoulos, T Dinh, Zolzaya Dashdorj, Themis Palpanas, and Luciano Serafini. Identification and characterization of human behavior patterns from mobile phone data. In *International*

Conference the Analysis of Mobile Phone Datasets (NetMob 2013), Special Session on the Data for Development (D4D) Challenge, 2013.

[159] Duncan Pedersen. Political violence, ethnic conflict, and contemporary wars: broad implications for health and social well-being. *Social science & medicine*, 55(2):175–190, 2002.

[160] Gabriel Pestre, E Letouzé, and E Zagheni. The abcde of big data. assessing biases in call-detail records for development estimates. Annu. Bank Conf. Dev. Econ., June, 2016.

[161] Jan H Pierskalla and Florian M Hollenbach. Technology and collective action: The effect of cell phone coverage on political violence in africa. *American Political Science Review*, 107(2):207–224, 2013.

[162] Robert Putnam. Social capital: Measurement and consequences. *Canadian Journal of Policy Research*, 2(1):41–51, 2001.

[163] Robert D Putnam. The prosperous community. *The american prospect*, 4(13):35–42, 1993.

[164] Robert D Putnam. Bowling alone: America's declining social capital. *Journal of democracy*, 6(1):65–78, 1995.

[165] Robert D Putnam, Robert Leonardi, and Raffaella Y Nanetti. *Making democracy work: Civic traditions in modern Italy.* Princeton university press, 1994.

[166] John M Quigley. Urban diversity and economic growth. *The Journal of Economic Perspectives*, 12(2):127–138, 1998.

[167] Lincoln Quillian. Segregation and poverty concentration: The role of three segregations. *American Sociological Review*, 77(3):354–379, 2012.

[168] AE Raftery and VE Akman. Bayesian analysis of a poisson process with a change-point. *Biometrika*, pages 85–89, 1986.

[169] Gyan Ranjan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mobile Computing and Communications Review*, 16(3):33–44, 2012.

[170] Ray Reagans and Ezra W Zuckerman. Networks, diversity, and productivity: The social capital of corporate r&d teams. *Organization science*, 12(4):502–517, 2001.

[171] Sean F Reardon and Glenn Firebaugh. 2. measures of multigroup segregation. *Sociological methodology*, 32(1):33–67, 2002.

[172] Sean F Reardon, John T Yun, and Tamela McNulty Eitle. The changing structure of school segregation: Measurement and evidence of multiracial metropolitan-area school segregation, 1989–1995. *Demography*, 37(3):351–364, 2000.

[173] José Rodriguez-Avi, María José Olmo-Jiménez, Antonio Condesánchez, and Ana María Martínez-Rodríguez. A new regression model for overdispersed count data. In *of the 29-th European Meeting of Statisticians*, page 256, 2013.

[174] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proc. 19th Int. Conf. on WWWW*, pages 851–860, 2010.

[175] Jose San Pedro, Davide Proserpio, and Nuria Oliver. Mobiscore: towards universal credit scoring from mobile phone data. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 195–207. Springer, 2015.

[176] Leonie Sandercock. *Towards Cosmopolis: Planning for multicultural cities*. John Wiley, 1998.

[177] Leonie Sandercock. *Cosmopolis II: Mongrel cities in the 21st century*. Continuum, 2003.

[178] Jimy M Sanders and Victor Nee. Immigrant self-employment: The family as social capital and the value of human capital. *American sociological review*, pages 231–249, 1996.

[179] Saskia Sassen. *The global city: New York, London, Tokyo*. Princeton University Press, 1991.

[180] Saskia Sassen. *Guests and aliens*. New Press, 1999.

[181] Thomas C Schelling. Models of segregation. *The American Economic Review*, 59(2):488–493, 1969.

[182] Thomas C Schelling. Dynamic models of segregation. *Journal of mathematical sociology*, 1(2):143–186, 1971.

[183] Markus Schläpfer, Luís MA Bettencourt, Sébastian Grauwin, Mathias Raschke, Rob Claxton, Zbigniew Smoreda, Geoffrey B West, and Carlo Ratti. The scaling of human interactions with city size. *Journal of The Royal Society Interface*, 11(98):20130789, 2014.

[184] Steven L Scott. Bayesian analysis of a two-state markov modulated poisson process. *Journal of Computational and Graphical Statistics*, 8(3):662–670, 1999.

[185] Steven L Scott and Padhraic Smyth. The Markov modulated Poisson process and Markov Poisson cascade with applications to web traffic data. *Bayesian Statistics*, 7:671–680, 2003.

[186] Ferdinando Semboloni. Planning the evolution of a city: a case study of abidjan. *Third World Planning Review*, 21(2):201, 1999.

[187] Jacob N Shapiro and Nils B Weidmann. Talking about killing: Cell phones, collective action, and insurgent violence in iraq. Technical report, DTIC Document, 2011.

[188] Adam Smith and Joseph Shield Nicholson. *An Inquiry Into the Nature and Causes of the Wealth of Nations...* T. Nelson and Sons, 1887.

[189] Michael P Smith. *City, state, and market: The political economy of urban society*. B. Blackwell, 1988.

[190] Chris Smith-Clarke and Licia Capra. Information diffusion and economic development.

[191] Victor Soto, Vanessa Frias-Martinez, Jesus Virseda, and Enrique Frias-Martinez. Prediction of socioeconomic levels using cell phone records. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 377–388. Springer Berlin Heidelberg, 2011.

[192] William T Story, Fahmida Taleb, SM Monirul Ahasan, and Nabeel A Ali. Validating the measurement of social capital in bangladesh: A cognitive approach. *Qualitative health research*, 25(6):806–819, 2015.

[193] Patrick Sturgis, Ian Brunton-Smith, Jouni Kuha, and Jonathan Jackson. Ethnic diversity, segregation and the social cohesion of neighbourhoods in london. *Ethnic and Racial Studies*, 37(8):1286–1309, 2014.

[194] James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[195] Andrew J Tatem, Youliang Qiu, David L Smith, Oliver Sabot, Abdullah S Ali, and Bruno Moonen. The use of mobile phone data for the estimation of the travel patterns and imported plasmodium falciparum rates among zanzibar residents. *Malaria journal*, 8(1):287, 2009.

[196] Henry Theil. Statistical decomposition analysis; with applications in the social and administrative sciences. Technical report, 1972.

[197] Jeffrey Martin Timberlake and Mario D Ignatov. *Residential Segregation*. Oxford University Press, 2014.

[198] DB Tindall and Jeffrey J Cormier. Gender, network capital, social capital, and political capital: the consequences of personal network diversity for environmentalists in british columbia. *Social Capital: An International Research Program*, pages 282–307, 2008.

[199] Michele Tizzoni, Paolo Bajardi, Adeline Decuyper, Guillaume Kon Kam King, Christian M Schneider, Vincent Blondel, Zbigniew Smoreda, Marta C González, and Vittoria Colizza. On the use of human mobility proxies for modeling epidemics. *PLoS Comput Biol*, 10, 07 2014.

[200] The U.S. Ambassador to Cte dIvoire. Country operational plan (cop) 2016 strategic direction summary. in press.

[201] Waldo R Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970.

[202] Jameson L Toole, Yu-Ru Lin, Erich Muehlegger, Daniel Shoag, Marta C González, and David Lazer. Tracking employment shocks using mobile phone data. *Journal of The Royal Society Interface*, 12(107):20150185, 2015.

[203] Vincent A Traag, Arnaud Browet, Francesco Calabrese, and Frédéric Morlot. Social event detection in massive mobile phone data using probabilistic location inference. In *IEEE Third Int. Conf. on Social Computing*, pages 625–628, 2011.

[204] M. Traunmueller, G. Quattrone, and L. Capra. Mining mobile phone data to investigate urban crime theories at scale. In *International Conference on Social Informatics*, pages 396–411, 2014.

[205] Frederick Jackson Turner and Allan G Bogue. *The frontier in American history*. Courier Corporation, 2010.

[206] Eleonora P Uphoff, Kate E Pickett, Baltica Cabieses, Neil Small, and John Wright. A systematic review of the relationships between social capital and socioeconomic inequalities in health: a contribution to understanding the psychosocial pathway of health inequalities. *International Journal for Equity in Health*, 12(1):54, 2013.

[207] Eric M Uslaner. Democracy and social capital. *Democracy and trust*, pages 121–150, 1999.

[208] Gordon Walker, Bruce Kogut, and Weijian Shan. Social capital, structural holes and the formation of an industry network. *Organization science*, 8(2):109–125, 1997.

[209] Huayong Wang, Francesco Calabrese, Giusy Di Lorenzo, and Carlo Ratti. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 318–323. IEEE, 2010.

[210] Shr-Jie Wang, Christian Lengeler, Thomas A Smith, Penelope Vounatsou, Guéladio Cissé, and Marcel Tanner. Rapid urban malaria

appraisal (ruma) iii: epidemiology of urban malaria in the municipality of yopougon (abidjan). *Malaria journal*, 5(1):29, 2006.

[211] John T Watson, Michelle Gayer, and Maire A Connolly. Epidemics after natural disasters. *Emerging infectious diseases*, 13(1):1, 2007.

[212] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998.

[213] John Robert Weeks. Population: An introduction to concepts and issues. 1986.

[214] Amy Wesolowski, Caroline O Buckee, Linus Bengtsson, Erik Wetter, Xin Lu, and Andrew J Tatem. Commentary: containing the ebola outbreak-the potential and challenge of mobile network data. *PLoS currents*, 6, 2014.

[215] Amy Wesolowski, Nathan Eagle, Andrew J Tatem, David L Smith, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.

[216] Gary C. White and Robert E. Bennetts. Analysis of frequency count data using the negative binomial distribution. *Ecology*, 77(8):pp. 2549–2557, 1996.

[217] Michael J White. Segregation and diversity measures in population distribution. *Population index*, pages 198–221, 1986.

[218] Apichon Witayangkurn, Teerayut Horanont, Yoshihide Sekimoto, and Ryosuke Shibasaki. Anomalous event detection on large-scale gps data from mobile phones using hidden markov model and cloud

platform. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*, pages 1219–1228. ACM, 2013.

[219] M Wulder and B Boots. Local spatial autocorrelation characteristics of remotely sensed imagery assessed with the getis statistic. *International Journal of Remote Sensing*, 19(11):2223–2231, 1998.

[220] Michael Xie, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. Transfer learning from deep features for remote sensing and poverty mapping. *arXiv preprint arXiv:1510.00098*, 2015.

[221] Quonan Christian Yao-Kouassi. *A la recherche dune synergie pour la gestion des dchets mnagers en Cote dIvoire: Cas du districy d'Abidjan*. PhD thesis, Universit du Maine, Gographie Department, 2010.

[222] Tadafumi Yoshihara, Shoji Kasahara, and Yutaka Takahashi. Practical time-scale fitting of self-similar traffic with Markov-modulated Poisson process. *Telecommunication Systems*, 17(1-2):185–211, 2001.

[223] Iris Marion Young. *Justice and the Politics of Difference*. Princeton University Press, 2011.

[224] William Chad Young, Joshua E Blumenstock, Emily B Fox, and Tyler 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.

[225] Huiqi Zhang, Ram Dantu, and João W Cangussu. Change point detection based on call detail records. In *Intelligence and Security*

Informatics, 2009. ISI'09. IEEE International Conference on, pages 55–60. IEEE, 2009.

[226] Yang Zhang, Nirvana Meratnia, and Paul Havinga. Outlier detection techniques for wireless sensor networks: A survey. *Communications Surveys & Tutorials, IEEE*, 12(2):159–170, 2010.

Appendix A

Data and Materials

In this section, we introduce the datasets that have been used to evaluate our proposed methodology: (i) a mobile phone records dataset, obtained by Orange for the "D4D Challenge,"[1] and (ii) an event records dataset obtained from a variety of public sources (e.g., Armed Conflict Location and Event Data Project, United Nations Council and International Crisis Group reports, local and international news).

A.1 Call Details Records

Mobile telecommunication operators collect Call Detail Records (CDRs) for billing purposes. These records generally consist of a number of fields, including the encrypted originating and destination numbers, the timestamp of the call, the duration of the call and the location of the cell towers that the originating number was connected to (the cell tower ID of the destination number could be available if it belongs to the same operator). Typically, there are CDRs for voice events, SMS events, and data traffic events. Even though the data is typically encrypted and all personal information is removed, CDRs are not easily accessible due to regulatory and legal constraints. The datasets we are analyzing in this paper were

[1]http://www.d4d.orange.com/en/Accueil

provided as part of two different data challenges: Data for development (D4D) 2013 data challenge organized by Orange Telecom and MIT Media Lab[2] and Telecom Italia Big Data Challenge.

A.1.1 Data for Development Dataset, Côte d'Ivoire

The Orange dataset contains anonymized and aggregated calls between 5 million customers from December 1, 2011 to April 28, 2012 (referred to as Set 1 in the D4D Challenge). Specifically, the dataset contains the total volume (number of calls and SMSs) and duration of calls between each pair of cell towers over the entire period. The total number of cell towers is 1238; however, in the pre-processing phase, we have eliminated the cell towers, which are not present during the whole period, ending up with 970 cell towers to analyze. The exact locations of the towers are not provided by Orange Telecom, due to the company's operational confidentiality. Finally, it is worth noticing that the data consist of the total traffic between cell towers, and hence no individual data are ever accessed.

Table A.1: Summary statistics of the Côte d'Ivoire dataset and the country.

Total Area of the Country (Hectares)	32,246,300
Population of the Country	around 20 million
Number of Subscribers	5 million
Type of the Data	Cell tower to cell tower
Total Number of Calls	190,638,909
Time Period	7 weeks / 20 weeks

Mobile phone penetration is high enough (95%) [1] in Côte d'Ivoire to make such a dataset sufficiently representative of the population. Moreover, the network operator holds a dominant position in Côte d'Ivoire (48% of the market share). Table A.1 provides summary statistics for the country and the dataset.

[2] http://www.d4d.orange.com/en/Accueil

Table A.2: Data summary of D4D subsets.

	Subset1	**Subset2**	**Subset3**	**Subset4**
Num.User	-	50,000	500,000	5,000
Duration	5 Months	5 Months	5 Months	5 Months
Type	AntAnt.Call	UserAntenna	UserSubpref.	User-User

Specifically, the call data consist of (i) date, (ii) hour, (iii) initiating cell tower, (iv) destination cell tower, (v) aggregated number of calls, and (vi) aggregated duration of calls. The CDR data with an unassigned cell tower are deleted. From December 14, 2011 to January 19, 2012, half of the country had a shortage of energy, confirmed by the Orange Telecom authorities. The data collected during this period are deleted. Additionally, we have missing data from 22nd of January to 30th of January. To preserve the weekly periodicity without being affected by data collection issues, we analyze the data from December 5 to December 11, and from January 30 to March 11 (49 days in total).

To provide some geographical context, Figure A.1 shows the rough locations of cell towers in relation to the regional boundaries of Côte d'Ivoire (255 subprefectures). It shows the dense concentration of cell towers in and around the largest cities of each region. Abidjan, the economic capital of Côte d'Ivoire, has a significantly higher concentration of cell towers compared to the rest of the country.

In this section, we describe the datasets that have been used to evaluate our proposed methodology and that were part of D4D Côte d'Ivoire Challenge[3].

The dataset contains anonymized and aggregated calls between 5 million customers from December 1, 2011 to April 28, 2012 (referred to as Set 1

[3]http://www.d4d.orange.com/en/Accueil

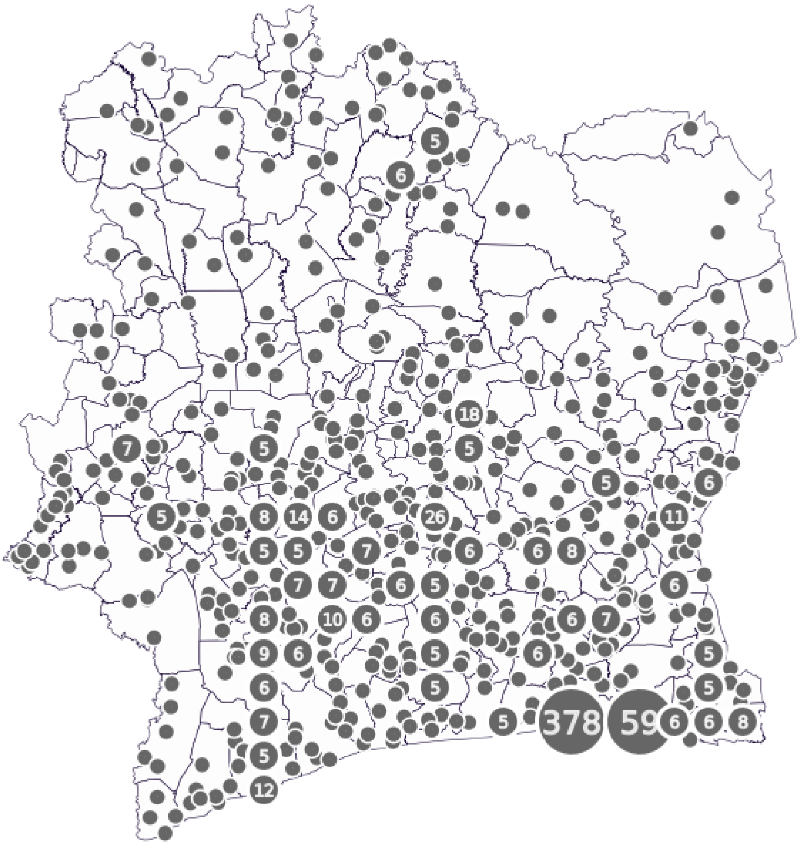

Figure A.1: Cell tower locations within subprefectures in Côte d'Ivoire.

in the D4D Challenge). Specifically, the dataset contains the total volume (number of calls and SMSs) and the duration of calls between each pair of cell towers over the entire period. The total number of cell towers is 1238. The exact locations of the towers are not provided by Orange, due to the company's operational confidentiality. However, they provide an approximate *latitude* and *longitude* location of each cell tower (obtained by adding noise to the actual location). The data consists of the hourly aggregated total traffic volume between cell towers, and hence no individual data are accessed.

Mobile phone penetration is high enough (95%) [1] in Côte d'Ivoire to make such a dataset sufficiently representative of the population. Moreover, the network operator holds a dominant position with 48% of the market share. The details of the data corpus of D4D are provided in Table A.2. While the D4D Challenge have shared 4 subsets of data, we focus our analysis on two of them, namely:

(1) Subset-1, which contains hourly aggregated data about the communication between cell towers with the following fields, (i) date, (ii) hour, (iii) initiating cell tower id, (iv) receiver cell tower id, (v) aggregated number of calls, and (vi) aggregated duration of calls;

(2) Subset-2, which contains information per customer, with the fields, (i) anonymized user id (ii) date (iii) hour (iv) initiating cell tower id.

We use Subset-1 to construct the communication network and Subset-2 to build the mobility network. We build both networks both at a city (Abidjan) and at a regional (sub-prefectures) level. The data to evaluate the model is richer at the city (Abidjan) level than at the sub-prefecture level. Hence, we mainly focus our analyses on Abidjan.

A.1.2 Telecom Italia Big Data Challenge

Specifically, the data used were provided by Telecom Italia Mobile (TIM), which is the largest mobile operator in Italy with 34% of the entire market share, during the TIM Big Data Challenge 2015 [4]. The data are coming from both TIM customers and roaming customers in the seven Italian cities.

The data collection period covers two months, from March 1, 2015 to April 31, 2015. The data collected cover seven Italian cities: Milan, Rome, Turin, Naples, Venice, Bari, and Palermo.

In order to keep users' privacy, the data are aggregated both spatially and temporally. For example, each city is divided into grids of different

[4] http://www.telecomitalia.com/tit/en/innovazione/archivio/bigdatachallenge2015.html

sizes, while temporal aggregation includes the call activities aggregated in 15-minutes time windows.

The schema of the data for Call is given as follows: *Timestamp - Grid ID - Number of Calls - Country Code - Duration of the calls.*

Appendix B

Emergency Event Lists from Côte d'Ivoire

In Table B.1 and B.2 we list emergency and social non-emergency events, gathered from a variety of public sources (e.g., Armed Conflict Location and Event Data Project, United Nations Council and International Crisis Group reports, local and international news). The cell towers IDs in the event region are given for the analyzed data period.

Date		Results MMPP		Results baseline		Cell T.	Emergency events	Social events
		E	S	E	S			
Dec. 11	7		X		X			Anniversary of death of Felix Houphouet Bougny
	11		X		X			New Parliament Selection
	15						Yopougon [63]	
	16						Yale, violent incident	
	17						Yale and Vavoua [63] violent incident	
	18						Vavoua [63]	
	25						Sikensi [63]	Christmas holiday
	26						Sikensi [63]	
Jan. 12	1							New Years Day
	4						Peite Guiglio	
	4						Boumi near Bouaké	
	5						Dobia	
	6						Toa Zeo near Dukou	
	6						Daloa	
	8							Baptisim of Lord Jesus fest
	9						Ayami	
	14							Arbeen Iman Huseyin fest
	15						Gagnoa	
	17							Annan and Clinton Visit
	18							Annan and Clinton Visit
	20						Touba	
	21						Abidjan Meeting	
	22						Konsou, Gnabra and Zedekan	CUP CIV & Sudan 17-18
	30		X		-			CUP CIV & Angola 19-20

Figure B.1: List of emergency and social non emergency events. MMPP and baseline results columns show the detected events with 'X' and the undetected ones with '-'. Specifically, we use 'E' for emergency events and 'S' for social non emergency events. For the emergency events not gathered from UN reports [6] we cite the source.

Date		Results MMPP		Results baseline		Cell Towers	Emergency events	Social events
		E	S	E	S			
Feb. 12	3	-		X		616, 336, 834, 964, 251	Bouaké-Katiola road	
	4		X		X			CUP CIV & Equar. Guina 20-21
	4		X		X			Mavlid an Nabi (Sunni)
	5		-		X			Day after prophet holiday
	8		-		-			CUP CIV & Mali 20-21
	9	X	X	-	-	398, 533, 426, 884, 534	Zibabo Yablo near Duekoue	Mavlid an Nabi (Shia)
	11	X		-		113, 16, 92	Arrah in eastern Cote d_Ivoire	
	12	X	X	X	X	113, 16, 92	Arrah in eastern Cote d_Ivoire	CUP CIV & Zambia 20:30-21:30
	13	X	X	X	X	113, 16, 92	Arrah in eastern Cote d_Ivoire	African cup recovery holiday
	16	X		-		764, 662, 698	Tueho village near Man	
	18	X		X		781, 1026, 27, 198, 436	Abidjan meeting	
	19	-		-		616, 336, 834, 964, 251	Bouaké-Katiola road	
	19	X		-		534, 1071, 645, 677	Westernpart of the country	
	21	-		-		872, 386, 1055	Zriglo in the border area with Liberia	
	22		-		-			Ash Wednesday festival
	26	X		X		239, 1154, 374	Bonon and Facobly election was attacked	
	27	X		-		546, 776, 916, 209, 27	Abidjan	
	29	X		X		285, 1128, 9, 819	Abobo	
	29	X		X		288, 963, 711	Séguéla	
Mar. 12	1	X		X		288, 963, 711	Séguéla	
	2					288, 963, 711	Séguéla	
	3	X		-		152, 980, 414	Daloa	
	5	-		-		42, 39	Agboville near Abidjan	
	8	X		-		616, 336, 834, 964, 251	Bouaké-Katiola road	
	11	X		-		895, 1034, 901, 410, 776, 306, 995	Assasination to Director General Police Academy, in Cocody 2 Plateaux	

Figure B.2: List of emergency and social non emergency events. MMPP and baseline results columns show the detected events with 'X' and the undetected ones with ''. Specifically, we use 'E' for emergency events and 'S' for social non emergency events. For the emergency events not gathered from UN reports [6] we cite the source.

Appendix C

Social Capital Auxiliary Files

Table C.1: Budgets of the 10 district (communes) of Abidjan in 1990, Source: Saint-Vil, J., Tableau de bord communal, Abidjan, DCGTx, 1991.

	Commune	**Budget, 1990 (CFA '000)**	**Revenue budget (CFA francs/person)**	**Capital budget (CFA francs/person)**
1	Abobo	734,524	1,405	4,262
2	Adjame	1,475,381	7,387	1,607
3	Attecoube	230,415	1,148	260
4	Cocody	969,471	5,038	2,494
5	Koumassi	731,787	2,420	744
6	Marcory	1,495,000	8,147	2,085
7	Plateau	2,213,876	164,200	25,633
8	Port-Bouet	1,496,000	7,099	1,767
9	Treichville	2,163,537	16,026	3,635
10	Yopougon	711,844	1,298	602

C.1 Descriptive statistics

Table C.2: Spatial distribution of settlements in Abidjan, left to right, poorer to better residential areas, source DGCTx, 1996 [221].

Hectar/%in region	Primitive Shelters	Slums	Shared Houses	Unshared Houses
Abobo	29.130	909.430	23.480	48.520
	2.88%	89.99%	2.32%	4.80%
Adjam	23.900	171.080	49.490	64.110
	7.74%	55.44%	16.04%	20.78%
Attcoub	37.810	251.090	5.270	49.820
	10.99%	73.01%	1.53%	14.49%
Cocody	72.830	94.430	143.780	846.930
	6.29%	8.15%	12.42%	73.14%
Koumassi	79.890	130.420	12.520	100.900
	24.68%	40.29%	3.87%	31.17%
Marcory	11.340	87.850	27.570	357.010
	2.34%	18.16%	5.70%	73.80%
Plateau	0.000	1.850	16.370	39.780
	0.00%	3.19%	28.22%	68.59%
Port-Bouet	385.310	28.140	20.700	153.580
	65.56%	4.79%	3.52%	26.13%
Treichville	2.960	65.830	23.110	45.640
	2.15%	47.87%	16.80%	33.19%
Yopougon	68.990	526.940	29.260	541.450
	5.91%	45.17%	2.51%	46.41%

Table C.3: Abidjan's communes participation in the 2011 elections.

	Commune	**# of Voters**	**Participation(%)**
1	Abobo	354,535	30.6
2	Adjame	102,423	35
3	Attecoube	82,234	30.4
4	Cocody	213,668	15.7
5	Koumassi	143,276	30.8
6	Marcory	83,748	20.6
7	Plateau	24,091	20.9
8	Port-Bouet	97,617	20.3
9	Treichville	46,986	36.8
10	Yopougon	450,717	19.5

Table C.4: People living with HIV(PLHIV) from PEPFAR[200] and confirmed malaria cases in Abidjan's communes from [210].

	PLHIV	Malaria
Abobo	23,506	71,437
Adjame	11,265	
Attecoube	7,879	35,714
Cocody	9,419	55,500
Koumassi	9,900	
Marcory	7,780	
Plateau	226	12,375
Port-Bouet	9,583	
Treichville	3,197	62,607
Yopougon	13,703	

Table C.5: Number of stolen cars in Abidjan city, From INS Annuaire des Statistiques Dmographiques et Sociales, 2009-2010.

Districts	2009	2010	Total
Abobo	82	46	128
Adjam	81	42	123
Attcoub	0	0	0
Cocody	198	142	340
Koumassi	0	0	0
Marcory	171	81	252
Plateau	0	0	0
Port-Bout	97	43	140
Treichville	0	0	0
Yopougon	164	74	238

Table C.6: Land Use Categories, (*) indicates taken for analysis.

Land Use Categories	**Definition**
1 informal settlement (*)	Housing areas which lack the legal status of the land
2 low density residential land	Residential areas where estimated population density is less than 70 persons/ha
3 medium density residential land	Residential areas where estimated population density is between 70 to 260 persons/ha
4 high density residential land	Residential areas where estimated population density is more than 260 persons/ha
5 industrial land (*)	Land areas for heavy chemical industry or factories, workshops, laboratories, car repair shops, tank storages
6 commercial/office land (*)	Land areas for commercial/service activities such as offices, markets, shopping centres, supermarkets, hotels, banks, petrol stations and other private services
7 health facility land (*)	Land areas for medical / health care such as hospital and clinic
8 educational land (*)	Land areas for education such as primary schools, high schools, universities, colleges, and other various schools
9 government office land	Land areas for the offices for local, national, other public organizations such as government offices, city halls, post offices, courthouses
10 sports and tourism facility (*)	Land areas for sports, gymnasiums, parks gardens, childrens playgrounds, green zones, tourist facilities, historic sites
11 transport facility land(*)	Land areas for transportation and storage such as stations, bus terminals track terminals, port facilities, airport facilities, car parks, depots, warehouses
12 security facility land (*)	Land areas for security facilities such as police stations, gendarmerie offices, military sites, prisons, fire stations
13 utilities land (*)	Land areas for public services such as water treatment plants, water tanks, sanitary facilities, waste disposal sites, final landfill sites, broadcast stations
14 cultural land (*)	Land areas for cultural and religious activities such as museum, art galleries, etc. garden, libraries, churches, mosques, temples
15 cemeteries	Large size and important cemeteries. Does not include a churchyard in a church.
16 forest	Forest area with native trees including original and secondary forest
17 grassland	Pasture land with sparse trees
18 agricultural land	Seasonal cultivated land for vegetables, grains and plantation area
19 riparian land	Swamp, Marsh area
20 others	No categorize items, non-utilized area, landslide and erosion, cliff

Table C.7: Descriptive statistics on degree centrality, based on the communication network.

	Degree	**Centrality**					
		Mean	**STD**	**Median**	**Mode**	**Min**	**Max**
1	Abobo	1200.2000	11.0774	1204	1205	1145	1205
2	Adjame	1195.5652	26.8867	1202	1202	1081	1202
3	Attecoube	1202.5385	3.7775	1203	1203	1194	1203
4	Cocody	1141.7889	206.9900	1198.5	1206	1	1206
5	Koumassi	1188.1176	24.6548	1199	1199	1109	1199
6	Marcory	1025.1034	262.4302	1109	1171	40	1171
7	Plateau	1001.5882	284.5240	1184	1196	200	1196
8	Port-Bouet	1067.7500	193.8588	1146	1146	376	1146
9	Treichville	1113.8500	131.0939	1145.5	1156	610	1156
10	Yopougon	1175.6974	147.7920	1206	1206	7	1206

Table C.8: Descriptive statistics on effective size, based on the communication network.

	Effective Size						
		Mean	**STD**	**Median**	**Mode**	**Min**	**Max**
1	Abobo	624.5552	39.4500	629.4597	472.3369	472.3369	680.0668
2	Adjame	609.0047	61.9073	621.2195	392.8463	392.8463	672.6013
3	Attecoube	619.9858	20.8695	618.6310	579.4393	579.4393	647.4868
4	Cocody	560.8717	119.0959	595.6323	19.4635	19.4635	660.9705
5	Koumassi	654.2057	14.3398	654.5719	631.0656	631.0656	677.1960
6	Marcory	573.9489	145.8651	626.9734	11.4110	11.4110	663.8848
7	Plateau	427.7789	205.7866	546.9742	26.6545	26.6545	646.7823
8	Port-Bouet	607.6660	111.2927	648.1647	185.6975	185.6975	681.6141
9	Treichville	612.7983	81.1506	641.0399	283.7057	283.7057	656.3072
10	Yopougon	599.4598	75.3283	621.9715	219.1734	219.1734	693.3558

Table C.9: Descriptive statistics on efficiency, based on the communication network.

	Efficiency						
		Mean	**STD**	**Median**	**Mode**	**Min**	**Max**
1	Abobo	0.5206	0.0292	0.5241	0.4129	0.4129	0.5625
2	Adjame	0.5090	0.0435	0.5175	0.3637	0.3637	0.5568
3	Attecoube	0.5160	0.0164	0.5147	0.4857	0.4857	0.5387
4	Cocody	0.4761	0.0759	0.4977	0.1436	0.1436	0.5463
5	Koumassi	0.5513	0.0177	0.5503	0.5276	0.5276	0.5887
6	Marcory	0.5521	0.0572	0.5682	0.2853	0.2853	0.5916
7	Plateau	0.3937	0.1269	0.4622	0.1339	0.1339	0.5354
8	Port-Bouet	0.5685	0.0244	0.5732	0.4952	0.4952	0.5959
9	Treichville	0.5496	0.0335	0.5620	0.4659	0.4659	0.5875
10	Yopougon	0.5019	0.0498	0.5146	0.3011	0.3011	0.5725

Table C.10: Descriptive statistics on local clustering coefficient, based on the communication network.

	Local	**Clustering**	**Coeff.**				
		Mean	**STD**	**Median**	**Mode**	**Min**	**Max**
1	Abobo	0.9458	0.0042	0.9448	0.9400	0.9400	0.9638
2	Adjame	0.9460	0.0070	0.9445	0.9387	0.9387	0.9676
3	Attecoube	0.9451	0.0023	0.9451	0.9413	0.9413	0.9503
4	Cocody	0.9483	0.0114	0.9450	0.9381	0.9381	0.9980
5	Koumassi	0.9465	0.0038	0.9459	0.9413	0.9413	0.9566
6	Marcory	0.9582	0.0128	0.9546	0.9411	0.9411	0.9923
7	Plateau	0.9610	0.0200	0.9498	0.9399	0.9399	0.9965
8	Port-Bouet	0.9565	0.0112	0.9540	0.9413	0.9413	0.9867
9	Treichville	0.9538	0.0098	0.9530	0.9405	0.9405	0.9809
10	Yopougon	0.9457	0.0078	0.9437	0.9377	0.9377	0.9844

Table C.11: Descriptive statistics on betweenness centrality, based on the communication network.

	Betweenness	**Centrality**					
		Mean	**STD**	**Median**	**Mode**	**Min**	**Max**
1	Abobo	1128.866	31.203	1133.893	987.517	987.517	1189.227
2	Adjame	1123.036	68.566	1134.741	851.771	851.771	1195.990
3	Attecoube	1147.459	68.040	1130.950	1104.335	1104.335	1370.012
4	Cocody	1054.691	227.319	1122.525	0.541	0.541	1175.043
5	Koumassi	1125.058	63.911	1139.323	911.252	911.252	1196.076
6	Marcory	841.657	333.818	959.499	0.074	0.074	1208.695
7	Plateau	789.809	427.570	1075.218	1.821	1.821	1160.059
8	Port-Bouet	896.124	305.849	1033.649	55.760	55.760	1181.863
9	Treichville	950.254	220.239	1001.399	189.344	189.344	1152.663
10	Yopougon	1143.794	188.827	1143.143	1174.250	286.681	2378.507

Table C.12: Descriptive statistics on degree centrality, based on the mobility network.

	Degree Centrality						
		Mean	**STD**	**Median**	**Mode**	**Min**	**Max**
1	Abobo	417.4889	77.5115	420	318	224	318
2	Adjame	533.1739	112.6534	497	471	271	471
3	Attecoube	466.3077	98.2356	455	340	340	340
4	Cocody	370.2584	96.8092	387	448	4	448
5	Koumassi	406.3529	79.4827	438	477	266	477
6	Marcory	285.6897	112.8827	274	371	5	371
7	Plateau	324.1765	170.7148	420	447	12	447
8	Port-Bouet	276	114.3112	265	36	36	36
9	Treichville	369.2500	111.0386	357	348	85	348
10	Yopougon	394.8800	106.6520	413	331	57	331

Table C.13: Descriptive statistics on effective size, based on the mobility network.

	Effective Size						
		Mean	**STD**	**Median**	**Mode**	**Min**	**Max**
1	Abobo	369.0861	75.8249	375.5944	184.2824	184.2824	542.8057
2	Adjame	474.5761	115.4581	435.3778	210.9621	210.9621	722.4065
3	Attecoube	412.8434	95.8663	398.4168	290.8184	290.8184	622.3746
4	Cocody	319.6934	88.6299	336.8421	2.0098	2.0098	437.3385
5	Koumassi	366.3059	72.6580	394.8219	235.1907	235.1907	464.3018
6	Marcory	256.2028	102.3382	251.9341	3.3112	3.3112	425.9249
7	Plateau	276.8271	156.1693	361.1846	7.6797	7.6797	453.7765
8	Port-Bouet	253.2924	104.8842	243.6615	31.4159	31.4159	444.7033
9	Treichville	331.5060	101.2939	323.2903	66.6192	66.6192	502.7404
10	Yopougon	350.8463	102.1479	371.5441	38.0114	38.0114	552.2725

Table C.14: Descriptive statistics on efficiency, based on the mobility network.

	Efficiency						
		Mean	**STD**	**Median**	**Mode**	**Min**	**Max**
1	Abobo	0.88066	0.02174	0.88092	0.81613	0.81613	0.91690
2	Adjame	0.88360	0.03352	0.87763	0.77846	0.77846	0.93334
3	Attecoube	0.88201	0.01798	0.87884	0.85535	0.85535	0.92204
4	Cocody	0.84963	0.06088	0.86694	0.50245	0.50245	0.88889
5	Koumassi	0.90097	0.00890	0.90292	0.88418	0.88418	0.92004
6	Marcory	0.88453	0.05031	0.89827	0.66224	0.66224	0.92348
7	Plateau	0.81130	0.08416	0.86089	0.63997	0.63997	0.89276
8	Port-Bouet	0.91572	0.01453	0.91819	0.87266	0.87266	0.93362
9	Treichville	0.89355	0.03030	0.90023	0.78375	0.78375	0.92374
10	Yopougon	0.88051	0.03786	0.89164	0.66687	0.66687	0.92546

Table C.15: Descriptive statistics on local clustering coefficient, based on the mobility network.

	Local clustering coeff.						
		Mean	**STD**	**Median**	**Mode**	**Min**	**Max**
1	Abobo	0.66559	0.09156	0.66483	0.46873	0.46873	0.86787
2	Adjame	0.55492	0.12910	0.57877	0.32386	0.32386	0.89550
3	Attecoube	0.62562	0.10985	0.64987	0.39461	0.39461	0.76729
4	Cocody	0.75589	0.08795	0.74844	0.60690	0.60690	1.00000
5	Koumassi	0.69004	0.07901	0.65119	0.57454	0.57454	0.83479
6	Marcory	0.80526	0.08263	0.82275	0.62668	0.62668	1.00000
7	Plateau	0.77635	0.13031	0.72754	0.58432	0.58432	0.95072
8	Port-Bouet	0.77236	0.08219	0.80957	0.57527	0.57527	0.85397
9	Treichville	0.73730	0.09869	0.76769	0.52780	0.52780	0.86499
10	Yopougon	0.68755	0.10680	0.67569	0.45786	0.45786	0.96742

Table C.16: Descriptive statistics on betweenness centrality, based on the mobility network.

	Betweenness Centrality						
		Mean	**STD**	**Median**	**Mode**	**Min**	**Max**
1	Abobo	2642.070	2259.401	1876.346	283.580	283.580	10144.262
2	Adjame	7693.997	8454.466	3099.174	49.721	49.721	35592.771
3	Attecoube	4076.686	5433.026	2104.138	965.325	965.325	21049.299
4	Cocody	904.282	795.863	737.297	0.000	0.000	3795.2082
5	Koumassi	1431.543	1029.157	1435.936	78.938	78.938	3303.143
6	Marcory	497.009	819.316	203.666	0.000	0.000	3254.942
7	Plateau	845.653	962.073	583.866	0.016	0.016	3913.121
8	Port-Bouet	861.108	1212.989	243.059	0.879	0.879	4312.045
9	Treichville	1067.158	1365.724	406.400	4.147	4.147	4875.294
10	Yopougon	2021.051	2263.924	1322.205	0.394	0.394	11101.363

Table C.17: Network Density for communication and mobility networks.

	Network	Density	
		Communication Network	**Mobility Network**
1	Abobo	2.96E-05	0.000484243
2	Adjame	7.74E-06	0.000124894
3	Attecoube	2.47E-06	3.86E-05
4	Cocody	0.000113705	0.001746538
5	Koumassi	4.23E-06	6.73E-05
6	Marcory	1.17E-05	0.000176336
7	Plateau	1.61E-05	0.000203787
8	Port Bouet	8.42E-06	0.000119206
9	Treichville	5.85E-06	8.98E-05
10	Yopougon	8.22E-05	0.001271446