УНИВЕРЗИТЕТ У НОВОМ САДУ

ПРИРОДНО МАТЕМАТИЧКИ ФАКУЛТЕТ
ДЕПАРТМАН ЗА ГЕОГРАФИЈУ, ТУРИЗАМ
И ХОТЕЛИЈЕРСТВО

ПРИМЕНА BIG DATA АНАЛИТИКЕ ЗА ИСТРАЖИВАЊЕ ПРОСТОРНО-ВРЕМЕНСКЕ ДИНАМИКЕ ЉУДСКЕ ПОПУЛАЦИЈЕ

ДОКТОРСКА ДИСЕРТАЦИЈА

BIG DATA ANALYSIS APPLIED IN SPACE-TIME HUMAN DYNAMICS RESEARCH

PHD THESIS

Ментор:
др Минучер Месарош
др Сања Брдар

Кандидат:
Оливера Мулић

Нови Сад, 2023. године

УНИВЕРЗИТЕТ У НОВОМ САДУ **ОБРАЗАЦ – 5а**
НАВЕСТИ НАЗИВ ФАКУЛТЕТА ИЛИ ЦЕНТРА

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА[1]

Врста рада:	Докторска дисертација
Име и презиме аутора:	Оливера Мулић
Ментор (титула, име, презиме, звање, институција)	Др. Минучер Месарош, ванредни професор, Природно-математички факултет Др. Сања Брдар, научни сарадник, Институт БиоСенс
Наслов рада:	Примена Big Data аналитике за истраживање просторно-временске динамике људске популације
Језик публикације (писмо):	Енглески (енглески алфабет)
Физички опис рада:	Унети број: Страница_____124________ Поглавља______8_________ Референци____140________ Табела________7_________ Слика________43_________ Графикона____0__________ Прилога______2__________
Научна област:	Техничко-технолошке науке
Ужа научна област (научна дисциплина):	Геонауке
Кључне речи / предметна одредница:	Big Data, подаци, динамика људске популације, дистрибуирано процесирање, теорија графова, географски простор, виртуелни простор, телекомуникације
Резиме на српском језику:	Са све већом и већом количином података која је доступна везано за динамику људске популације, постаје све више изазовно да се спроведе истраживање у овој области које би донело ново знање. У данашње време људи масовно живе у великим градовима где би знање о људској динамици, навикама и понашању могло значајно да унапреди организацију градова, енергетску ефикасност, транспорт и свеукупно квалитетнији и више одржив животни стил. Динамика људске популације може да се посматра са више аспеката, али сви они имају три заједничка елемента: време, простор и количину података. Људска активност и интеракције не могу се посматрати одвојено од просторне и временске компоненте јер се све дешава *негде* и у *неко време.* Такође, са великим присуством паметних телефона данас су доступни терабајти података о људској динамици. Иако су подаци осетљиви због приватности корисника, прави власници података су заправо телеком компаније, или компаније друштвених мрежа или неке друге компаније које развијају корисничке апликације за паметне телефоне. Ако би се такви подаци отварали за јавност или научну заједницу морали би прво

[1] Аутор докторске дисертације потписао је и приложио следеће Обрасце:
5б – Изјава о ауторству;
5в – Изјава о истоветности штампане и електронске верзије и о личним подацима;
5г – Изјава о коришћењу.
Ове Изјаве се чувају на факултету у штампаном и електронском облику и не кориче се са тезом.

	да буду анонимизовани. Други изазов везан за кориснички генерисане податке је величина података. Подаци су обично веома велики меморијски (енг. „Volume“), долазе из различитих извора и у различитим форматима (енг. „Variety“) и генерисани су реалном времену и мењају се веома брзо (енг. „Velocity“). Ово су три „V“ Великих података, и такви подаци захтевају посебан приступ аналитици са специјално дизајнираним алатима за Аналитику великих података. У оквиру истраживања које је презентовано у овој тези објединили смо Аналитику великих података, Теорију графова и просторно-временски зависне податке о људској динамици.
Датум прихватања теме од стране надлежног већа:	31.10.2019.
Датум одбране: (Попуњава одговарајућа служба)	
Чланови комисије: (титула, име, презиме, звање, институција)	Председник: Др. Миро Говедарица, редовни професор, Факултет Техничких Наука Члан: Др. Минучер Месарош, ванредни професор, Природно-математички факултет Члан: Др. Сања Брдар, научни сарадник, Институт БиоСенс Члан: Др. Даниела Арсеновић, ванредни професор, Природно-математички факултет Члан: Др. Данијела Тешендић, ванредни професор, Природно-математички факултет Члан: Апостолос Н. Пападопоулос, ванредни професор, Faculty of Sciences, Aristotle University of Thessaloniki
Напомена:	

UNIVERSITY OF NOVI SAD
FACULTY OR CENTER

KEY WORD DOCUMENTATION[2]

Document type:	Doctoral dissertation
Author:	Olivera Mulić
Supervisor (title, first name, last name, position, institution)	Dr. Minučer Mesaroš, associate professor, Faculty of Sciences Dr. Sanja Brdar, research associate, BioSense Institute
Thesis title:	BigData analysis applied in space-time human dynamics research
Language of text (script):	English language (english alphabet)
Physical description:	Number of: Pages______124_________ Chapters_____8__________ References____140_______ Tables_________7_______ Illustrations_____43_______ Graphs_______0______ Appendices____2_________
Scientific field:	Technical and technological sciences
Scientific subfield (scientific discipline):	Geosciences
Subject, Key words:	BigData, data, human dynamics, distributed computing, graph theory, geographical space, virtual space, telecommunications
Abstract in English language:	With the rapid growth of the volume of available data related to human dynamics, it became more challenging to research and investigate topics that could reveal novel knowledge in the area. In present time people tend to live mostly in large cities, where knowledge about human dynamics, habits and behaviour could lead to better city organisation, energy efficiency, transport organisation and overall better quality and more sustainable living. Human dynamics could be reasoned from many different aspects, but all of them have three elements in common: time, space and data volume. Human activity and interaction could not be inspected without space and time component because everything is happening somewhere at some time. Also, with huge smartphone adoption now terabytes of data related to human dynamic are available. Although data is sensitive to personal information, true owners of the data is either telecom operator company, social media company or any other company that provides the applications that are used on the mobile phone. If such data is to be opened to public or scientific community to conduct a research with it, it needs to be anonimized first.Another challenge of user generated data is data set volume. Data is usually very large in size (Volume), it comes from different sources and in different formats (Variety)

[2] The author of doctoral dissertation has signed the following Statements:
5б – Statement on the authority,
5в – Statement that the printed and e-version of doctoral dissertation are identical and about personal data,
5г – Statement on copyright licenses.
The paper and e-versions of Statements are held at he faculty and are not included into the printed thesis.

	and it is generated in real-time and it evolves very fast (Velocity). These are three V's of Big Data, and such data sets need to be approached with specially designed Big Data technologies.In the research presented in this thesis we assembled Big Data technologies, Graph Theory and space-time dependent human dynamic data.
Accepted on Scientific Board on:	31.10.2019.
Defended: (Filled by the faculty service)	
Thesis Defend Board: (title, first name, last name, position, institution)	President: Dr. Miro Govedarica, full professor, Faculty of Technical Sciences Member: Dr. Minučer Mesaroš, associate professor, Faculty of Sciences Member: Dr. Sanja Brdar, research associate, BioSense Institute Member: Dr. Daniela Arsenović, associate professor, Faculty of Sciences Member: Dr. Danijela Tešendić, associate professor, Faculty of Sciences Member: Dr. Apostolos N. Papadopoulos, associate professor, Faculty of Sciences, Aristotle University of Thessaloniki
Note:	

Big Data Analysis Applied in Space-Time Human Dynamics Research

Olivera Mulić

Faculty of Sciences
University of Novi Sad

This dissertation is submitted for the degree of
Doctor of Philosophy (PhD)

February 2023

I would like to dedicate this thesis to my grandmother, who held education in high regard, more than anybody I know. She toughed me that the only appreciation we can seek in this life is through knowledge we harvest in our minds and kindness we cherish in our hearts.

For my dear grandmother,
for Knowledge and Love

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 50 figures.

Olivera Mulić
February 2023

Acknowledgements

Firstly, as my PhD journey started at BioSense Institute in Novi Sad, I would like to express my sincere gratitude to professor Vladimir Crnojević who recognised my potential for scientific research and who gave me one great advice at the very beginning of my PhD research, when I asked him what topic should I explore during PhD he told me: *Explore whatever is the most interesting to you. In the end, only your sincere interest and love for the topic will get you through it.* I would also like to thank other colleagues from BioSense who made my years of academic career memorable.

Next, I would specially like to thank professor Apostolos N. Papadopoulos from Aristotle University of Thessaloniki for his wholehearted help and support all the way through my PhD, from very beginning to the end. I would also like to thank whole DELAB for warm welcoming me during my research visits and for making me feel like at home. I am grateful for the opportunity to be a part of COST Action IC1406 which supported my collaboration with Aristotle University through grants for short term visits.

Furthermore, I wish to express my gratitude to my supervisors Dr. Minučer Mesaroš and Dr. Sanja Brdar for all their advice, guidelines, patience and commitment.

Finally, I am grateful to my family, my parents Nebojša and Biljana and my sister Jelena for their love and understanding. Very special thanks to my dear husband Mihajlo for all his support, understanding, advice, for deeply caring for my well being and for being so proud of me. Mihajlo, I couldn't do this without you.

Olivera Mulić,
in Novi Sad, February 2023

Abstract

With the rapid growth of the volume of available data related to human dynamics, it became more challenging to research and investigate topics that could reveal novel knowledge in the area. In present time people tend to live mostly in large cities, where knowledge about human dynamics, habits and behaviour could lead to better city organisation, energy efficiency, transport organisation and overall better quality and more sustainable living. Human dynamics could be reasoned from many different aspects, but all of them have three elements in common: time, space and data volume. Human activity and interaction could not be inspected without space and time component because everything is happening *somewhere* at some *time*. Also, with huge smartphone adoption now terabytes of data related to human dynamic are available. Although data is sensitive to personal information, true owners of the data is either telecom operator company, social media company or any other company that provides the applications that are used on the mobile phone. If such data is to be opened to public or scientific community to conduct a research with it, it needs to be anonimized first. Another challenge of user generated data is data set volume. Data is usually very large in size (Volume), it comes from different sources and in different formats (Variety) and it is generated in real-time and it evolves very fast (Velocity). These are three V's of Big Data, and such data sets need to be approached with specially designed Big Data technologies. In the research presented in this thesis we assembled Big Data technologies, Graph Theory and space-time dependent human dynamic data.

Апстракт

Са све већом и већом количином података која је доступна везано за динамику људске популације, постаје све више изазовно да се спроведе истраживање у овој области које би донело ново знање. У данашње време људи масовно живе у великим градовима где би знање о људској динамици, навикама и понашању могло значајно да унапреди организацију градова, енергетску ефикасност, транспорт и свеукупно квалитетнији и више одржив животни стил. Динамика људске популације може да се посматра са више аспеката, али сви они имају три заједничка елемента: време, простор и количину података. Људска активност и интеракције не могу се посматрати одвојено од просторне и временске компоненте јер се све дешава *негде* и у *неко време.* Такође, са великим присуством паметних телефона данас су доступни терабајти података о људској динамици. Иако су подаци осетљиви због приватности корисника, прави власници података су заправо телеком компаније, или компаније друштвених мрежа или неке друге компаније које развијају корисничке апликације за паметне телефоне. Ако би се такви подаци отварали за јавност или научну заједницу морали би прво да буду анонимизовани. Други изазов везан за кориснички генерисане податке је величина података. Подаци су обично веома велики меморијски (енг. „Volume“), долазе из различитих извора и у различитим форматима (енг. „Variety“) и генерисани су реалном времену и мењају се веома брзо (енг. „Velocity“). Ово су три „V“ Великих података, и такви подаци захтевају посебан приступ аналитици са специјално дизајнираним алатима за Аналитику великих података. У оквиру истраживања које је презентовано у овој тези објединили смо Аналитику великих података, Теорију графова и просторно-временски зависне податке о људској динамици.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

"Everything is related to everything else, but near things are more related than distant things."

The first law of geography, Waldo Tobler

From early stages of human development the question of “where” something is happening or existing emphasised the importance of local neighbourhood in human lives. The richer the local neighbourhood were, people living in there were more likely to survive and to develop in wider context. People always gathered in groups because that ensured access to resources, but with the development of society, people tend to connect even beyond they initial “group” such as village or city, they wanted to connect to other “groups” and exchange resources and knowledge [92]. So, grouping, gathering and connecting came as a natural characteristic of human population. Human dynamics is therefore reflected through mobility, connectivity and grouping. With technological advancements modern life became more dynamic in every possible way, which made the human dynamic research even more complex. Lee et al. [70] even imply that modern technology is so deeply embedded in our lives that it is affecting not only the forms of existence, but also the forms of thinking. With the complexity and new technologies came new solutions as well, because now we have available immense amount of data related to human mobility and overall behaviour.

Today, we have almost 80% of the population living in urban areas [100], where people are constantly interacting with technology, space, services etc. While interacting with technology using smart phones, digital footprints of human behaviour are collected and stored. Location information, type of activity, duration, personal data, etc. are collected in those digital footprints, and while in most of the cases that data is used by companies to provide, enable and enhance their services, it could be used also to inspect other ways of human dynamics that were previously unable to be observed. While the idea of inspecting human dynamics using data about their activity is not novel, traditional methods of data collection such as counts and surveys are unable to provide insights on wider human dynamics, whereas the data from mobile phones collected in digital footprints could fill up that gap and provide sufficient information [55].

In this research, we use connectivity and mobility data to explore the novel aspects of human dynamics in urban spaces. We use modern Big data technologies to exploit the full value of the data. Our five main hypotheses are:

1. Patterns in human dynamics and behaviour are strongly correlated with day time and day type distinguishing between hours within a day, but also between working days and non working days.

2. Patterns in human dynamics are geographically mapped and clustered within spatial units with similar location semantic.

3. Human dynamics and interactions, represented through connectivity and mobility networks, can be further analyzed by means of graph theory and clustering to extract informative properties at the level of cities and its units.

4. Land use type and location functions strongly impact the prediction of connectivity and mobility properties that are correlated with human dynamics.

5. Big data technologies need to be utilized for designing more efficient data processing workflows.

We defined main hypotheses of this thesis by observing interconnections of human dynamics and spatial location semantic in urban areas through data, we further extended the related findings from other authors, and we fortified our assertions with experimental results. Relation between human activity in urban areas and day time is observed before by analysing traffic mobility data [60]. Cottineau et al. [37] and Calabrese et al. [28] investigated how mobile phone data could be used to estimate socio-econimic organisation of cities and to derive indicators for urban sensing. Extensive work was done by Brdar et al. [23] to investigate country wide telecom connectivity and mobility data and to extract meaningful knowledge from the data using Network science and Graph theory. Release of a new, content rich mobile phone data set from Telecom Italia [13] encouraged us to dig deeper into the area of mobile phone data analytics with the aim to unveil novel knowledge and methods to analyse human dynamics.

We observe patterns in human dynamics depending on day period and day type such as peak hours, working and non working hours, night hours, weekday, weekend and holiday. Locations with different semantic are related to different human behaviour and different digital footprints. As human dynamics evolves through time, the connectivity/mobility network also evolves and network properties could be used to quantify the human dynamics.

1.1 Data sources

Smartphone producers and telecom companies estimated that in the 2022 more than 83% of the worlds population would own a smartphone. Massive usage of smartphones in our every day lives led to even greater collection of data related to our activities. Smartphones collect data about our activities and location through applications that we use and through telecom network. That data is sensitive due to GDPR[1] because it contains personal information, and companies are obliged to preserve it with most precaution because the leakage of such data would lead to user privacy violation.

Industry-like telecommunications which include mobile service providers exist for a long time now and therefore their work is conditioned with strict legal regulation. Telecom companies are obligated to preserve user privacy while providing their services which includes personal and location data collection. Therefore, telecom companies are not allowed to share the data they collect with third parties, not even for research purpose. Despite those constrains, research groups together with some telecom companies made and effort to enable the research on telecom data while preserving the privacy by introducing strict non-disclosure agreements before sharing the data. Also, there are examples of special data sets that are opened for public access but before releasing those data sets were exposed to anonymization procedure.

Unlike telecom industry, internet services and applications are relatively new and evolving rapidly, so the legal regulations there are less strict and not well-defined. Usually, internet based applications that are collecting some personal data, such as usernames, locations, credit card numbers, etc. are obliged only to collect user consent prior to providing their services, but since the absence of such consent would lead to disablement of service, users are left without real choice. Even when they get the user consent to collect their personal data the companies are still obligated to keep those data private because otherwise they might loose a lot of customers and users. Similar like with telecom companies, internet based companies that would like to share their data for research purpose must follow strict non-disclosure procedure and apply data anonymization before releasing.

[1] GDPR regulative in Europe https://gdpr-info.eu/.

Unlike telecom and internet user generated data, geographical data tends to get more opened and more accessible to wider public than before. Good example of this is remote sensing data which were 10, 20 years ago very difficult to obtain free of charge, but today are accessible more than ever even in reasonably high spatial resolution (10+ meters). Many publicly available products were derived from remote sensing data, such as Copernicus Urban Atlas with detailed land use classes, OpenStreetMap products, etc. Despite the trend to make the geographical data more accessible and more available, companies like Google and ESRI are still not very likely to open their data for free. They charge high rates for their data APIs such as Google's Geocoding API, which is the most detailed product of such kind on the market at this point. Large companies like Google and ESRI are not very interested in sharing their data products for social good research, because they already have R&D departments inside their companies, where they explore the data with the aim to develop new products and services for the market. At this point, scientific community is mainly left with open data sets to conduct the research, or in some special cases the University can have special agreement with national telecom provider or any other state institution to use their data for research and non-comercial purpose.

Another great challenge with data sets such as user generated telecom data and large scale geographical data is the computational power and resources. Such data is usually very large in size and very complex in relationships it preserves, which puts another challenge for the research community to address those obstacles using advanced Big Data and HPC technologies.

1.1.1 Mobile phone data

Mobile phone service providers collect large amount of data to monitor user interactions. Each time a user is using a mobile device (for sending an SMS or performing a call), a Call Detail Record (CDR) is created in the database of the service provider. CDRs are primarily made for billing purpose, but their value exceeds much that cause since they contain great information about user behaviour. Telecom user generated data is private, the owners of the data are telecom companies who are providing services to the customers and they are obligated to keep high level of security to the data in order to preserve users privacy. Mainly because of the sensitive nature of the data and strict regulations rules that telecom companies need to address to obtain and keep their licences, not many telecom operators are willing to share the data even for social good research purpose. In recent times there were some initiatives to open the data for research and in that case telecom operators are instructed to follow rigorous procedures for data anonymization to preserve privacy such that anonymized records cannot be linked to subscribers under any normal circumstances. Furthermore, before releasing any data to third parties, data sets are usually aggregated on temporal and/or spatial scales. For example, the numbers of calls as well as the duration of calls between any pair of antennas are aggregated hourly and movement trajectories are provided with reduced spatial resolution [3]. Differential privacy paradigm adds noise to original data up to the level not affecting the statistics significantly to preserve users' privacy. Another approach, suggested by the Open Algorithms (OPAL) initiative, proposes moving the algorithm to the data [71]. In their model, raw data are never exposed to outside parties, only vetted algorithms run on telecom companies' servers. Telecom data typically include spatial and temporal parameters to map device **activity, connectivity** and **mobility**. Here are described some examples of good practice when releasing the data to the third parties for research purpose.

Telecom Italia have opened some data for the **BigData challenge** organized in 2014 [13]. They aggregated telecom data for time period of two months, 1st November 2013 to 1st January 2014, for city of Milan and Province of Trentino. Spatial aggregation is done to the level of regular grid, where the area of Milan is composed of a grid overlay of 1,000 (squares with size of about 235 × 235 meters) and Trentino is composed of a grid overlay of 6,575 squares. This grid is projected with the WGS84 (EPSG:4326). Temporal aggregation is done in 10 min resolution. There are many types of CDRs and Telecom Italia has recorded the following activities:

- Received SMS a CDR is generated each time a user receives an SMS
- Sent SMS a CDR is generated each time a user sends an SMS
- Incoming Call a CDR is generated each time a user receives a call
- Outgoing Call a CDR is generated each time a user issues a call
- Internet a CDR is generated each time a user starts an Internet connection or ends an Internet connection. During the same connection a CDR is generated if the connection lasts for more than 15 min or the user transferred more than 5 MB.

They provided two type of data sets, one referring to telecom **activity** and one referring to telecom **connectivity**. The first type of data represents the activity of Trentino and Milan, showing all the aforementioned telecommunication events which took place within these areas. The data provides information of Telecom Italia's customers interacting with the network and of other people using it while roaming. The example of telecom activity data is presented in Figure 1.1. Areas with higher activity are presented in lighter color, which unambiguously revels urbanized zones and main transport roads.

The second type of data represents connectivity. Two CDR data sets were released which represent the interaction intensity between different locations: one from a particular area

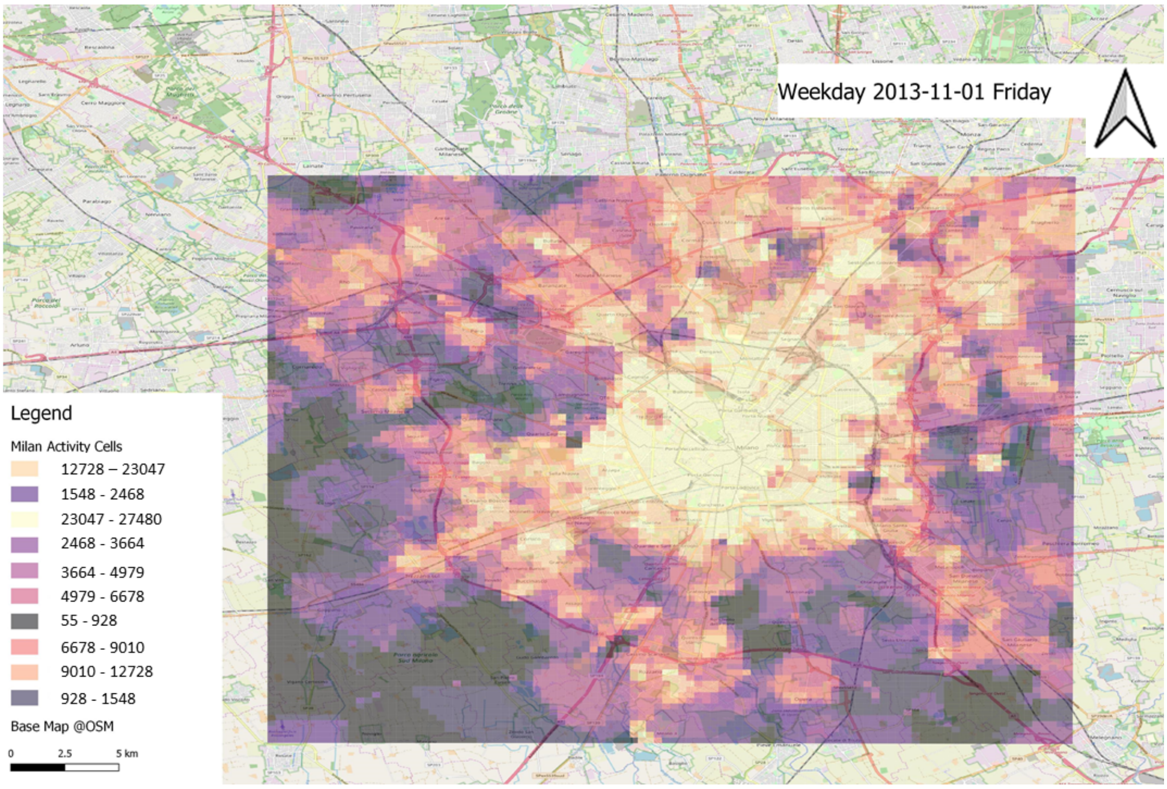


Fig. 1.1 Telecom activity over the city of Milan

(Trentino/Milan) to any of the Italian provinces and one quantifying the interactions within the city/province (e.g., Milan to Milan). Since Telecom Italia only possesses the data of its own customers, the computed interactions are only between them. This means that (at most) 34% of population's data is collected, due to Telecom Italia's market share.

The Orange, French telecom provider that operates worldwide has organized **"Data for Development" (D4D) challenge** and opened some CDR data of Orange's mobile phone users in Ivory Coast in 2013 [19]. The goal of the challenge is to help address society development questions in novel ways by contributing to the socio-economic development and well-being of the Ivory Coast population. The data sets are based on anonymized CDRs of phone calls and SMS exchanges between five million of Orange's customers in Ivory Coast between December 1, 2011 and April 28, 2012. The data is released in the form of four different data sets, where each of them represent different **connectivity** and **mobility** patterns. To preserve the privacy of the users different aggregation methods are performed on data, so released data sets are either temporally or spatially aggregated or random set of users is selected for specific data set. First data set represent antenna to antenna telecom traffic aggregated on hourly basis. When the aggregation is performed all personal information is lost from the data, leaving just information about hourly amount of traffic that occurred between each separate pair of antennas. This data can reveal much about overall pulse of the city, and how different parts of the city perform during the day in regards to people flow and activity [82]. In the literature this type of data is referred as **connectivity** data. Second data set contains information about individual trajectories for 50 000 users for two week time windows with precise antenna location information. The privacy is preserved by changing sample of 50 000 users each two weeks, so that single person could not be present in the period longer than two weeks. Third data set contains information about individual trajectories for 500 000 users over the entire observation period with sub-prefecture location information. Privacy is preserved by revealing only spatially aggregated trajectories between sub-prefectures, without information about the precise antenna that registered the communication. In the literature the data that is related to users' trajectories is referred to as **mobility** data. Fourth data set contains a sample of communication ego graphs for 5 000 randomly selected users. For these users, communications within their second order neighborhood have been divided into periods of two weeks spanning the entire observation period. For constructing an ego centered graph, one consider first and second order neighbors of the ego and communications between all individuals. The privacy is preserved by assigning to each selected user depersonalized identifier that are identical for all time slots but are unique for each subgraph. That is, a user who is part of the communication graph of two different users has a different identifier in the two graphs. In the data set is present a total of 5 000 connected graphs in every time period.

The response to the first D4D challenge has been overwhelming with over 80 research teams around the world submitting research projects using the mobile network data and correlating it against other localized or international data sets to tackle some of the biggest development challenges. With such encouraging outcome The Orange telecom decided to organize next **D4D challenge in the year 2014 for Senegal**, where they are also present in the market. They opened the data sets based on CDRs of phone calls and text exchanges between more than 9 million of Orange's users in Senegal between January 1, 2013 to December 31, 2013 [41]. They performed similar aggregation methods like in the 2013 D4D challenge to preserve the privacy of the users and released three data sets. First data set contains aggregated antenna to antenna traffic for 1666 antennas on an hourly basis. Second data set contains information about users mobility on a rolling 2 week basis for a whole year period with behavioral indicators at individual level for about 300 000 randomly sampled users. Third data set contains one year of aggregated mobility data at arrondissement level with behavioral indicators at individual level for about 150 000 randomly sampled users. Similar like in the first D4D challenge, to preserve the privacy of the users while keeping the valuable information in the data The Orange telecom had to carefully tune the relation between temporal and spatial aggregation and selection of random users for limited time period.

Mobile phone data proved to be very valuable source of information about people displacement when dealing with global crisis following natural disasters events like floods, earthquakes or epidemics, or humanitarian crises caused by war or political instability. After the Syrian Civil War started in 2011-12, civilians in increasing numbers sought refuge in neighboring countries. By May 2017, Turkey had received over 3 million refugees — the largest refugee population in the world. About 30% of them live in government-run camps near the Syrian border. Many have moved to cities looking for work and better living conditions. They face problems of integration, income, welfare, employment, health, education, language, social tension, and discrimination [102]. In 2017 Türk Telekom opened some data sets for the non-profit project **"The Data for Refugees (D4R) Challenge"** with the aim to ultimately improve the conditions of the Syrian refugees in Turkey by providing a special database to scientific community for enabling research on some urgent problems. The data sets are based on anonymized mobile CDRs of phone calls and SMS exchange between Türk Telekom users who are flagged as "refugee" for one year period. Following the examples of previous D4D challenges, Türk Telekom performed spatial and temporal aggregation to anonymize the data before sharing it to the third parties. They provided three types of data sets. First data set represent antenna to antenna aggregated telecom traffic on hourly basis. Total number and duration of calls, number of SMS exchanged, as well as any personal identifier is hidden during aggregation process. Second data set contains information about fine grained mobility of randomly selected group of users observed for a period of two weeks. After two weeks a fresh sample of active users are drawn at random. The data is provided for the whole one year sampling period. The users are represented by random numbers in the data set, and no personal information is stored. To protect privacy, new random identifiers are chosen for every two-week period, and if a user is sampled in more than one period, these records cannot be associated with each other. Third data set contains information about mobility of randomly selected subset of users for the entire one-year period, but with the reduced spatial resolution. The entire country is divided into the electoral prefectures, and for each call record, only the prefecture information is provided. The users identifiers are randomly assigned.

When data challenges are organized, the data is provided for participants under special rules, and usually only selected groups get the opportunity to access the data and to work with it. Data challenges are global initiatives where research groups world wide can participate. In some cases, the data set stays open even after the challenge has finished. Besides data challenges there are other models that could be used by Telecom operators to share the data for research purpose. Telecom operators can cooperate with research groups under special non-disclosure agreements, where the research group gain privileged rights that cannot be transferred, to work with the data. Mobile phone data can reveal the approximate location of a user and its mobility trace based on geographical location of the cell tower which registered the traffic. In [40] the authors proposed a novel computational framework that enables efficient and extensible discovery of mobility intelligence from large-scale spatial-temporal data such as CDR, GPS and Location-Based Services data. In [57] the authors focus on usage of CDRs in the context of mobility, transport and transport infrastructure analysis. They analyzed CDR data associated with cell towers together with Open Street Map road network to estimate users mobility. CDR data can provide a generalized view of users' mobility, since data is collected only when the telecom traffic happens. To illustrate mobility data set we created Figure 1.2 that presents a map with mobility traces across the city of Novi Sad on 3rd July 2017, for the time interval between 6am and 12pm extracted from raw CDR data through aggregation of visited locations' sequences of anonymous users. Data originate from the Serbian national operator, Telekom Srbija, released under non-disclosure agreement. From mobility traces we can detect a few locations in the city that acts as trajectory hubs. To preserve the privacy of users, Telekom Srbija selected random sample of users, assigned them depersonalized identifiers and changed the random sample each day for the observed time period from 3rd July 2017 to 11th July 2017. Also, only selected cell towers are present in the data. Even with such a rigid data preprocessing, a small group of anonymized users is present over entire period (about 2000 users).

With the pervasive adoption of smartphones in modern societies, in addition to CDRs, there is now a growing interest in xDRs, Extended Data Records. They enclose information on visited web sites, used applications, executed transactions, etc. Coupled with cell tower triangulation, applications can infer fine-grain phone locations [61], thus making data volumes even larger.

Along with CDRs and xDRs there are other types of mobile phone data that can reveal much about user behaviour. In 2012 Nokia Research Center organized **Mobile Data Challenge (MDC)**, a large-scale research initiative aimed at generating innovations around smartphone-based research, as well as community-based evaluation of related mobile data analysis methodologies [69]. For the purpose of MDC they provided large scale data (more than 80GB of data) of diverse types such as location data (GPS, WLAN), motion (accelerometer), proximity (Bluetooth), communication (phone call and SMS logs), multimedia (camera, media player), and application usage (user-downloaded applications in addition to system ones) and audio environment. To preserve the privacy of the users they performed various anonymization techniques like truncation for location data, hashing of phone numbers,

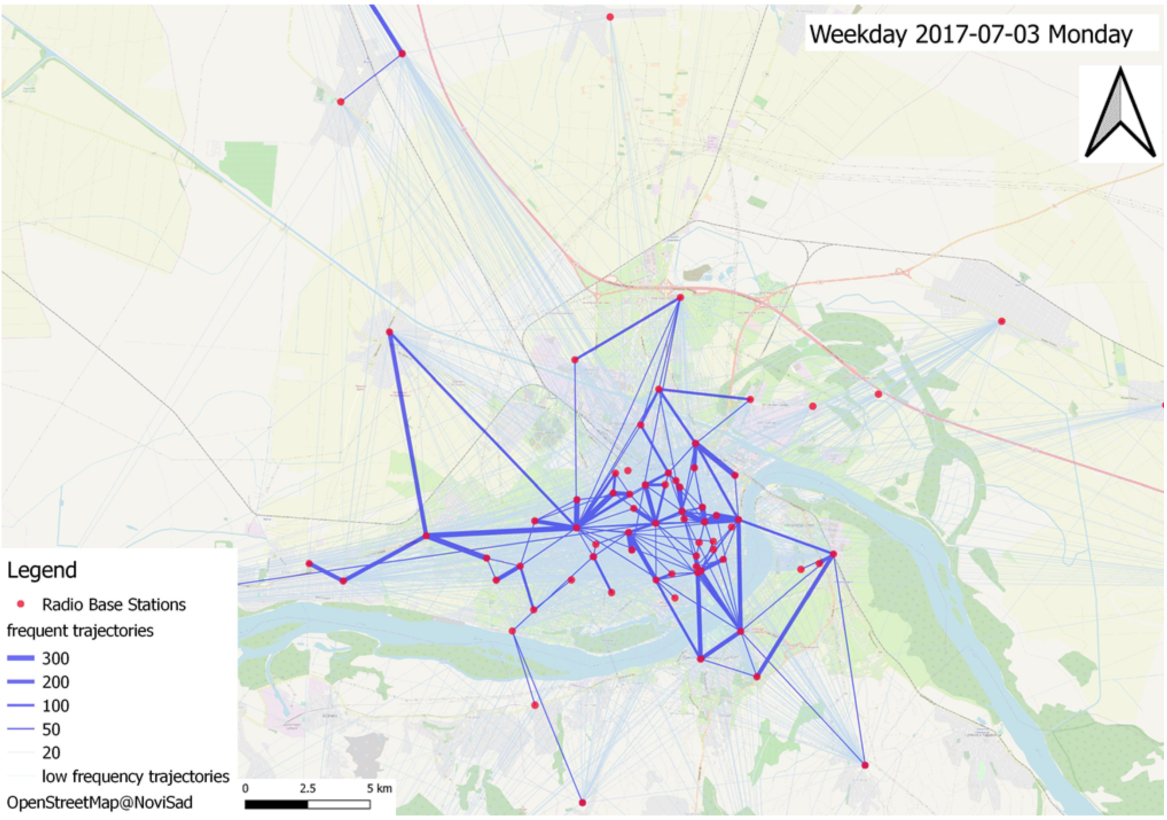


Fig. 1.2 Frequent trajectories from CDR data over the city of Novi Sad

names and MAC addresses. Also the cell ID and the location area code (LAC) of the cellular networks were anonymized using the hashing technique. The challenge was received enthusiastically by the research community with total number of 108 worldwide submissions.

The outcomes of global initiatives to expose mobile phone data in privacy preserving manner to scientific community proved that such data sets have unprecedented value in human dynamic research and computational social science. Examples of good data sharing practices show that with carefully designed and applied anonymization techniques privacy of the users is guaranteed while the data still has great research potential.

1.1.2 Location-based social media data

Social media has become an essential element of our every day life, for many years now. Most of the time we are using social media on our mobile devices, and most of those platforms require users location to provide, customize or enhance the content and services they are providing. Thus, social media companies became a key player in delivering location-based services since they know where the users are, they know what are they looking for, and they have a channel to through which they deliver services or customized content. Social media companies use the vast amount of data they collect on users location and interest to provide business to business services for other companies, but their data is usually not publicly available. Some companies however, like Twitter and Foursquare, share limited amount of data on demand through open source API. Many studies proved that location-based social media is valuable data source when conducting studies related to urban environment, human dynamics, social computing, etc.

Twitter provides access to their public data through API. Application programming interface, aka. API, is the way to access some platform programmatically, software to software. When someone wants to access Twitters data, they are required to register an application. This way, public tweets and replies could be harvested from Twitter platform and be used further for sentiment analysis, marketing analysis, etc. Hawelka et al. [58] used one full year of geo-located tweets all over the world between Januray 1, 2012 and December 31, 2012 to analyze human mobility. As a result of the study they were able to compare mobility profiles of countries worldwide. Another study showed how can we use Twitter data to understand human mobility [64]. Jurdak et al. analysed sequences of geo-tagged tweets to characterise the movement patterns of individuals. They have discovered that human mobility patterns extracted from geo-tagged tweets have similar overall features as observed in mobile phone records. Zagheni et al. [136] also used Twitter data to analyse human mobility with the focus on international and internal migration patterns.

Besides Twitter, another popular Location-Based Social Media platform is Foursquare. Foursquare is a social network application founded in 2009. In the application, users are able to notify their friends about their current location through check-ins for which they can receive virtual rewards. Apart for that, it allows users to a leave note about their experience in the specific venue, which can be utilized for building a recommendation system. With its initiatives to open some of data they collect, Foursquare attracted researches to explore their rich source of information and evaluate its potential for understanding social behaviour, mobility, and propose location intelligence services. Preoţiuc-Pietro et al. [93] used Foursquare data to cluster users based on their behaviour and to predict users' future movement. Cranshaw et al. [38] conducted the study to cluster the city zones based on the diverse usage of urban area evaluated through Foursquare data. Noulas et al. used Foursquare data to analyse and to predict the next venue user will visit [81]. The greatest advantage of using Location-Based Social Media and application data is:

- (a) there are no privacy concerns (users either agreed to privacy policy or they are using social media services in public mode),

- (b) the data have worldwide coverage, depending on the penetration within the nations of specific social media platform.

Despite this great advantages, there are some obstacles too. First is that the data is in the ownership of the companies, and only limited amount of data could be accessed without charges, or non at all. Second obstacle is that the data is usually biased towards younger population who is using the social media to a greater extent. Considering rapid growth on the market of Location-Based Social Media services we can anticipate that in the future this will be even more valuable data source in human behaviour and mobility studies.

1.1.3 Geographical open data

Data about geographical space was always challenging to collect and to visualize. Since the time being people had a need to know and to understand the geographical space around them, because the topology, the climate and natural resources meant life for living world including people. In the modern time, when life is more concentrated in the urban areas, people have even greater need for space and resources management. Remote sensing evolved to be the major discipline in Geosciences, with focus on observing the Earth surface from distance. Remote sensing is the set of methods and technologies for detecting and monitoring the physical characteristics of an area by measuring its reflected and emitted radiation at a distance (typically from satellite or aircraft). Special sensors collect remotely sensed images, which are further developed into various data products.

In order to better understand and to observe changes that are happening to the physical geographical space all over the world, many countries are developing space programs to monitor the Earth from satellite. European Space Agency (ESA) initiated the one of the most ambitious Earth observation program at present time - Copernicus, with the aim to provide accurate, timely and easily accessible information to improve the management of the environment, understand and mitigate the effects of climate change and ensure civil security. ESA is developing a new family of satellites, called Sentinels, specifically for the operational needs of the Copernicus program. The Sentinels are equipped with diverse optical and radar sensors to monitor the weather, deliver day and night snapshots of the Earth and high-resolution optical images for land services, provide the data for the ocean, coast and atmospheric monitoring.

Copernicus services are focused on six main categories: land management, the marine environment, atmosphere, emergency response, security and climate change. Copernicus Land Monitoring Service (CLMS) is a part of Copernicus program and it provides geographical information on land cover [2]. One of the CLMS products are data sets for Land cover and land use mapping. For the scope of this thesis, category Land cover and land use mapping will be further described.

CLMS produces land cover / land use (LC/LU) information in the CORINE Land Cover data, High Resolution Layers, Biophysical parameters and European Ground Motion Service. The CORINE Land Cover is provided for 1990, 2000, 2006, 2012, and 2018. It is vector based data set which includes 44 land cover and land use classes. The CORINE Land Cover data also includes a land-change layer, highlighting changes in land cover and land use for the observed 6 to 10 year period. Besides vector data there are high resolution raster data sets available which provides information about different land cover characteristics.

Derived from CORINE Land Cover there are four categories of thematic data, Urban Atlas, Riparian Zones, Natura 2000 and Costal Zones. For the scope of this thesis, the Urban Atlas data set is used.

Besides data from major satellite providers, there are efforts to collect data by community, which would be open both for users and contributors. OpenStreetMap is global open source project based on the community of mappers that contribute and maintain data about roads, trails, restaurants, railway stations and much more, all over the world [3]. OpenStreetMap provides vector based spatial data that could easily integrate in web sites, mobile apps, hardware devices, programs, etc. Besides vector data, OpenStreetMap provides set of base maps for GIS software such as QGIS, and web pages.

1.2 Thesis contribution and structure

The main contribution of this thesis is a novel approach to human dynamics research that uses advanced Big Data technologies to extract meaningful results from user generated data. User generated data such as telecom CDR data and social networks data can reveal a lot about human dynamic, but to this point it was unavailable for wider academic public. Another important aspect of user generated data is its size and complexity, which makes it computationally and algorithmic very challenging to process. Finally, spatial semantic is inseparable of human dynamic and we proved its relation and quantified its mutual effect by using advanced Machine Learning techniques. Therefore, diverse expertise is needed to approach this challenging topic and we propose a multidisciplinary solution.

[2] Copernicus Land Monitoring Service

[3] OpenStreetMap

To the best of our knowledge, this is the first time that telecom CDR data is used together with spatial data to inspect the dynamic of human behaviour and habits in urban space, mostly by utilising Graph Theory and Machine Learning.

Some of the results of the research work conducted in this thesis were published in world-class journals, such as ISPRS "International Journal of Geo-Information" and presented in international conferences such as NetMob, International Conference on Big Data Analytics and Knowledge Discovery, World Conference on Information Systems and Technologies, etc.

Chapter 2 gives a high level overview of state of the art work in the Human Dynamics research, together with fundamentals of Network Science and Graph Theory. In the Chapter 3 is presented the overview of the advanced Big Data technologies that are used in this research. In the Chapter 4 is presented the case study of Evolving Connectivity Networks where the strong correlation between day time/type and human dynamic is explained. In the Chapter 5 we dive into the topic of community detection in connectivity networks and present the methodology for community detection based on HPC, that would enable us to overcome the performance issue. Chapter 6 gives the most detailed results of our research work related to human dynamics and urban space evaluated through telecom connectivity network. In the Chapter 7 are presented the results of a novel case study that explores human dynamics through mobility networks evaluated from social media data. Chapter 8 summarizes our findings and provides some conclusions of the topics studies in this thesis.

Chapter 2

Network science for studying human dynamics

In this Chapter, we present the state-of-the-art related work in the Human Dynamics research and fundamentals of Network Science. Network Science [12] is a relatively young academic field which studies complex networks such as telecommunication networks, computer networks, web networks, biological networks, social networks etc. The field assembles the multidisciplinary knowledge and relies on theoretical concepts and methods including graph theory from mathematics, statistical mechanics from physics, data mining and visualisation from computer science, inferential modeling from statistics, and social structure from sociology. Any complex relational data could be presented in the form of a network, which opens up a new potential to inspect the data and relations in completely different way using methods from Network Science. We live in the era of Big Data and complex systems that grow hopelessly complicated.

If we think about our society and how it requires the cooperation between millions of individuals, or communications infrastructure that integrates billions of devices, or chemical reactions in our bodies, we can collectively call these systems *"complex systems"* based on the fact that it is difficult to derive their collective behaviour from a knowledge of the system's components [12]. Given the important role that complex systems play in our daily lives, in science and economy, their deepest understanding, description and prediction is one of the major intellectual and scientific challenges of the 21st century [12].

The overall emerging availability of data related to complex systems, such as large scale spatial data, human dynamics data related to activity, connectivity and mobility of individuals, large scale trading and trajectory data, communication networks data, IoT[1] devices data, etc. gives us the possibility to explore these systems in the way that couldn't be imagined before. Now more than ever, is important to study the complex systems and to derive the novel knowledge from it that would help develop a better, sustainable and technologically advanced society. A long time ago, Francis Bacon declared that **"Knowledge is power"**, simple quote worth thousands words that is present now more than ever because we live surrounded by complex systems that require novel knowledge to be understood.

2.1 Human dynamics research

Human dynamics is mostly reflected through mobility, connectivity and grouping. Previously, it was almost impossible to collect the data about those aspects of human dynamics, but with the high presence of new technologies in our every day life, abundance of spatially and timely referenced data became available and ready to explore. As the technology evolved and changed our every day life and ways people are communicating, interacting, moving and spending time, so the Human Dynamics research changed and evolved and became more multidisciplinary oriented area, because many diverse expertise need to be utilized to conduct such research.

With an increasing number of devices and services that collect data about people activity, connectivity and mobility in the last few years, human dynamics research became the key topic in computational social science [77]. Human dynamics research evolves with changing every day circumstances in which people live such as natural and urban environment, emerging new technologies, climate change and society. High presence of modern information and communication (ICT) technologies including location-aware devices, various sensors and mobile technology in every day life have great impact on shaping human activity and interaction patterns [107]. Big Data, collected through many devices and services, present unprecedented opportunities for human geography to transform our understanding of the social world and human behaviour [101].

Singleton et al. discussed about relationship and interplay between modern Geography and Data Science highlighting the fact that there has never been more abundant geographic data than today and that human dynamic can not be understood without inspections of those data sources [111].

[1] The Internet Of Things (IoT) describes the network of devices that are embedded with sensors, software, and other technologies for the purpose of connecting and exchanging data with other devices and systems over the internet.

Advances in smart technologies, led by artificial intelligence, machine learning, internet services, are transforming many facets of society, such as human dynamics and human mobility [108]. Shaw et al. discussed about obstacles of conventional GIS conceptual framework to capture the new reality where the physical and virtual, objective and subjective, territorial and topological worlds are increasingly coupled and entangled for most human activities from local to global scales [108].

One of the richest data sources about human daily based activities is mobile phone data [18]. Many diverse applications with significant social impact are developed based on mobile phone data, such as urban sensing and planning [15, 28], traffic engineering [7, 60, 27], predicting energy consumption [22], disaster management [75, 90, 127], epidemiology [23, 72, 126], deriving socio-economical indicators [89, 114]. Among other applications, CDR data could be used to provide national-scale poverty estimates by deriving features from CDR data that are able to account for the variance in socioeconomic status [113].

More specifically, Ratti et al. in their review [97] highlighted the potential of using mobile phone data for urban planning. Soto et al. [112] used *Call Detail Records* to extract the information to automatically identify land use behaviors in urban environments. They used fuzzy c-means to cluster the Radio Base Station signatures and detect the class representatives of the land use in urban environment. Grauwin et al. used mobile phone data to detect land use classes in three different cities, New York, London and Hong Kong [56]. Furno et al. conducted comparative analysis between ten different cities, in which they constructed specific mobile traffic signatures to determine dynamic patterns of human presence in urban areas [51]. Rios and Muñoz used a big mobile phone data set with 880 million records in a case study for Santiago, Chile for land use pattern detection. They used the latent variable clustering technique in detecting clusters of residential, office area, leisure-commerce and rush hour pattern areas [99]. Pei et al. used hourly relative pattern and the total call volume trough semi supervised fuzzy c-means clustering approach in inferring land use types in Singapore, showing that the accuracy decreased with the increase in heterogeneity of land use and density of cell phone towers [91]. Furno et al. combined simultaneously Call Detail Records and vehicle GPS traces for revealing land use context in French and Italian cities [50].

Unveiling complex ties between land use and human dynamics properties derived from mobile phone data is an active area of research. The latest results demonstrate relations between dominant land use for each Voronoi zone and corresponding human activity represented as aggregated CDRs [10], as well as land-use composition of city's neighborhoods and the time series of CDR intensities [16]. Both studies utilized clustering to group similar land uses on one side and human dynamics properties on the other side and finally estimated agreement between clustering results obtained from this two data sources. Another novel study quantified by regression models how urban land use influences the commuting flows [73].

Noyman et al. conducted the study that suggests a methodology of "reversed urbanism" to urban planning and decision making. The methodology considers human behaviour patterns extracted from mobile phone data as a key element of urban design and their association with the functionality of urban areas [85]. One recent study conducted by Cottineau et al. showed the relation of mobile phone data indicators such as number of calls, active days, duration of calls, entropy, etc. and socioeconomic organization of cities [37]. They showed how mobile phone data together with census and administrative data could be used for urban development. Kandt et al. provided an extensive overview of the current challenges in urban policy making and highlighted the need to utilize Big Data analytics and technologies to better understand the modern human dynamics that consists of many levels beyond just simple geographical space [65].

2.2 Networks and their representation

Traditionally the study of complex networks has been strongly related to the graph theory. While graph theory initially focused on regular graphs, since the 1950s many large scale complex networks with disorderly structure and topology have been described as random graphs [6]. When we talk about *networks* we often imply to *graphs* as their natural mathematical structure. Although a network and a graph is essentially the same thing, there is subtle difference between the two terminologies. The *network* consists of the *nodes* connected by *edges*, and they often refer to real systems such as WWW, or metabolic network or telecommunications, etc. In contrast, when we use the terms *graph, vertex, edge*, we usually discuss the mathematical representation of these networks [12]. However, in practise and among professionals this distinction rarely made, so the terminologies are considered as synonyms [12].

2.2.1 Graph theory

Graph theory is field in mathematics that study graphs, which are mathematical structures used to model pairwise relations between objects. Its roots go back to 1735 in Königsberg, the capital of Eastern Prussia, a thriving merchant city of its time. The trade supported by its busy fleet of ships allowed city officials to build seven bridges across the river Pregel that surrounded the town. Five of these connected to the mainland the elegant island Kneiphof, caught between the two branches of the Pregel. The remaining two crossed the two branches of the river. This peculiar arrangement gave birth to a contemporary puzzle: Can one walk across all seven bridges and never cross the same one twice? Despite many attempts, no one could find such path. The problem remained unsolved until 1735, when Leonard Euler, a Swiss born mathematician, offered a rigorous mathematical proof that such path does not exist [2].

Euler represented each of the four land areas separated by the river with single point. Next he connected with lines each piece of land that had a bridge between them. He thus built a graph, whose nodes were pieces of land and links were the bridges. Then Euler made a simple observation: if there is a path crossing all bridges, but never the same bridge twice, then nodes with odd number of links must be either the starting or the end point of this path. Indeed, if you arrive to a node with an odd number of links, you may find yourself having no unused link for you to leave it. In the Figure 2.1 is presented the the map of the city with bridges and the simplified interpretation of the problem the way Euler saw it.

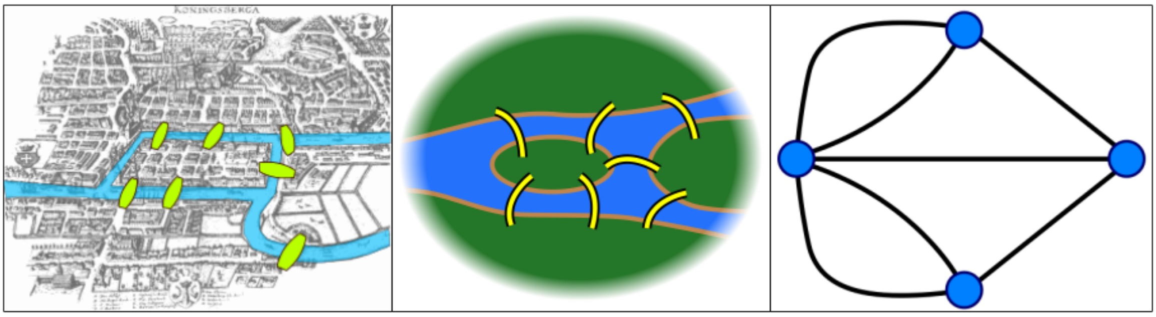


Fig. 2.1 Seven Bridges of Königsberg mathematical problem

A walking path that goes through all bridges can have only one starting and one end point. Thus such a path cannot exist on a graph that has more than two nodes with an odd number of links. The Königsberg graph had four nodes with an odd number of links, so no path could satisfy the problem. Eulers proof was the first time someone solved a mathematical problem using a graph. We can observe that first problem that is defined and solved using graphs have cartographical nature.

Graphs are simplified way to represent relationships between any pair of objects and spatial relationships follow the same rule. Today many problems are modelled using graphs. The most popular example is the Internet which consists of routers bind together with connections. Also, social networks have graph structure, where each individual is a node and links are the “relationships”. One less obvious but the most surprising example of a

[2] The Seven Bridges of Königsberg problem

graph model is protein – protein interaction in cells. In biology, protein interactions create a network which consists of proteins and binding interactions between them. There are many more examples of natural, social and technological phenomena that are modelled using graphs. Evolution of graph theory and its applications created one completely new field of science, the network science. Any network, natural or artificial, consist of nodes and links between them. Graph theory gave us mathematical tools to model, study and explain networks and that way to better understand the world around us, and inside of us as well.

Before presenting some applications and properties of graphs, we need to introduce some basic terminology. A *graph* G is the tuple (V, E) which consists of a finite set V of *vertices* and a finite set E of *edges*, where each edge is a connection between pair of vertices [95]. The two vertices associated with an edge *e* are called the *end-vertices* of *e*. An edge between two vertices *u* and *v* are often denoted by (*u, v*). The set of vertices of a graph G is denoted by V(G) and the set of edges of G by E(G). Let *e = (u, v)* be an edge of a graph G. Then the two vertices *u* and *v* are said to be *adjacent* in G and the edge *e* is said to be *incident* to the vertices *u* and *v*. The vertex *u* is also called a *neighbor* of *v* in G and vice versa. The graph in Figure 2.2 has seven vertices *a, b, c, d, e, f, g* and ten edges. Vertices *a* and *b* are end vertices of edge *(a, b)*. So, *a* and *b* are adjacent. Vertices *b, c* and *f* are the neighbors of the vertex *a*.

Graphs are data structures that have applications in many science and engineering disciplines, as they represent the structure of networks. They have a central role in the analysis of mobile phone data collected by service providers. Due to their mathematical formalism and the variety of existing graph-based algorithmic techniques, they can be used efficiently and effectively in social networks, transportation, airlines, supply chain, web and bioinformatics to solve specific problems.

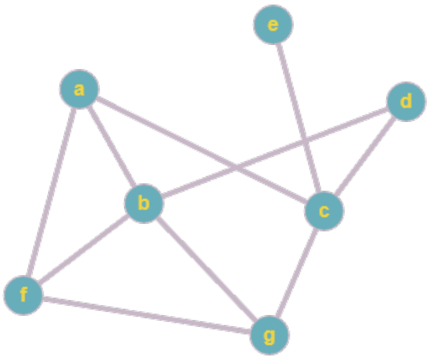


Fig. 2.2 Example of a graph.

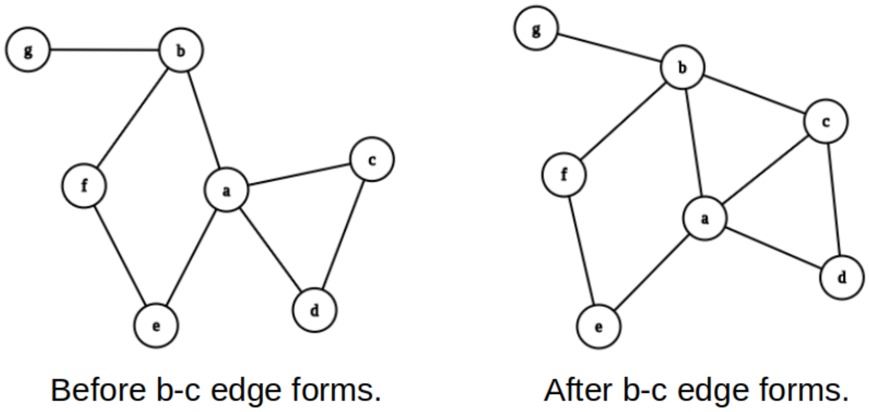
2.3 Graph algorithms

Graph algorithms are developed to better understand the properties and even hidden patterns that exist in graphs. Detecting hidden patterns and relationships in the graphs is of crucial importance to understand the properties of the network we study.

There is a huge number of algorithms that are used for graph analytics. Finding shortest paths, strongly connected components and clustering coefficient are some of the most important tasks when analysing graphs.*The path* in a graph is a way to get from vertex *a* to *b* by using edges between vertices that are connecting *a* and *b*. *The shortest path* or *geodesic path* is the distance between two nodes that traverses minimum number of edges. *Eulerean path* is a path that traverses each link exactly once. *Hamiltonian path* is a path that visits each node exactly once. The *diameter* of a network, denoted by d_{max}, is the maximum shortest path in the network. In other words, it is the largest distance recorded between any pair of nodes [12]. Breadth-First Search (BFS) Algorithm is a frequently used algorithm in network science, which is used to identify shortest path between any two vertices in a graph. Connectivity is very important property of the network. In a sense of a graph, connectivity describes whether a vertex is reachable from some other vertices in a graph or not. Graph can be fully connected, meaning that from any vertex in the graph any other vertex is reachable, or graph can be disconnected with some vertices or even subsets of vertices completely isolated. When graph has large number of vertices, we need to apply algorithms to detect connected components in the graph. Connected component is a subset of vertices such that every vertex in the subset has a path to every other and the subset is not part of some larger set with the property that every vertex can reach every other. Dividing a graph into its components is global way of describing its structure. Breadth-First Search algorithms can be used also for finding connected components in the graph.

With knowing the graph structure, the number of connected components and its position in the graph we can gain better insight in how the information is spread through the graph, and that could be of great practical value to explore the nature of the network that is represented with the graph. If the graph represents telecom network, with knowing the connected components we would know how would information spread between individuals. If the graph represents some biology phenomena, with information about connected components we would know which part is isolated and which is well connected and with that knowledge we could predict the spreading of viruses or the mutation in the cells.

When analysing social networks and the way information is spread between people, one basic principle occurred, the principle of *triadic closure*. The principle of *triadic closure* could be described in the way that if two people in social network have a friend in common, then there is an increased likelihood that they will become friends at some point in future. The principal of *triadic closure* seems as a natural mechanism to make new connections, especially in social networks [17]. In the visual representation of the graph this situation looks like there are two edges with one vertex in common, and if an edge is created between two vertices at the end of those edges that new edge is “closing the triangle”, Figure 2.3. Then all three nodes are connected to each other and they are forming a triangle in the network.



Before b-c edge forms.

After b-c edge forms.

Fig. 2.3 Example of triadic closure in a graph

The basic role of triadic closure in social networks has motivated the formulation of simple social network measures to capture its prevalence. One of these is the **clustering coefficient**. The clustering coefficient of a vertex *a* is defined as the probability that two randomly selected friends of *a* are friends with each other [12]. In other words, it is the fraction of pairs of *a's* friends that are connected to each other by edges. In general, the clustering coefficient of a vertex ranges from 0 (when none of the vertex's friends are friends with each other) to 1 (when all of the vertex's friends are friends with each other). The more strongly triadic closure is operating in the neighbourhood of the vertex, the higher the clustering coefficient will tend to be.

Although the idea behind triadic closure and the clustering coefficient came from analysing social networks, it could be applied as a general rule for different type of networks. In general, clustering coefficient is a measure that represent the local density of the network. The more densely interconnected the neighbourhood of vertex is, the higher is its local clustering coefficient [12]. The clustering coefficient of a vertex is defined as 2.1

$$C_i = \frac{2 * L_i}{k_i(k_i - 1)} \tag{2.1}$$

where L_i is the number of links between the k_i neighbours of vertex i.

The clustering coefficient in general refers to clustering of local vertex, that means it is local measure of graph density. The degree of clustering of a whole network is captured by the *average clustering coefficient*, representing the average of all clustering coefficients of all vertices in the network. Average clustering coefficient is defined as 2.2.

$$\langle C \rangle = \frac{1}{n} \sum_{i=1}^{n} C_i \tag{2.2}$$

Besides average clustering coefficient, there is a *global clustering coefficient* which measures the total number of closed triangles in the network. Global clustering is defined as a fraction between number of triangles and number of connected triples 2.3.

$$C_{\Delta} = \frac{3 * NumberOfTriangles}{NumberOfConnectedTriples} \tag{2.3}$$

The average clustering and the global clustering are not equivalent, but both measurements represent the degree of a network's global clustering. In the following chapters we will describe some graph properties and procedures that are used to unveil the structure of the network.

2.3.1 Graph properties and centrality measures

To inspect graph and to extract knowledge from the relationships within, we can analyse global and local graph properties [36]. Global graph properties provide information on a global structure of the graph and further allow comparisons among graphs. Some of the commonly applied global graph properties are the number of vertices and edges, maximum weight, radius, diameter, max clique size and average clustering coefficient. The number of vertices represent the size of the graph, while the number of edges is related to the density of the graph. Graph with small number of vertices comparing to very high number of edges is considered a dense graph. Maximum weight is the property applied only to weighted graphs and it refers to the edge with maximum weight in the graph. To understand the meaning of the radius and diameter of the graph we need to know the eccentricity of a vertex. The eccentricity of the vertex a is the greatest distance between a and any other vertex. We can think of the eccentricity as a measure of how far a vertex is from the most distant vertex in the graph. The radius of a graph is the minimum eccentricity of any vertex, while the diameter is the maximum eccentricity of any vertex in the graph. A clique in the graph is a subset of vertices such that every two distinct vertices in the clique are adjacent. A clique could be applied only to undirected graphs. Max clique size is the number of vertices that participate in the largest clique in the graph. Average clustering coefficient is the average of all clustering coefficients in the graph.

Unlike global graph properties that provide unique measure per graph, local graph properties are used to uncover localized patterns within the graph. Local graph properties are calculated per vertex and they reveal structure of the graph in the close neighbourhood of the vertex. Some of the most commonly calculated local graph properties are clustering coefficient, node degree and centrality. Clustering coefficient of a vertex is a measure of local density of the network. Node (or vertex) degree is the number of edges connected to the node. Node centrality measures [86] quantify the importance of graph nodes. Among the most widely used node centrality measures are:

- *Degree Centrality* (DC): the number of edges incident to the node,
- *Betweenness Centrality* (BC): the number of shortest paths passing through the node,
- *PageRank* (PR): related to the probability that a random walker will visit the node, and
- *Core Number* (CN): a value associated with a node that quantifies how well the node is connected with respect to its neighborhood (we will provide more details on this).

BC and PR are the most computationally intensive, whereas DC and CN require linear time to compute. More specifically, the complexity of CN in undirected and unweighted graphs is $\mathcal{O}(n+m)$, where n is the number of nodes and m is the number of edges .

Core number. The core number of node u, $CN(u)$, is the maximum value of k such that u belongs to the k-core of the graph G.

The k-core. The k-core of a graph G is the maximal subgraph H such that every node u in H has at least k neighbors in H, i.e., the degree of all nodes in H is at least k.

The core decomposition concept has many applications in diverse scientific areas. A broad coverage of topic related to the application of the core decomposition is extensively described in the work [76].

2.3.2 Graph filtering

In the recent years, a huge amount of data related to social networks, biology, Internet, communication networks, transport and mobility has become available for scientific research. In some domains, those large-scale networks have even become a critical element for new services development. Common property of these networks is very large size and large quantity of complex relationships. When modeling such complex network we often obtain very large, densely connected and weighted graph. Extracting relevant information from large, complex, weighted networks can be very challenging task. When working with such networks, reduction in size is necessary step towards knowledge extraction.

Choosing the right reduction method for complex networks is not a trivial task because it leads to the trade-off between the level of network reduction and the amount of relevant information we need to preserve [105]. The major obstacle is that in many cases the probability distribution P(ω) that any given link is carrying a weight ω is widely distributed. Due to such wide probability distribution P(ω), there is a lack of characteristic scale and any method based on the thresholding would destroy the network structure. Especially multiscale real-world networks are sensitive to this feature because the weights are locally correlated on edges incident to the same node and nontrivially coupled to topology [14]. Thus, the reduction techniques that highlight the relevant structures, hierarchies and local perspective are needed when working with very large, complex networks.

In statistical mathematics, reduction or filtering techniques aimed at revealing the relevant information from the massive and noisy data sets are very popular and successful. Common example is the Principal Component Analysis which is used to identify hidden patterns by reducing the dimensions of multivariate data [62]. In this research we are focused on reduction techniques that preserve the statistically relevant backbone of complex weighted networks. We refer to network reduction as the construction of network that contains far fewer data (in our case, number of links) but preserves relevant features of the original network. There are two main types of reduction schemes: coarse-graining and filter/pruning. In the case of coarse-graining techniques, nodes that are sharing a common attribute could be gathered together in the same class, group or community and then substituted by a single new unit that represents the whole group in the new network. Coarse-graining techniques could be thought of as zooming in/out of the system. On the contrary, something completely different is happening when a filter is applied to the network. In this case, there is no change to the observation scale of the network, instead the reduction is done by discarding the elements (nodes and edges) that are not carrying relevant information about the network. A well-known technique that is used to detect the core of the network is the k-core decomposition [76], where the filtering rule operates upon the connectivity of the nodes.

In the case of weighted networks, the problem of reduction gets even more complex due to the high heterogeneity of the network and local correlation of the weights. The basic reduction techniques such as extraction of the minimum spanning tree and application of global threshold on the weights, could potentially destroy the network structure. Due to heterogeneity in weights distribution across network there is a need to introduce some measure of the fluctuation of the weights attached to a given node, at the local level in relative terms, so that each node could independently assess the importance of its connections. With the aim to reduce the network size, but to preserve the structure of the network and importance of the edges at local level we used the method for statistical filtering, the Disparity filter [105].

Disparity filter

The Disparity filter is a method for extraction of the relevant connections in weighted networks, by discarding the edges that are not carrying the statistically significant information in the network. The Equation (2.4) is used to calculate the *link significance*, α_{ij}, where α

denotes the significance *threshold*, p_{ij} is the probability of having a connection between nodes i and j, and n is the number of nodes in the network.

$$\alpha_{ij} = 1 - (n-1)\int_0^{p_{ij}} (1-x)^{n-2} dx < \alpha \tag{2.4}$$

The disparity filter proceeds to evaluate the significance of the connections in the network at local level, and identify which edges should be preserved. We note that smaller values of α_{ij} denote more significant edges. Therefore, filtering is applied by keeping all edges where $\alpha_{ij} \leq \alpha$ and thus removing all edges where $\alpha_{ij} > \alpha$. By changing the significance level denoted in α parameter, we can filter out the edges progressively focusing on more relevant edges.

By tuning the significance level α we are obtaining more or less restrictive subsets of the network, revealing the core backbone gradually. This strategy is justified whenever we have the network with very high heterogeneity and locally correlated weights. Otherwise, the results of the significance pruning would be similar to the global threshold algorithm.

2.4 Clustering

In the previous chapter we mentioned coarse-graining techniques for network size reduction that consist of joining together nodes that are sharing the same attributes, and then replacing the group of nodes with the single node representing the group. Described process is called graph clustering, or more often community detection if we are working with social networks. Clusters, communities, or modules within the graph are groups of vertices which share the common properties. Another characteristic of clusters is that the vertices within the cluster have higher probability of being connected to each other than to members of other clusters, though other patterns are possible. Network, graph clustering or community detection is complex problem because there are no universal protocols on the fundamental ingredients, like the definition of cluster/community itself, nor on other crucial issues, like the validation of algorithms and the comparison of their performances [48].

Graph clustering or community detection although complex and ill-defined problem, has drawn great attention in the recent years within the scientific community and became one of the most popular topics of modern network science. Identifying communities within the graph has great benefits because it offers the valuable insight on network structure and can answer some more common questions such as: how the information is spreading through the network? With knowing the community structure, we can focus on regions that have some degree of autonomy within the graph. The lack of universal protocols and definitions related to community detection leave a lot of freedom to propose diverse approaches to the problem, which are often tightly bonded to the particular research problem. Such ambiguity leads to development of diverse algorithms and analytical tools that could be used for community detection, i.e. graph clustering.

2.4.1 Methods for clustering

Graph clustering is a complex task, with many open questions related to clusters quality and validation. There are many algorithms for graph clustering, adjusted for specific problems and they can be grouped in categories based on different criteria.

In general, the only preliminary information that is used to evaluate the clusters in the network is the network structure, i.e. which pairs of nodes are connected by an edge and which are not (possibly including the weights over edges). Any additional insight about network structure, such as the number of expected clusters would be beneficial to reduce the huge space of possible solutions. Among many possible pre-detected inputs, the number q of expected clusters plays a great role. Many popular algorithms require the specification of expected clusters number q before they can run.

One of the widely used algorithms is K-Means clustering. In the K-Means clustering, the objective function is the sum of squares of the Euclidean distances of data points to their closest representative. Therefore, when initiating the algorithm, one needs to provide some randomly selected points as representatives, so that the algorithm can proceed. This approach is widely used due to it's simplicity, and relatively small complexity, but it is highly sensitive to the selection of representative points at the beginning. Usually, the user inputs only the number of expected clusters q, and the algorithm chooses the representative points randomly. It is clear to notice that the choice of the points can vary from each iteration.

Unlike K-Means clustering, there are methods that can infer the number q of clusters during the course of the algorithm execution. Graph clustering methods based on optimisation have received the greatest attention in the literature. These techniques are relaying on the stop function that indicate the quality of the partitions. The goal is to find the extremum, usually the maximum of a stop function over the space of all possible clusterings [48]. The algorithm stops when reaching the local extremum of a stop function, and returns the partitions set from the previous level.

There are many possible quality functions that could be used as a stop function in the algorithm, but one of the most popular is the *modularity* described in [79]. Modularity is used to estimate the quality of a partition of the newly generated network of clusters. It can be found in the literature in many different forms depending on the specific use case, but the most general form is (2.5),

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{i,j} - P_{i,j}\right) \delta(c_i, c_j) \tag{2.5}$$

where m is the number of edges of the network, the sum runs over all pairs of vertices i and j, $A_{i,j}$ is the element of the adjacency matrix, $P_{i,j}$ is the null model term and in the Kronecker delta at the end C_i and C_j indicate the clusters of i and j. The term $P_{i,j}$ indicates the average adjacency matrix of an ensemble of networks, derived by randomising the original graph. Therefore, modularity measures how different the original graph is from such randomisation. The idea came from the observation that by randomising the network structure clusters are destroyed, so the comparison between the actual structure and its randomisation reveals how non-random the cluster structure is. A standard choice is $P_{i,j} = \frac{k_i k_j}{2m}$, with k_i and k_j being the degrees of i and j, and corresponds to the expected number of edges joining vertices i and j if the edges of the network were rewired to preserve the degree of all vertices, on average. Including this approximation we can redefine the general form of the modularity to classic, more common one (2.6).

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{i,j} - \frac{k_i k_j}{2m}\right) \delta(c_i, c_j) \tag{2.6}$$

Although the form of modularity defined in (2.6) is widely used, other choices of the null model term are possible allowing us to incorporate specific features of the network structure. It is important to highlight that because of the delta term in the equation (2.6), the only contributions to the sum come from vertex pairs belonging to the same cluster. Modularity maximisation is NP-hard problem, so one can expect only to find decent approximations of the modularity maximum, and many diverse approaches have been proposed in the literature. Simple but effective approach is proposed in [79], due to which the modularity has become the best known and most studied object in network clustering. Despite the rapid success of the method, it quickly became clear that the measure has its downside. One example is the existence of high-modularity partitions even in random graphs where there are no natural grouping.

The problem of graph clustering gets even more complex for large scale graphs, where demands for finite computational complexity can be great limitation for analysis. To enable evaluation of community structure in large scale networks, Blondel et al. proposed the method based on maximizing the modularity function [20]. The modularity of a partition is a scalar value between -1 and 1, and it measures the density of links inside communities, i.e. clusters as compared to links between communities, i.e. clusters. In the case of weighted networks it is defined as (2.6), where $A_{i,j}$ represents the weight of the edge between i and j, $k_i = \sum_j A_{i,j}$ is the sum of the weights of the edges attached to vertex i, c_i is the cluster to which vertex i is assigned, the Kronecker delta is 1 if $u = v$ and 0 otherwise and $m = \frac{1}{2}\sum_{i,j} A_{i,j}$. The final form of the modularity function used in algorithm proposed by Blondel et al. [20] is (2.7).

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{i,j} - \frac{\sum_j A_{i,j} \cdot \sum_i A_{j,i}}{2m}\right]\delta(c_i, c_j) \tag{2.7}$$

Unfortunately, evaluating the cluster structure based on the maximization of the modularity, is an NP-hard problem. To provide an efficient solution, the algorithm proposed in [20] uses an iterative process that involves shrinking the graph, every time modularity converges. The algorithm starts by assigning each node to its own cluster and starts moving the nodes to neighbouring clusters forming a dendogram, while calculating the modulurity in phases. In each phase, each node is assigned to a neighboring cluster that maximizes the modularity of the graph. As long as nodes are moving around clusters and modularity grows, we keep on executing this process. When there are no more changes, a shrinking process is applied. Upon shrinking the graph, each cluster produced during the previous phase, it is assigned to the same *super node* of the new graph. The same process is applied to the new graph. The algorithm terminates when the modularity detected in the new graph is less than the modularity detected in the previous one. The set of cluster groupings that maximize the modularity is returned as an answer. The outline of the technique is depicted in Algorithm 1. It is evident that this algorithm may reach a local maximum. However, in general it performs very well, it is efficient and the quality of the evaluated cluster structure is high.

Alongside above described methods for clustering that are distinguishing whether the number of clusters q is known in advance or not, there are other methods such as *consensus clustering, spectral methods, hierarchical clustering*, etc. [48]

2.4.2 Clustering validation

Due to the existence of many diverse approaches and algorithms for graph clustering, and the lack of formal definitions and criteria for evaluating the goodness of partitions there is a need to validate the cluster structure. Clustering validation techniques are used to compare between different algorithms, approximate the number of clusters, etc. Based on the validation criteria there are three categories of measures for clustering validation: external, internal and relative. External validations are used to compare the evaluated clusters to the ground-truth labels in the case when they exist. Internal validation use only the underlying data set to measure how well obtained clusters satisfy compactness, connectedness and/or separation criterion [98]. Relative validations are used to compare the results of the same algorithm but realized considering different values of parameters. Many variants and combinations of those criteria

Algorithm 1 LOUVAIN $(G(V,E))$

```
Input: the graph $G$
Result: the communities of $G$
$n \leftarrow |V|$                                        ▷ Number of graph nodes
$done \leftarrow false$
while not done do
    assign each $u \in V$ to a different community
    while there is a change do
        for every node $u \in V$ do
            $C \leftarrow$ a community that maximizes modularity    ▷ $C$ is a neighboring community
              or $u$'s community
    if newModularity $>$ oldModularity then
        $G \leftarrow$ shrink graph based on communities  ▷ Each community becomes a super node
          in the new graph
    else
        return communities
```

exists, here we will describe the four most widely used measures for cluster validation, i.e. Purity, Entropy, Rand Index and Adjusted Rand Index.

The *Purity* of a cluster measures the extent to which each cluster contains elements from primarily one class [139].

Definition - Purity. Given a set S of size n and a set of cluster C of size k, then, for a cluster $c_i \in C$ of size k_i, the purity is $p(c_i) = \frac{\max_i(n^i_j)}{k_i}$, where n^i_j is the number of elements of the j-th class assigned to the i-th cluster. The overall purity is defined as $P(C) = \sum_{i=1}^{k} \frac{k_i}{n} \cdot p(c_i)$

Entropy is an evaluation method that assumes that all elements of a set have the same probability of being picked and, by choosing an element at random, the probability of this element to be in a cluster can be computed [123].

Definition - Entropy. Given a set S of size n and a set of clusters C of size k, then, by assuming that all elements in S have the same probability of being picked, the probability of an element $s \in S$ chosen at random to belong to cluster $c_i \in C$ of size k_i is $p(c_i) = \frac{k_i}{n}$. Then, the overall entropy associated with C is $H(C) = -\sum_{i=1}^{k} p(c_i) \cdot \log_2(p(c_i))$

The *Rand Index (RI)* is a measure used to determine the similarity between two data clusterings [96].

Definition - Rand Index. Given a set $S = \{s_1, s_2, ..., s_n\}$ of n elements and two groupings $X = \{X_1, X_2, ..., X_r\}$ and $Y = \{Y_1, Y_2, ..., Y_t\}$ then the Rand index is $RI = \frac{a+b}{\binom{n}{2}}$, where $a = |S'|$

with $S' = \{(s_i,s_j)|s_i,s_j \in X_k, s_i,s_j \in Y_l,\}$ and $b = |S''|$ with $S'' = \{(s_i,s_j)|s_i \in X_{k_1}, s_j \in X_{k_2}, s_i \in Y_{l_1}, s_j \in Y_{l_2},\}$ for some $1 \leq i,j \leq n, i \neq j, 1 \leq k,k_1,k_2 \leq r, k_1 \neq k_2, 1 \leq l,l_1,l_2 \leq t, l_1 \neq l_2$,

Adjusted Rand Index (ARI) is a cluster evaluation method that calculates the fraction of correctly classified (respectively misclassified) elements to all elements by assuming a generalized hypergeometric distribution as null hypothesis [123]. ARI is the normalized difference of the Rand Index and its expected value under the null hypothesis [120]. ARI uses the contingency table.

Definition - Adjusted Rand Index. Given a set S of n elements and two groupings $X = \{X_1,X_2,...,X_r\}$ and $Y = \{Y_1,Y_2,...,Y_s\}$, the overlap between X and Y, can be summarized in a contingency table $[n_{ij}]$, where $n_{ij} = |X_i \cap Y_j|$, $a_i = \sum_{j=1}^{s} n_{ij}$ and $b_j = \sum_{i=1}^{r} n_{ij}$. Using the contingency table, the Adjusted Rand Index is defined in Equation (2.8).

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \frac{\sum_i \binom{a_i}{2} \cdot \sum_j \binom{b_j}{2}}{\binom{n}{2}}}{\frac{1}{2}(\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}) - \frac{\sum_i \binom{a_i}{2} \cdot \sum_j \binom{b_j}{2}}{\binom{n}{2}}} \tag{2.8}$$

Chapter 3

Distributed data analysis

Big Data is extremely valuable to produce new breakthrough in economy and science, but it arises with many challenges, such as difficulties in data capture, data storage, data analysis and data visualization [33]. User generated network data such as telecom or social media data are no exception in the concept of Big Data. Network data analysis can be computationally very demanding, due to the complex network structure, size of the network, and overall algorithms complexity when executing some algorithms. To handle computationally intensive tasks, a potential solution is to use multiple resources and apply parallel or distributed computing techniques, aiming at reducing the overall processing time.

In the past few years we have seen the major and rapid change in computer systems, due to the growing data volumes and stalling processor speeds more and more applications require to scale out to distributed systems. The system is distributed if the processing is done in many distant machines, and only the final result get back to main machine, i.e. the user. It is clear that these systems require stable, broadband internet connection in order to function, and some communication costs are implied, but the overall benefits exceed much these downsides.

3.1 Hadoop MapReduce

We live in the Big Data era, where everything around us (devices, online systems, vehicles, machines, etc.) is generating some data that could be analysed. To analyse and process this huge amount of data and to extract the meaningful knowledge from it, there is a need of deploying data intensive application and storage clusters [53]. These type of systems and applications have of course great requirements such as fault tolerance, parallel processing, data distribution, load balancing, scalability and highly availability. To deal with such complex requirements, Google introduced MapReduce programming model [54]. MapReduce is a programming model for distributed computing that consist of two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job. Great advantage of MapReduce model is that it is easy to scale the processing of the data over multiple computing nodes. Apache Hadoop is an open source implementation of MapReduce system [67]. Hadoop is designed to process, handle and combine extremely large sets of unstructured, structured and semi-structure data. It is Java-based programming framework that manages the processing of large data sets in distributed or clustered environment. The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a MapReduce programming model. Hadoop splits files into large blocks and distributes them across nodes in a cluster. An HDFS cluster follows typical Master-Slave architecture, with single Namenode, a master server that manages the file system namespace and regulates access to files by clients and many distributed Datanodes, which manage storage attached to the nodes that they run on. The base of Hadoop framework is composed of the modules:

- **Hadoop HDFS** - a distributed file system that stores the data on commodity machines, it acts as storage layer,
- **Hadoop YARN** - resource manager used for jobs scheduling,
- **Hadoop MapReduce** - an implementation of the MapReduce programming model for large-scale data processing, it acts as application layer, where other applications are possible.

Hadoop architecture is descriptively presented in Figure 3.1. Hadoop is the most popular and open source implementation of MapReduce programming model. It is a software framework for reliable, scalable, parallel and distributed computing. Another great advantage of Apache Hadoop framework is that it empowers the processing of Big Data on commodity hardware, therefore decreasing the cost of overall system. Hadoop enables parallel processing of huge amount of data, automatic data partitioning, distribution, fault tolerance and load balancing. The rapid growth rate of users and systems generated data first posed challenges on big companies like Facebook, Google, Amazon and Yahoo. These companies need to execute Big Data analytic on daily basis to deduce demands and queries of their users. With such challenging demands, traditional tools and applications quickly become unable to meet the requirements. Therefore, Hadoop was designed to manage such demanding applications.

Many enterprises, industries and Universities works on parallel and distributed computing but for many of them Big Data processing and analysis is not an easy task and it requires much effort to manage. For such user group, the open source system like Hadoop that can handle demanding processing is the great opportunity to deliver some value in the context of their tasks and expertise.

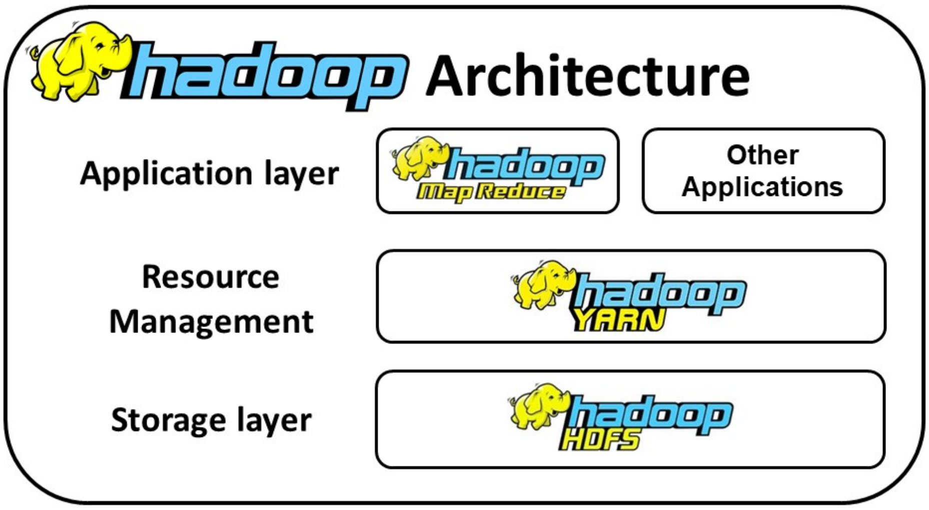


Fig. 3.1 Hadoop architecture

3.2 Apache Spark

Not long after the Hadoop was launched and quickly became the most popular tool for Big Data analysis, another framework for challenging data intensive processing is introduced - Apache Spark. Apache Spark started as a research project at UC Berkeley in the AMPLab [1], which focuses on big data analytics. The goal of the research was to create programming model that supports much wider class of applications than MapReduce, while maintaining its automatic fault tolerance. Concretely, MapReduce turned to be inefficient for multi pass applications that require low-latency data sharing across multiple parallel operations. These applications are very common in analytics, and include:

- **Iterative algorithms** - including many machine learning algorithms and graph algorithms like PageRank,
- **Interactive data mining** - where a user would like to load data into RAM across a cluster and query it repeatedly,

[1] https://amplab.cs.berkeley.edu/

- **Streaming applications** - applications that maintain aggregate state over time.

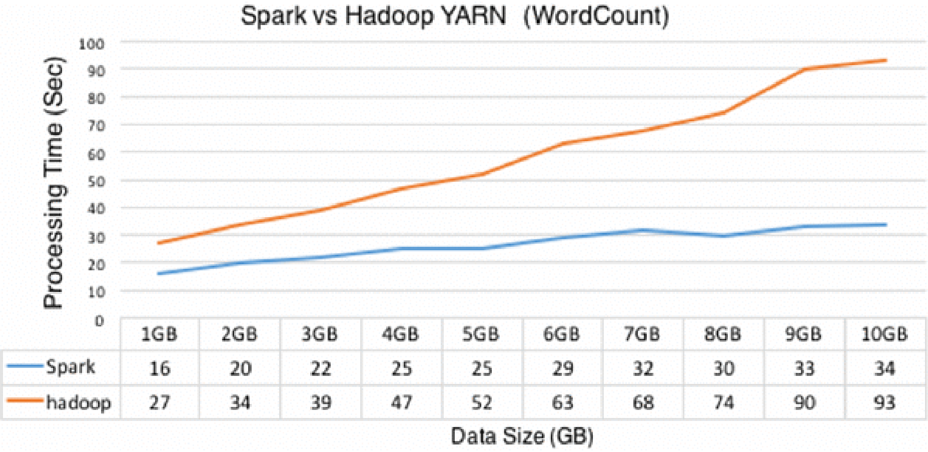
Traditional MapReduce and DAG (Directed Acyclic Graph) engines are inefficient for these applications because they are based on acyclic data flow: an application has to run as a series of distinct jobs, each of which reads the data from the stable storage, and writes it back to the storage. Repeated I/O operations that involve reading and writing to the storage bring in significant cost to already costly data intensive processing. Unlike Hadoop MapReduce, Spark exploits main memory as much as possible, being able to persist data across rounds to avoid unnecessary I/O operations. Besides that, Spark offers an abstraction called **RDD** (*Resilient Distributed Dataset*) to support data intensive applications efficiently [137]. Essentially, an RDD is a distributed collection, i.e., a dataset that is split to several partitions and each partition is located in one of the cluster machines. Note that an RDD is *immutable*, and therefore its contents cannot be altered. To change the contents of an RDD another RDD must be created. RDDs can be stored in memory between queries without requiring replication. Instead, they rebuild lost data on failure using lineage: each RDD remembers how it was built from other datasets (by transformations like map, join or groupBy) to rebuild itself. RDDs allow Spark to outperform existing models by up to 100x in multi pass analytics. RDDs may correspond to files in HDFS, files in the local file system, or datasets in other systems like HBase, Cassandra, Amazon S3, and many more.

After Hadoop made a substantial success with its MapReduce model, many scientific and industry professionals wanted to see the proof that Apache Spark can really outperform it. Yang et al. [132] did a small performance test using Word Count application with different data sizes, just to verify that Spark has better performance than MapReduce. First they randomly generated the ten test files of sizes between 1GB and 10GB, and put them into HDFS. Next they run Word Count over each file by Spark and Hadoop individually. As presented in Figure 3.2, Spark outperforms Hadoop regardless of the file size. It is worth mentioning that the difference between them increases when the test file size increases.

Apache Spark is a unified distributed engine with a rich and powerful API for Scala, Python, Java and R [66]. Spark jobs are executed based on a master-slave model in cooperation with a cluster manager such as YARN [2] or MESOS [3]. Spark jobs execution scheme is presented in Figure 3.3. A Spark application consists of a *Driver* program that is responsible for executing the `main` function. The driver detects parts of the application that will be processed by *Worker Nodes* in parallel. The Driver communicates with the *Resource Manager* (e.g., YARN or MESOS) in order to get access to cluster resources. The main execution component is the *Executor* which corresponds to a Java Virtual Machine (JVM) that runs

[2] http://hadoop.apache.org/

[3] http://mesos.apache.org/



	1GB	2GB	3GB	4GB	5GB	6GB	7GB	8GB	9GB	10GB
Spark	16	20	22	25	25	29	32	30	33	34
hadoop	27	34	39	47	52	63	68	74	90	93

Fig. 3.2 Spark vs. Hadoop MapReduce performance test [132]

on a Worker Node. Distributed processing is achieved by the communication between the Driver and the Executors. A Spark job is split into a directed acyclic graph (DAG) of stages where each stage is a collection of tasks, and then each task is sent to an Executor which is responsible for the execution.

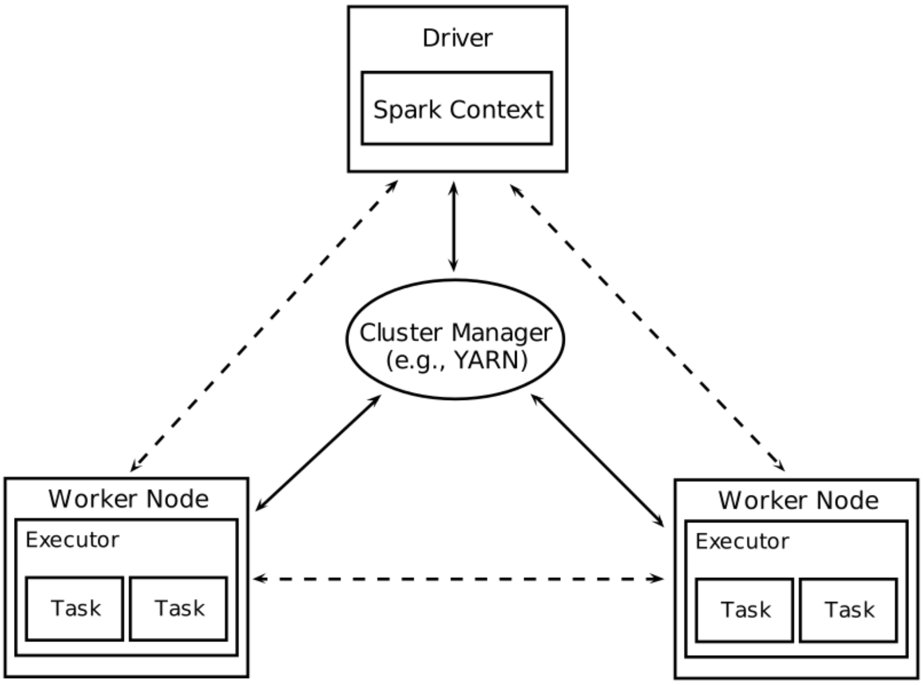


Fig. 3.3 Apache Spark execution scheme

Unlike Hadoop, which is closed system operating on three different layers, each of which are Hadoop components, Apache Spark allows many different technologies to connect to its core engine, and exploit its power of resilient, distributed, fault tolerant processing. Architecture of the Spark based system is descriptively presented in Figure 3.4.

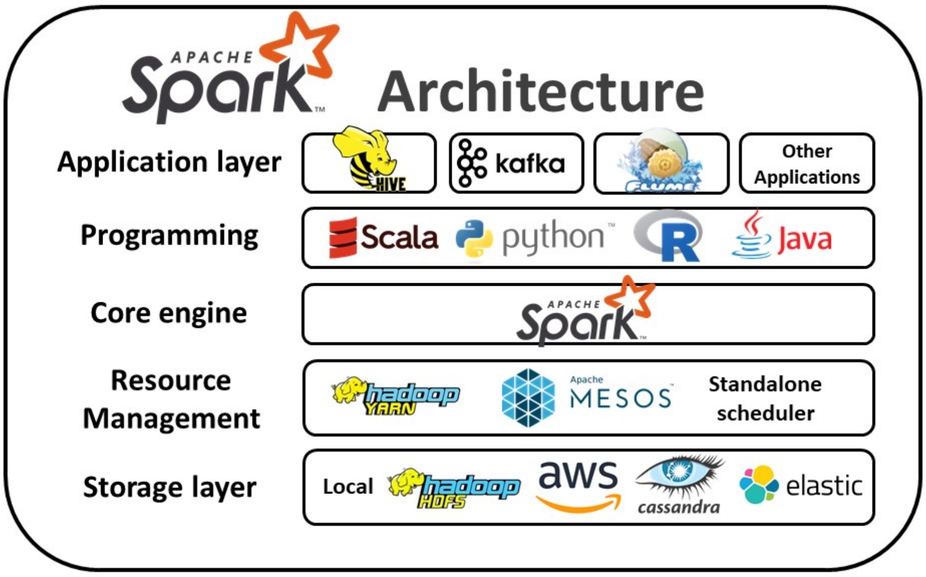


Fig. 3.4 Apache Spark system architecture

General architecture of a system built around Apache Spark core engine consist of:

- **Application layer** - applications that use Apache Spark core engine with the request to perform complex processing,
- **Programming layer** - Apache Spark supports four different programming languages, Scala, Python, Java and R,
- **Core engine** - Apache Spark core engine written in programming language Scala, this is where the main abstraction of data structure happens, execution DAG is created and the execution is triggered,
- **Resource management layer** - layer that manages cluster resources, such as the nodes, memory and CPU power,
- **Storage layer** - layer where the data is stored, it could be local or distributed.

Apache Spark supports many diverse applications that can connect to its core engine and exploit its power. Many of them are also from the Apache family, such as Apache Hive which is data warehouse engine that facilitates reading, writing, and managing large data sets residing in distributed storage using SQL, or Apache Kafka which is a distributed streaming platform used to feed the streaming data into the system for further analytics. Similar example of the application that can connect to Spark is Apache Flume which is distributed service for efficient manipulation with log data, although it is not restricted to log data. Besides applications from the Apache family of technologies some other, more traditional applications such as MySQL or PostgreSQL can also connect to Spark and feed the data into it. Spark also supports NoSQL databases like Cassandra and MongoDB. As we can see from the previous brief elaboration, many different technologies in the application layer can connect to Apache Spark core and feed the system with the data, in the end Apache Spark is all about the data.

Apache Spark provides high-level programming APIs in Scala, Python, Java and R. That means that a user can develop a software application in any of the supported programming languages and deploy it to Apache Spark cluster. Scala is native language for Spark, because the Spark core is written in Scala, and the latest updates are always first available in Scala, but usage of other languages are not falling behind since there is a large development community around it.

Apache Spark core engine is the heart and soul of the Spark based system. It is written in programming language Scala, and it runs on JVM. Spark core uses a master-slave architecture, where the Driver program runs in the master node and distributes the tasks to the Executors running on various slave/worker nodes, Figure 3.3. Spark core is the place where the main data abstraction happens, after loading the data into the Spark we proceed working with distributed data structures such as RDD or DataFrame, which are unique for Spark. In the Spark core there are available some specific libraries such as GraphX for dealing with graphs, MLlib for machine learning, Spark SQL for structured data analysis and Structured Streaming for stream data processing. The power of Apache Spark fast processing lies in the extensive usage of main memory (RAM) without heavy disk I/O operations. Apache Spark performs the computation in the main memory of the worker nodes and does not store mid-step results of computation on disk. Besides extensive RAM usage, Spark owes its speed to the process of so called "lazy evaluation". Spark supports a collection of parallel operations, transformations and actions. Transformations are operations over an RDD used to make a new resulting RDD (remember that and RDD is a immutable data structure, therefore we can not change an RDD we can just make the new one), where actions are operations that launches a computation on and RDD and then returns the result to the driver program. Lazy evaluation means that the transformations are executed only when the action is called, Figure 3.5. Spark keeps track of the lineage graph of transformations which is used to compute each RDD on demand and to recover lost data if a failure happens [103]. Lazy evaluation enables Spark to optimize the computations and entire data flow from end to end.

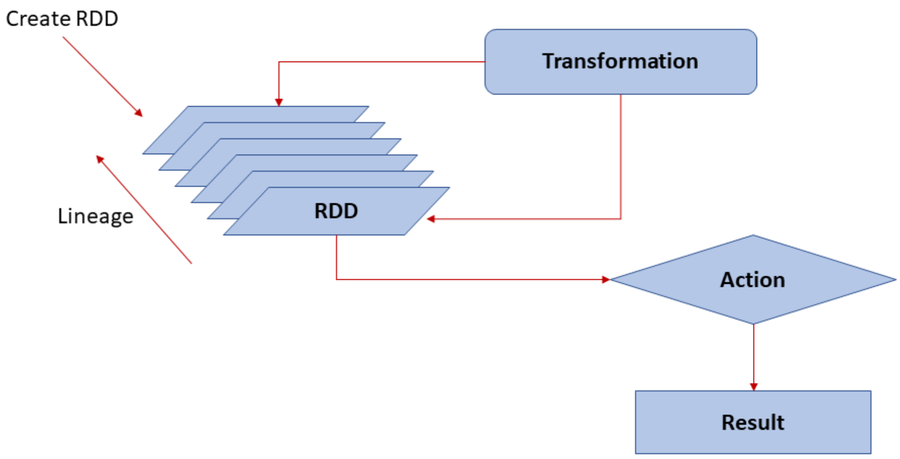


Fig. 3.5 Lazy evaluation of RDDs

The Driver program of the Spark application communicates to the nodes in a cluster using resource manager such as Hadoop YARN, Apache Mesos or Standalone Scheduler. YARN is a resource negotiator included with Apache Hadoop. YARN decouples the programming paradigm of MapReduce from its resource management capabilities, and delegates many scheduling functions (e.g., task fault-tolerance) to per-application components. Apache Mesos is a fine-grained resource negotiation engine that supports sharing and management of a large cluster of machines between different computing frameworks, including Hadoop, MPI, Spark, Kafka, etc. The main difference between YARN and Mesos is the resource negotiation model. Whereas YARN implements a push-based resource negotiation approach, where clients specify their resource requirements and deployment preferences, Mesos uses a pull-based approach, where the negotiator offers resources to clients which they can accept or decline. Spark offers also a built-in scheduler called Standalone scheduler, where deployed applications will run in FIFO (first-in-first-out) order, and each application will try to use all available nodes.

Similar like for application layer, for storage Spark supports multiple options, each of which has its own trade-offs and operational details. The most common storage systems used for Spark are distributed file systems based on Hadoop's HDFS and key-value stores such as Apache Cassandra [31]. As many Big Data applications require the data to be stored in the cloud, Spark offers a connector to AWS services, Microsoft Azure and Google Cloud. Depending on the use case, many scenarios are possible but before choosing a storage option, it is recommended to evaluate the performance of its Spark connector and to evaluate the available management tools.

We can conclude that the systems based on Apache Spark core engine are very complex and highly diverse regarding the applications. With so many supported technologies in application layer, programming, resource management and storage, Apache Spark became nearly default choice when working with Big Data, regardless of the domain or system requirements.

3.3 Big Data applications using Hadoop and Apache Spark

In the recent time, exponential growth of available data generated from users, devices and services lead to the new perspective on data analytics. The new paradigm, so called "Big Data" created the environment for progressive development of new technologies, such as Hadoop and Apache Spark. The main expectation of the new technology is to enable fast processing of massive data while taking care of fault tolerance, scalability, availability, load balancing in distributed environment. First Apache Hadoop made a breakthrough with its MapReduce model in 2006, shortly after Apache Spark was introduced in 2009 with even better performance, and many other technologies were developed to enrich the ecosystem of Big Data technologies.

Hadoop, as a pioneer in Big Data technologies quickly became popular within scientific community. Zhao et al. [140] conducted experiments on Hadoop to improve scalability of Collaborative Filtering algorithm. Another domain that craved for efficient Big Data solutions is GIS. High availability of ubiquitous positioning technologies has enabled capturing spatial data at an unprecedented scale and rate, which lead to new challenges in data processing. Aji et al. [5] used Hadoop as a data warehousing system for running large scale spatial queries. Their comparative experiments showed that performance of Hadoop based GIS outperforms traditional SDBMS for computationally intensive queries. Bioinformatics researches and scientists are confronted with the issue of Big Data analysis which will only increase in the coming years. Taylor [116] published an overview of the Hadoop framework and its applications in bioinformatics in 2010. In his work he extensively elaborate on diverse use cases of Hadoop and similar frameworks in bioinformatics, highlighting the need for such tools in the domain where the data is growing and the underlying algorithms are complex.

Apache Spark if often a choice when working with very large data sets, because of its high performance on in-memory distributed data processing. Dang et al. used Apache Spark to build a framework for mobility data analysis [40]. Main focus of their work is on building software products to provide mobility intelligence from spatio-temporal data where data sets are large, diverse and time evolving. To meet the performance demands while handling both batch and stream processing modes, they used Apache Spark platform. Increasing volume of available spatial data lead to the development of novel applications which introduced new challenges in spatial data analytics. The first challenge is related to system scalability, the system must be capable to effectively digest, store and retrieve the data when necessary, while in the same time must be able to deliver the response to user's request in no time - which is second challenge, interactive performance. In the research paper [135], Yu et al. presented the *GeoSpark*, an in-memory cluster computing system for processing large-scale spatial data. *GeoSpark* extends the core of Apache Spark to support spatial data types, indexes and operations, enabling the good run time performance over massive spatial data. Besides spatial data, there are other data sources that provide very large data sets. One example is the data from mobile devices. Mobile devices became global sensing systems, generating exabytes of data per month. Such data contains useful information for solving many security, advertising, urban monitoring, healthcare, etc. problems. Alsheikh et al. used Apache Spark to build a framework for distributed learning [8]. They presented the usage of Apache Spark to train deep learning models on large samples of data from mobile devices.

Apache Spark and Apache Hadoop are often compared between each other performance wise, because both frameworks are used in similar use cases, when data is to larger to be handled by traditional computing methods, but there are some key difference between them. Apache Hadoop is built mainly for batch processing, when we need to exploit Hadoop Distributed File System (HDFS) in order to handle very large data sets, whereas Apache Spark supports batch, real-time and streaming processing. Apache Spark is based on the principles similar to those of MapReduce engine [42], but it's main advantage in processing speed is due to in-memory processing, while on Hadoop the processing is on disk. Because of in-memory processing, Apache Spark is limited by memory and should be used only in those infrastructures where sufficient memory is available. One of the main advantages of Apache Hadoop is it's distributed file system, HDFS, while Spark uses other platforms for storage. However it is possible to use Spark over HDFS [115], but it can process data that are stored also in HBase, Cassandra and Hive. Truica et al. [118] used Spark together with Hive to perform community detection over CDR data. When working with Big Data that comes from live, user or system generated data source such as mobile CDR data, Internt traffic data, IoT data, etc. scalability is mandatory. Both Apache Spark and Hadoop are easily scalable systems, where scalability is achieved by adding more nodes to the cluster with one notice that in the case of Spark those nodes need to be memory rich.

Witayangkurn et al. [128] proposed a data analysis pipeline for CDR data based on open-source solutions for Big Data. The pipeline that they built relies on Apache Hadoop ecosystem for data intensive tasks. Complementary to CDR data, mobile phone trajectory data also requires Big Data solutions to exploit the full value of it. Yang et al. [133] proposed a framework based on Apache Spark to process and anonymize mobile trajectories. Qin et al. [94] presented a case study of "smart tourism" that utilizes CDR data to provide real-time information about tourists activities and Big Data processing framework based on Apache Spark SQL engine.

Big Data technologies have shown great promise for large scale data analytics, including spatial data too. Spatial data is traditionally considered as "big" data, but with the rapid growth of available spatial data in the recent years it expanded not in the size of single record but in the overall volume of records. Shangguan et al. [106] developed software development kit that exploits powerful capabilities of Spark, but with spatial data - SparkSpatialSDK. Before Spark was extended to work with spatial data, there were some efforts to exploit Big Data frameworks for spatial data analytics. Aji et al. [5] presented Hadoop-GIS - a scalable and high performance spatial data warehousing system for running large scale spatial queries on Hadoop. In their paper they highlighted two major challenges for massive spatial data analytics: i) volume of spatial data and ii) high computational complexity of spatial queries. Eldawy et al. [45] went even one step further in enabling Hadoop to work with spatial data, they built SpatialHadoop, which is MapReduce framework with native support for spatial data.

For few years now, the commercial sector is widely using and developing Big Data solutions based on Apache Spark, while the scientific community is still behind. The lack of scientific research and benchmark studies related to Apache Spark and similar Big Data frameworks, lead to growing gap between science and practice. Brdar et al. [24] investigated the issue of Big Data processing and analysis through the lenses of mobile cellular networks. They elaborated on diverse data types and sources, Big Data architectures, analysis and data fusion, with the focus on novel frameworks such as Apache Spark and Hadoop.

While Apache Spark is entering scientific research in a modest way, the industry is using it to great extend for diverse data intensive tasks and applications. Many well established companies such as Amazon, Nokia, many Telecom operators and new data oriented companies are using Apache Spark to develop their solutions [4].

[4] https://spark.apache.org/powered-by.html

Chapter 4

Evolving connectivity networks

Telecom connectivity networks reflect the way people are communicating using voice call. When one person calls another, the call is routed through the network to deliver the signal from originating base station to terminating. Originating base station is in shortest geographical proximity to the caller, while terminating base station is in the shortest geographical proximity to the callee. When building telecom connectivity networks we are using only originating and terminating base stations, the intermittent base station used to route the call through the network are not considered in this research. Telecom connectivity network are specially interesting because they conceal the information about peoples inter connections in space and time. As network evolve through time, so the inter connectivity structures are changing also. In this chapter, we will describe the nature of time evolving networks, how they can be measured and evaluated and finally how the network community structure forms and changes through time. Also, in this chapter we will present how day time/type effects connectivity network and how location semantic is related to the dynamics of connectivity network which all refer to our first and second hypothesis.

4.1 Introduction

Connectivity graphs inferred from mobile phone data uncover pulse of human interaction. In the recent years many innovative applications based on this rich data emerged, such as urban sensing, transport planning, social analysis and monitoring epidemics of infectious diseases [18] [78]. Anonymous mobile communication data from telecom operators can be utilized for sensing activities occurring within a city and can further fit into wider vision of smart cities that aims at monitoring and optimizing urban landscapes. Several studies explored mobile phone data in the context of urban sensing. Cici et al. analyzed aggregated cell phone activity per unit area that allowed them to detect seasonal patterns (weekday/weekend), anomalous activities and to segment a city into distinct clusters [34]. In another study, authors examined interactions among city inhabitants and visitors and identified the city's hotspots [11]. Mobile phone data can be also used to derive city land use information [91]. Here we utilized graph theory to study connectivity patterns on a city scale. We focused on the dominant backbone of networks - the most significant part of overall communication interaction. Pairwise communication was aggregated over spatial units of a city and one day time intervals and analysed throughout two months period. In our graphs nodes are spatial units and links were drawn if communication strength between units was significant. This allows us to study the backbone connectivity graphs as evolving structures and to examine temporal and spatial dynamics within a city. As network evolves through time due to fluctuations of users activity, connectivity graphs are changing as well and by measuring the graph properties we can unveil the hidden patterns of communication. We measured global and local graph properties and here present a part of obtained results.

4.2 Data

Mobile phone service providers collect large amount of data for every user interaction. Every time a user makes interaction using mobile phone (SMS or call), one Call Detail Record (CDR) is created in Telecom operator database. As described in 1.1, Telecom Italia opened the set of activity and connectivity CDR data for the purpose of BigData challenge. In this research we used CDRs that refer to the voice communication inside city of Milan for a time period of two months (November and December 2013).

Telecommunication interaction between mobile phone users is managed by Radio Base Stations (RBS) that are assigned by the operator. Every RBS has unique id, location and coverage map that provide approximate user's geographical location. CDRs contain the time of the interaction and the RBS which handled it. In available data collection CDRs are spatially aggregated on the grid containing 10 000 cells and temporally aggregated on time slots of ten minutes. We used telecommunication interactions set that comprises measured intensities between different cells. Only cells that spatially intersect with administrative area of Milan city were selected. The city is divided into 88 administrative zones[1]. The map with zones is presented in Figure 4.1, where each zone is represented by unique id placed in its center.

Urban zones are divided by distance from 'Piazza del Duomo' and represent homogeneous neighborhoods in terms of social composition and structure [30]. The investigated area includes Milan's municipality with total population of nearly 3.2 million residents and

[1] City of Milan public data

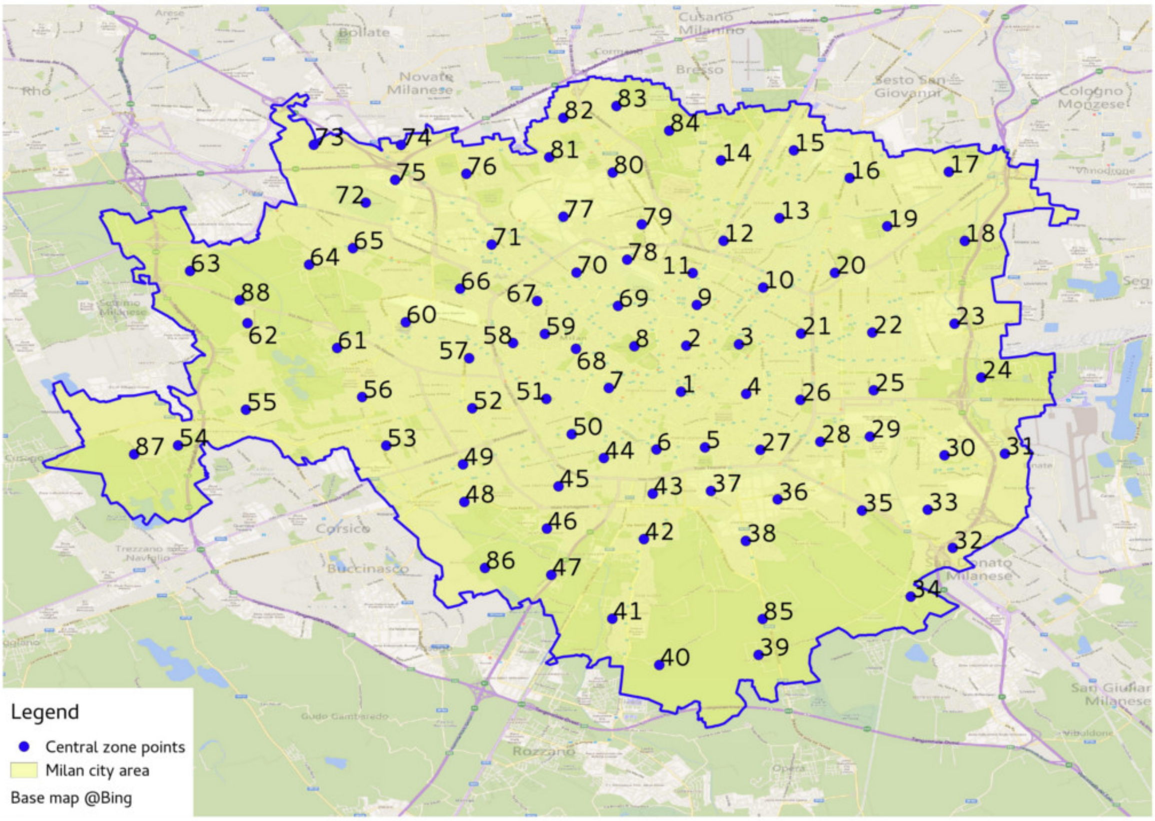


Fig. 4.1 Administrative zones in Milan city

total surface area of 1576 square kilometers. Urban zones are formed in the way to reflect similarity in socioeconomic structure, population distribution, work and income profile. Carlucci et al. [30] described in their work the population growth in the urban area of Milan city from 1976 to 2017, and emphasized the population density change in the specific city zones which indicate even greater importance of local neighbourhood analysis.

4.3 Methodology

To create connectivity graphs, we further aggregated communication from grid cells to 88 zones of Milan. Each grid cell is assigned to corresponding zone, thus our graphs refer to zones interactions. The connectivity matrices for 61 days were made in pairwise manner. Matrix element on the position [i, j] represents aggregated communication strength between zone i and zone j. After creating the connectivity matrices the filtering was performed to eliminate weak links while preserving the core backbone network structure. Graph filtering method is previously described in the Section 2.3.2. In this research we used the Disparity filter to denote significant links. We can recall to the equation 2.4 used with the disparity filter, where α denotes significance threshold, p_{ij} is the probability of having link between nodes i and j, and k is the number of nodes.

In our case $k = 88$ and α was set to 0.05. After the weak links were eliminated, graph structure for each day was created from remaining links in connectivity matrix. The final graphs are sparse and suitable for visual analytics. We presented links with QGIS and selected four typical graphs, Figure 4.2. The first is typical weekday, where the strongest communication links tend to appear and concentrate near city center and maximum communication strength is much higher than observed on the weekends. The second is the holyday (Friday 2013-11-01), where the strongest communication links tend to disperse and the maximum communication strength is very low. The third graph presents typical weekend. The strongest communication links tend to disperse across city but the maximum communication strength is higher than on the holyday. Finally, the Christmas day is presented in the fourth graph. Its structure is similar as the one presented for another holiday. The strongest links are dispersed across residential parts of the city and the overall communication strength is low, which is typical for holydays.

Along with visual inspection of graphs across two month period we quantified graphs properties and did deeper analysis of their changes during time and identified interesting weekday/weekend distinctions. We performed both, global and local, graphs analysis [36].

4.4 Results

Global graph properties provide information on a global structure and further allow comparisons among graphs. We calculated the number of edges, maximum weight, radius, diameter, max clique size, average clustering for all inferred graphs. We compared global graph properties on different day types: weekdays and weekends and discovered differences. We observed that number of edges is higher on the weekdays than on weekends, Figure 4.3, which is expected result if we consider that sever amount of voice traffic is happening due to work driven reasons. Distributions of measured diameters unveil that weekday graphs have lower diameter compared to weekend, indicating faster information flow during work days and larger “connectivity distance” in weekends.

Local graph properties uncover localized patterns in the graphs, considering local connections of each node. We measured numerous local properties such as clustering coefficient, node degree, page rank, betweenness centrality, etc. Here we selected three nodes (Zone 1, 71 and 23, see Figure 4.4) and presented changes in time of betwenness centrality and PageRank.

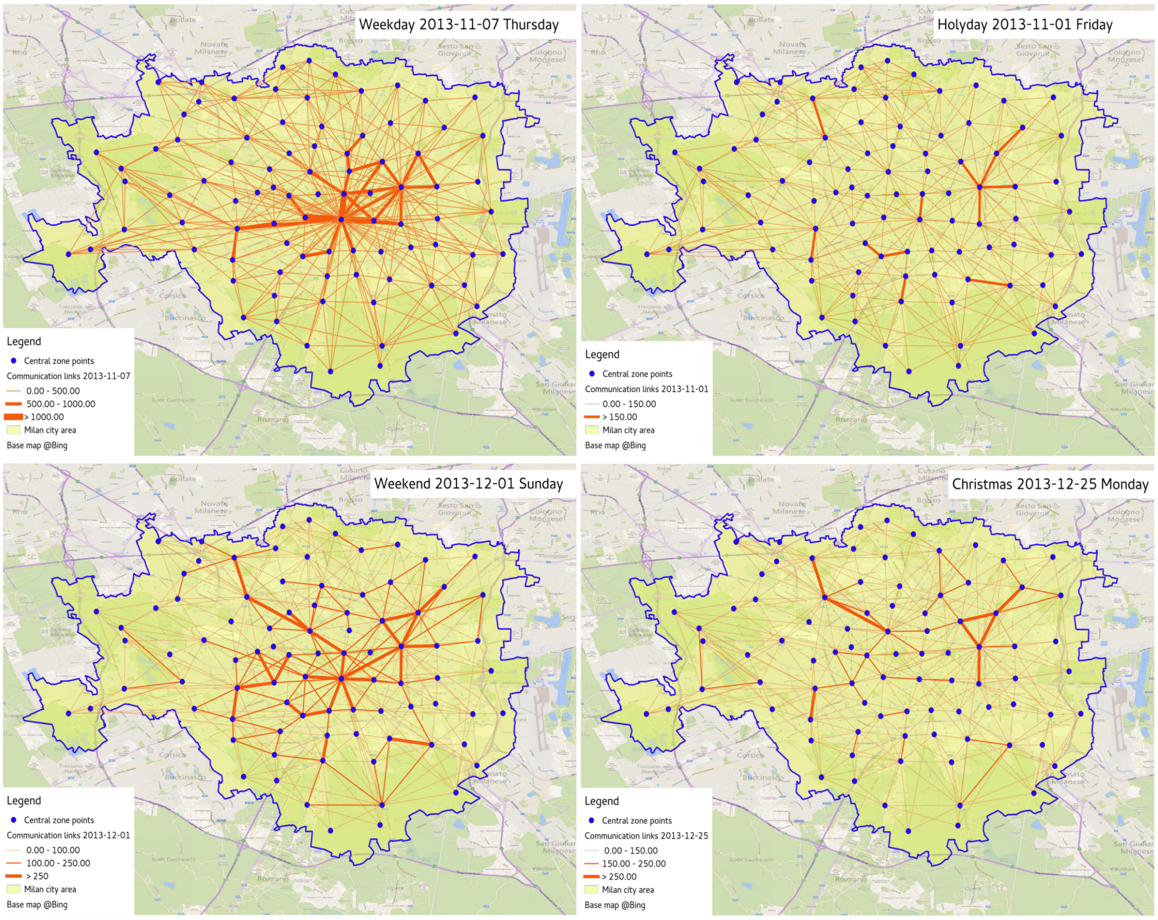


Fig. 4.2 Backbone connectivity graphs in the city of Milan

Betweenness centrality is a graph centrality measure that calculates the number of shortest paths that pass through examined node. In general case, for every pair of vertices there is at least one shortest path between the vertices such that the number of edges that the path passes through is minimal. In the case of weighted graphs, sum of the wights of the edges in the path should be minimal. Betweenness centrality is the most general measure of centrality in the graphs, applied in wide variation of problems. In network science it represent the degree to which the nodes stand between each other. In the case of telecom connectivity networks, betweenness centrality is an indicator of how much control each node have over the network. The higher the betweenness centrality is, more information is passing through the examined node, meaning that location of related RBS is either extremely important in the city or that something extraordinary is happening there at the time.

Measurements of the betweenness centrality for selected nodes 1, 71 and 23 are presented in Figure 4.5. Zone 1 had very regulated periodical pattern where we can clearly distinguish between week days and weekend. During work days there is high, stable values

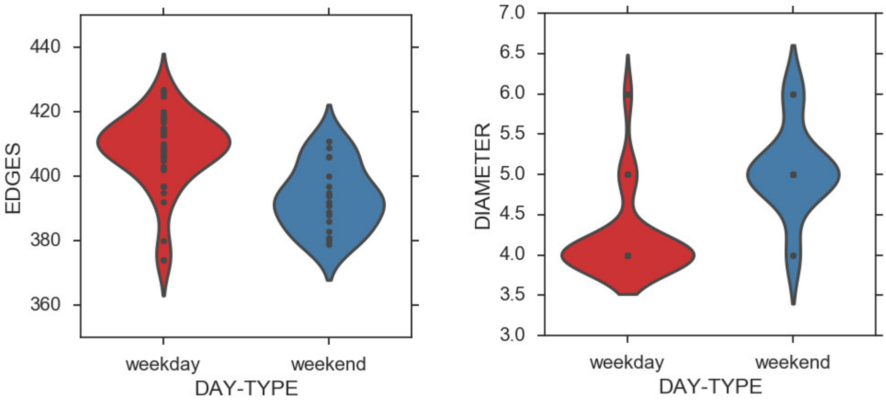


Fig. 4.3 Edges diameter

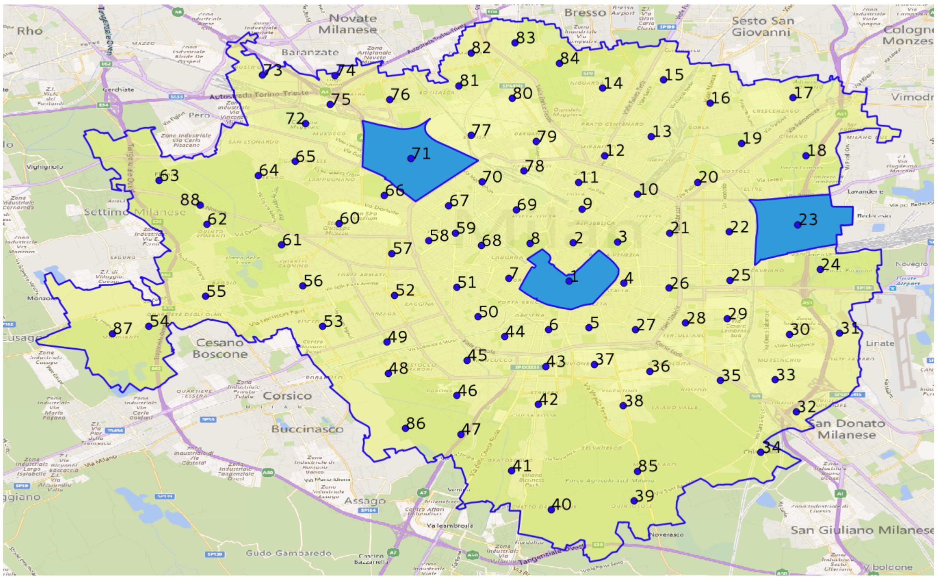


Fig. 4.4 Zones 1, 71, 23

of betweenness centrality, while it significantly drop during weekend, indicating that this location is more work driven than residential. This comes as no surprise since the Zone 1 is the core city centre of Milan city, and it is touristic, administrative, shopping, retail and catering centre. It is important to highlight that absolute values of betweenness centrality for the Zone 1 are significantly higher comparing to Zones 71 and 23, even during "low activity" days, meaning this location is important for the overall network no mater of the days pattern. Zone 71 has high fluctuations in betweenness centrality indication that the activity in this area is dynamically changing. Some higher picks can be observed during weekends and holydays (end of the December), but without firm regular structure. Such pattern of betweenness centrality indicate "come and go" locations, where there are no facilities to attract more structural communication flow. Zone 23 has very interesting pattern of betweenness centrality, the values are very small and stable indicating there is no much activity in the area in general. Nevertheless, Zone 23 that covers part of Lambrate district has unusual jump in betweeness centrality at 15th December, that could be due to event The Lambrate Bicycle Film Festival.

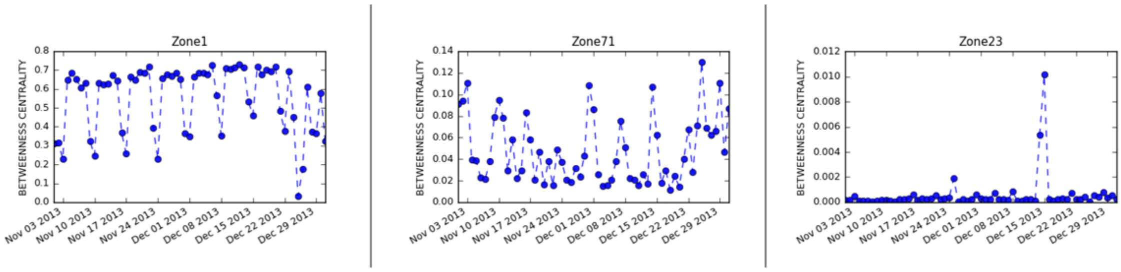


Fig. 4.5 Betweenness centrality

Page Rank is common measure to evaluate the importance of a node in the graph. Page Rank algorithm [87] is used in Google Search engine to rank the web pages, therefore the Page Rank is a way of measuring the importance of the website pages. Page Rank works by counting the links that point toward a node in a graph, considering the quality of a link but also considering the links that node makes toward other nodes in a graph. Simplified, the Page Rank algorithm will evaluate how well the node is positioned in the graph. Higher Page Rank indicate the more important nodes, so called "hubs" in the graph. Original Page Rank algorithm is developed for unweighted graphs, but the algorithm is later modified to consider weighted edges as well [131].

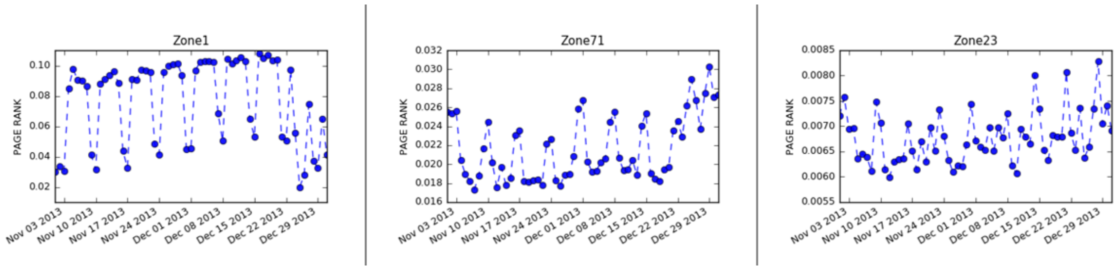


Fig. 4.6 Page Rank

Results of Page Rank for selected nodes 1, 71 and 23 are presented in Figure 4.6. Zone 1 Page Rank has similar behaviour like betweenness centrality, where we can clearly notice periodicity and work day - weekend pattern. Higher values of Page Rank indicate the node is a "hub" in the graph, meaning that Zone 1 has central role in the network, which is well aligned to geographical position of Zone 1 - central position in the city. Zone 71 also shows similar behaviour for Page Rank like for betweenness centrality, there is significant pick during weekends following the noticeable drop during work week, while overall higher values occur during holydays (end of December). Interestingly, Page Rank pattern for Zone 23 differs considerably form betweeness centrality implying that different graph properties can provide complementary information. There are high fluctuations in Page Rank values for Zone 23, without clear periodicity. Occasional picks that are occurring on specific days need additional investigation to be explained properly. Overall smaller values of Page Rank for Zone 23 indicate the zone has less central position in the graph, which is also aligned to its geographical position. Zone 23 covers part of Lambrate district in the Northeast area of Milan city 4.4 which is traditional industrial part of the city that does not attract higher mobility flow of people contrary to Zone 1 which is historical city centre.

4.5 Discussion

Our analysis of the connectivity backbone networks in the city provided new insights into social interactions and their changes across the city zones in different day types. Through the lenses of graph theory we discovered properties that can serve for detecting the patterns and deviations from typical observations. We presented how graph properties can be used as a measure to evaluate the change in communication patterns through time and space. We also examined how specific local graph properties such as betweenness centrality and Page Rank can indicate the importance of node in the communication network but also the geographical context of the location that the node represent. Our findings imply the promising potential of graph analysis in the telecommunication network, that could help reveal the underlying patterns of human interactions, develop new services, assist urban planning and decision making, etc. Further analysis and examination would include more graph-based formalism in identifying strong temporally consistent links, patterns of change and evolving graph sequences [10] as well as unveiling underlying social pulse that is reflected in mobile phone data.

Our results showed that depending on the day type, whether it is a working day, weekend or holiday, the connectivity network will form differently, the strongest links with most intensive communication will occur at different places which indicate our first hypothesis, how day type effects human dynamics because those connectivity links are formed from user generated telecom traffic. Also, the location semantic plays a crucial role in the dynamic of connectivity network because the strongest links are associated to densely populated urban area where the historical center of the city is. Also, how location semantic matter shows our experiments with graph properties that showed unusual high peak in the area where at the moment of peak the social event was happening. Our experiments proved that location semantic really effects the forming of connectivity networks which indicate our second hypothesis.

Chapter 5

Community detection in telecom connectivity networks

Community detection is one of the most popular topics of modern network science [12]. It is based on the concept of graph clustering described in the Section 2.4. In the recent time, the increased popularity of social networks, severe growth of the Internet and other similar systems that include components that interact together shed a light to this research topic due to the need for computationally efficient solutions. Communities are essentially the subsets of graph nodes that form the strongest links between each other, inside the community, comparing to the links they form with the nodes outside of the community. This is very important property of the network because it shows the potential to information spreading across the network and within the specific groups. In the case of telecom networks, where the data is georeferenced, the formed communities are also related to specific geographical space. Nevertheless, community detection is complex problem in graph mining and requires great computational power for the large networks. In this chapter, we presented the results of community detection of telecom connectivity network and proposed the solution based on HPC to increase the computational performance. Our results showed that forming of communities is strongly correlated to location semantic which indicate our second hypothesis.

5.1 Introduction

Community detection is based on graph mining, which is a heavily active research direction with numerous applications [4, 35]. The aim is to discover the meaningful communities in a large networks [46]. In the majority of real-life applications, graphs are extremely sparse usually following power-law degree distribution. However, the original graph may contain groups of vertices, called *communities*, where vertices in the same community are more well-connected than vertices across communities.

A network example with communities is depicted in Figure 5.1. Evidently, the number of communities depends on the criterion being used to differentiate between communities. Note that this is also the case for clustering algorithms, where different definitions of the cluster concept lead to different clustering algorithms. This means that different techniques in general produce different results.

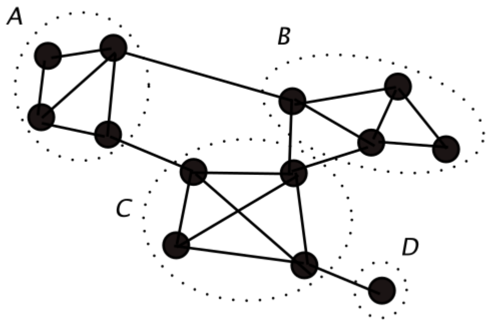


Fig. 5.1 A network example with four possible communities.

The efficiency of community detection algorithms is heavily dependent on the size of the input graph, i.e., the number of vertices and/or the number of edges, and also on its structural complexity. In addition to the main processing task that must be performed, preprocessing is also a significant step that in many cases is computationally intensive. To handle both preprocessing and main processing efficiently, a potential solution is to use multiple resources and apply parallel or distributed computing techniques, aiming at reducing the overall processing time.

In this research we focus on the analysis of real world telecom connectivity network, based on CDR data. CDRs are in general large in size and form complex, very dense graph structure which impose great computational and algorithmic challenge for community detection analysis. To overcome this issue we need to use advanced Big Data technologies such as Apache Hadoop [109], Apache YARN [122], Apache Spark [66], and Apache Hive [117].

In our previous work [82] we have examined time evolving structure of telecom connectivity networks. In that work we have used a conventional DBMS and Python to extract knowledge from raw telecom data, which is described in detail in Chapter 4 of this thesis. The experiments were very time consuming and we could analyze only subset of the data in due time. Also, with such computational architecture we were unable to analyse very dense graph structures that are formed from the original data distribution, and to proceed we needed to perform the additional space aggregation to make graphs less dense. The trade off between graph density and spatial resolution was acceptable for the analysis of underlying graph patterns, but it wouldn't be of use for the community detection application. However, the results that we obtained have motivated us to apply different processing techniques over the telecom data, in order to speed up the experimental evaluation and to be able to analyze the complete data set from the original distribution without additional space aggregation.

The new approach for analyzing the data is based on Apache Spark. We measured the time needed for processing the data and compare it with our previous experience. The improvement in time is very significant. For example, data preprocessing required several hours, whereas with Spark, the preprocessing runtime is in the order of minutes. In general, our implementation considers the complete pipeline, from preprocessing the raw data to knowledge discovery. All necessary tasks are executed within Spark and results are stored in Hive, without the need to ETL data from one platform to another. Comparing to other processing methodologies that we have used, it is highly justified to use Apache Spark for telecom data analytics, and specifically, for community detection.

Besides using the advanced Big Data technologies to speed up the processing time, we performed graph sparsification through filtering which lead to more efficient discovery of communities since runtime depends heavily on the number of edges in the graph. On the other hand, by comparing the communities generated with and without filtering we observe that communities remain relatively stable in comparison to the ground truth (unfiltered graph).

5.2 Data

We performed community detection over connectivity data provided by the Semantics and Knowledge Innovation Lab (SKIL) of Telecom Italia [13], previously described in Section 1.1. The data refers to connectivity inside Milan city for two months period (November to December 2013). The data is preprocessed by Telecom Italia before releasing, to preserve the privacy of the users as well as sensitive telecom company information such as the exact location of the RBSs. Instead of exact locations, the RBSs are represented by cells in the regular grid, Figure 5.2. Telecom traffic from the real RBS network is aggregated and normalized to the level of regular grid, so we can observe each cell in the grid as one RBS in the real network. Therefore, voice call connectivity links that are occurring between each two RBS are now spatially spread so they form the link between each cell in the grid.

The grid consists of 10 000 cells, which reflect to 10 000 nodes in the graph of telecom connectivity network. Edges are now telecom connectivity links between each grid cell.

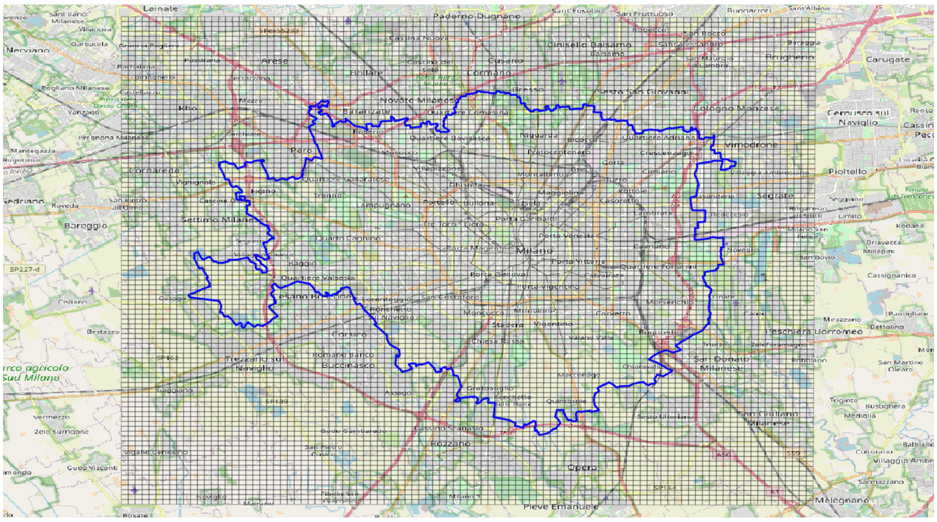


Fig. 5.2 Milan city telecom network grid

Graphs formed in such way are very dense, almost fully connected which adds to the computational complexity.

Community detection is done using the Louvain algorithm [21]. To minimize the graph and keep only the important nodes without losing the communities we applied filtering based on the disparity filter, Section 2.3.2. In our case, the number of nodes is 10,000. The value of the parameter α_{ij} (Equation (2.4)) defines the level of filtering. By changing the filtering level we obtain graphs with fewer edges. We are using these graphs to demonstrate the performance of community detection as we grow the number of edges in the input graph.

Since the data refers to the real world communication that was registered by telecom provider, each link in the graph is associated to its *weight* which is based on the number of calls exchanged between different areas of the Milan city. Original data is represented in the form of directed graph. For our experiments, two transformations are applied on the original data: i) the directed graph is transformed into an undirected graph where *weight* for both directions is summed and represent the *weight* of the undirected edge, and ii) the data is temporally aggregated on daily basis level (original data set has 10 minutes time resolution).

5.3 Methodology

Complex analysis such as community detection in telecom networks require usage of advanced HPC systems to deliver the results. On of the most frequently used frameworks for HPC, both in academic research and industry is Apache Spark.

Apache Spark is a unified distributed engine with a rich and powerful APIs for different programming languages [66], i.e. Scala, Python, Java and R. One of its main characteristics is that (in contrast to Hadoop MapReduce) it exploits main memory as much as possible, being able to persist data across rounds to avoid unnecessary I/O operations. Spark jobs are executed based on a master-slave model in cooperation with the cluster manager YARN. Apache Spark is computing framework that operates in distributed environment, while enabling users to develop models using high level programming API. Apache Spark can connect to diverse data sources, from relational databases, to HDFS and streaming data lakes, which makes it very flexible to build the computational architecture around it.

The proposed architecture for community detection in telecom networks is based on the Apache Spark framework with programming language Scala [124] and GraphX library [130]. We utilized Spark's DataFrame [32] as the most convenient data structure and HiveContext to exploit HiveQL [29] as query language.

Aggregated data is stored in the Apache Hive data warehouse which is installed on top of HDFS. DataFrames are used to load semi-structured telecom data, since our original corpus is in tsv format. HiveContext is used to enable SQL like semantic for data analytics queries. GraphX library is used to perform specific graph analytic tasks, such as community detection, since we are exploring connectivity data. The cluster resource manager used is YARN. Our methodology utilizes the following pipeline (Figure 5.3):

1. CDRs are aggregated in such a way that each graph node corresponds to a spatial area. This task has been performed by the mobile operator before releasing the data.

2. The original directed graph is aggregated to obtain an undirected one, as the orientation of edges is ignored in our case.

3. Filtering is applied in order to sparsify the network. To avoid the generation of a complete graph, links with small weight (i.e., number of mutual calls) were removed.

4. Community detection is applied using the LOUVAIN algorithm.

5. Visualization is applied.

From loading raw original data into Hadoop ecosystem, all computationally intensive tasks are performed within the system to obtain the best performance. Last step that includes visualizing results on geographical maps is performed using QGIS software since Spark had a weak support for spatial data visualization at the moment we created the proposed pipeline.

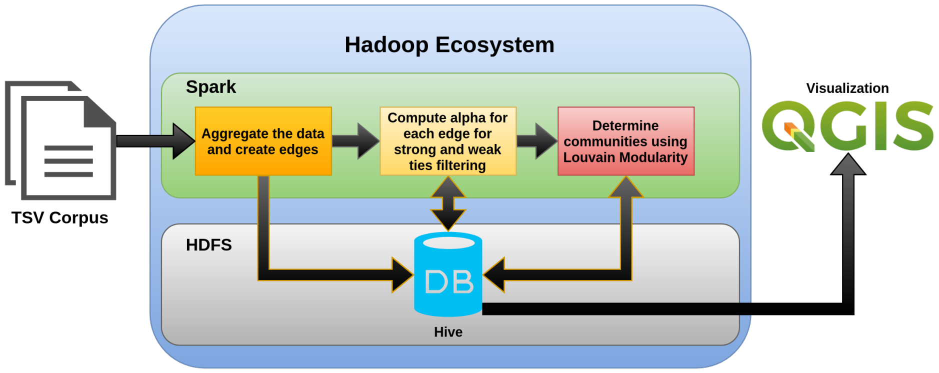

Hadoop Ecosystem
Spark
TSV Corpus
Aggregate the data and create edges
Compute alpha for each edge for strong and weak ties filtering
Determine communities using Louvain Modularity
Visualization
QGIS
HDFS
DB
Hive

Fig. 5.3 Architecture

5.4 System architecture and performance

Performance is the main bottleneck when applying complex analysis such as community detection in very large and dense networks. To enhance the performance while keeping the high accuracy of the results we can perform two things: *i*) first, we can apply graph filtering to sparsify the graph structure while keeping the most significant edges, therefore making the graph "lighter" for algorithm to handle it and *ii*) second, we can use some HPC system to enhance the algorithm runtime.

To get more understanding of how filtering affects community detection in both terms of algorithm runtime as well as the quality of the results, we performed the experiments to test the community detection with and without filtering, using Apache Spark. We performed the experiments over two diverse system architectures:

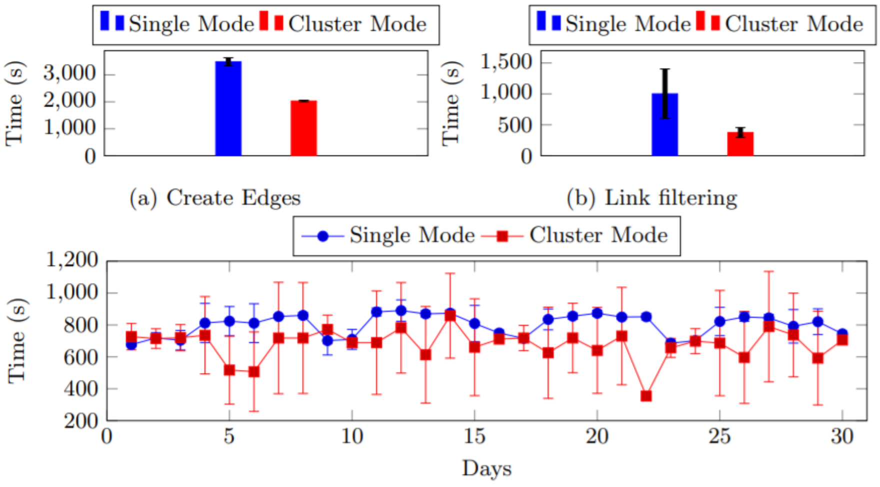
1. Multi core server machine running CentOS with 16 Intel Xeon E5-2623 CPUs with 4 cores at 2.60GHz, 126 GB RAM and 1TB HDD,

2. Cluster with 6 nodes running Ubuntu 16.04 x64, each with 1 Intel Core i7-4790S CPU with 8 cores at 3.20GHz, 16 GB RAM and 500GB HDD.

The Hadoop ecosystem is running on Ambari and has the following configuration for the 6 nodes: 1 node acts as HDFS NameNode and SecondaryName Node, YARN ResourceManager, and Hive Metastore and 5 nodes, each acting as HDFS DataNodes, YARN NodeManagers, Spark Client, and Hive Client.

For the experiments we fix the number of Spark executors to 16 with one vnode and 3GB memory each. The same settings have been used in both the Single Machine and Cluster Mode. Each experiment was run 10 times to compute an average and standard deviation for the time performance. We set the value for the Spark executors memory to 3 GB per executor, we activated 16 executors and set number of processors per executor to 1. We used the same setting in multi core single machine and in cluster. For each step of the processing pipeline we made Scala code, created .jar artifacts from the code, and submitted job to Spark using command line. For coding and debugging we used IntelliJ IDEA, and for building artifacts we used SBT. Submission of spark jobs was done through command line using .sh scripts.

The first experiment measures the performance of edge generation from raw data. The second experiment measures the performance of computing α values that denote statistical significance of the edge based on the disparity filter (see Equation (2.4)). The third experiment is related to the performance of Louvain algorithm over a filtered graph where $\alpha = 0.05$.



(a) Create Edges

(b) Link filtering

Fig. 5.4 Runtime evaluation results.

For each experiment we measured the average runtime over a series of 10 runs. For the first task (creating edges), the cluster environment showed 42% better runtime then the single mode case, whereas the standard deviation is relatively small in both cases. For the task related to computing α values, the cluster showed again better performance, since the computation is 63% faster. In this case, the standard deviation is higher for both environments in comparison to the previous task. Finally, the performance of community detection using Louvain algorithm differ for each graph, but in most cases the cluster showed the best performance. In most of the cases the mean values of runtime for Louvain algorithm are higher in the case of single machine, whereas the standard deviation is higher in the cluster environment. We hypothesize that these results are a direct consequence of how YARN's ResourceManager schedules the ApplicationManager and NodeManager and Spark's directed acyclic graph (DAG) execution engine optimizes the graph construction for the GraphX library. A comparison between statistics is illustrated in Figure 5.4.

5.5 Results

In addition to the main processing performed for community detection, a visualization is also applied for the illustration of the results. So far, all the steps of the pipeline are executed inside the Apache Spark engine. Different levels of filtering and visualization are done over graph made for Thursday, November 8, 2013, as the graph made for that date contains the highest number of edges.

For community visualization, the QGIS software has been used. We have chosen to present the set of communities generated by using the 8^{th} of November, because the network graph for this particular day contains the highest number of edges. Experiments are performed on the cluster environment for three different threshold values, i.e., $\alpha = \{0.001, 0.01, 0.05\}$. We used the results of community detection performed over unfiltered graph as the ground truth result, and compared it with the results when the filtering was applied. After filtering is performed, the Louvain community detection algorithm is executed using Apache Spark. The first level of filtering eliminates more then 50% of edges, while the runtime for Louvain clustering algorithm improves with a factor of 2.46. When $\alpha = 0.01$ almost 70% of edges are eliminated, and the algorithm runtime improves with a factor of 3.7. Filtering with $\alpha = 0.001$ eliminates almost 80% of the edges, the algorithm's runtime improves by a factor of 6.88.

The number of communities changes when filtering is applied. The results for 10 tests are presented in Table 6.1. The Louvain community detection algorithm converges to the same result for the same input graph. The number of communities is higher for the graphs where the filtering is stricter. That is expected, because the higher level of filtering gives the graph containing the strongest links. Moreover, as the number of nodes stays constant, the removal of edges tends to create more communities. In Figure 5.5 we observe the centrality pattern of the clustering for each graph. The structure of communities is denser in the central area of the city, while in the peripheral parts communities are more spatially spread. That is due to overall higher traffic of people in city center, which reflects to telecom network. We observed also that the number of communities produced from graphs where the filtering with $\alpha = 0.01$

Table 5.1 Number of nodes and edges after applying different filtering levels, number of communities and runtime community detection.

α	# nodes	# edges	# communities	time(sec.)
1	10,000	29,099,392	175	2,229.73 ± 340.14
0.05	10,000	12,942,551	174	906.43 ± 279.79
0.01	10,000	9,003,404	176	603.20 ± 174.28
0.001	10,000	6,043,769	186	324.21 ± 127.73

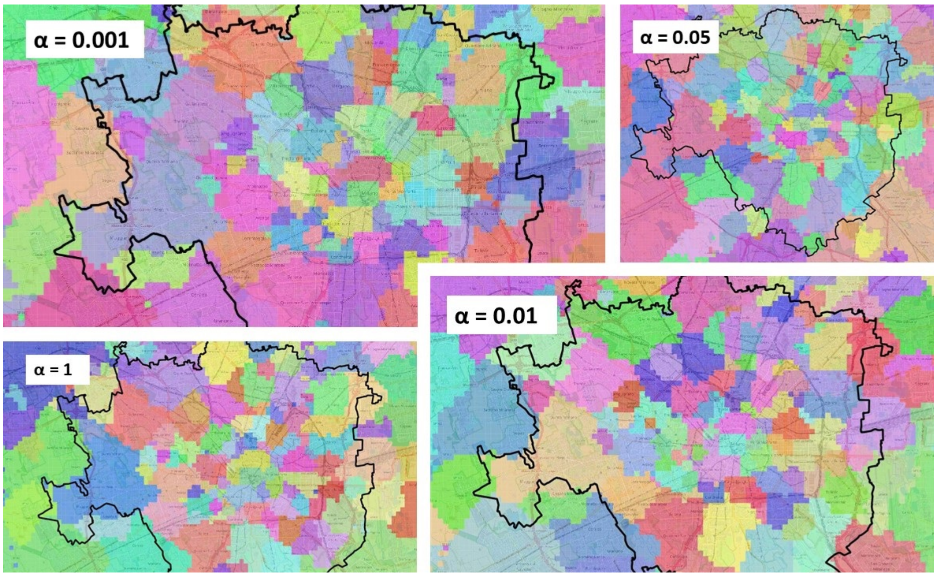


Fig. 5.5 Communities formed for different filtering thresholds.

and $\alpha = 0.05$ is applied does not differ much from the number of communities produced from the unfiltered graph. On the other hand, the runtime for filtered and unfiltered graphs differs significantly. Even with the less strict filtering applied, we get a major improvement in processing time, which justifies the use of filtering.

The community evaluation is done using *Purity*, *Entropy*, *Rand Index* and *Adjusted Rand Index* measures, described in Section 2.4.2. The results are illustrated in Table 5.2 and they are obtained by using Louvain algorithm without filtering as the ground truth. The *Purity* measures how many elements from the communities determined by the algorithm when filtering is applied belong to the communities from the ground truth. This measure is low because the number of communities differ between the ground truth and each different α threshold. The *Entropy* measures the probability of a node chosen at random to belong to a community. The high results of this measure describes how much we can, on the average, reduce the uncertainty about the cluster of a random element when knowing its cluster, the ground truth, in another clustering of the same set of elements. The *Rand Index* provides

Table 5.2 Cluster evaluation for different threshold values.

Evaluation Measure	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$
Purity	0.055	0.052	0.048
Entropy	0.902	0.909	0.912
Rand Index	0.981	0.985	0.987
Adjusted Rand Index	0.662	0.745	0.781

information on how a new clustering is compared to a correct one, the ground truth, and it is highly dependent on the number of clusters. This measure is high for all the tests because the *Rand Index* converges to 1 as the number of clusters increases which is undesirable for a similarity measure [49]. To address this problem we also computed the *Adjusted Rand Index* which shows a clearer classification for the filtering technique. For the threshold $\alpha = 0.05$, as well as for $\alpha = 0.01$, the number of communities remain stable and closer to the ones detected using the ground truth, although the number of edges decreases significantly (see Table 6.1).

5.6 Discussion

Mobile phone records offer many potentials for knowledge discovery with significant impact. In particular, community detection is a task related to networks and aims at the discovery of groups of nodes that are densely connected. In general, community detection is solved by executing graph clustering algorithms, which is a computationally intensive task, and therefore scalable algorithms should be applied to guarantee efficiency for large networks.

To explore the possibility of using community detection to evaluate the structure of real-world telecom connectivity networks we focused on applying community detection in a distributed environment. The first results has shown that parallelism is an important tool to attack scalability issues, since we can analyze larger graphs by using a cluster of machines. Moreover, we have shown that by applying sparsification through filtering, we may boost performance even further without penalizing the quality of the community detection result.

Our results showed that by using advanced HPC technologies we can increase the computing performance when running community detection algorithms over complex networks. We also showed that by applying graph filtering before community detection we can speed up the computation time while not damaging the core graph structure and the output result. Also, our results showed that communities form differently over different locations, for example, the communities formed over city center are smaller in size, but there is higher number of them, comparing to the communities formed over the peripheral parts of the city where they are larger in size, but more stable, which indicates our second hypothesis that location semantic matters in human dynamics.

Chapter 6

Human dynamics evaluated through telecom connectivity networks

Human dynamics has been the topic that challenges many diverse research areas for a long time. With an increasing number of devices and services that collect data about people activity, interactions and mobility in the last few years, human dynamics research became the key topic in computational social science [77]. Human dynamics research evolves with changing every day circumstances in which people live such as natural and urban environment, emerging new technologies, climate change and society. High presence of modern information and communication (ICT) technologies including location-aware devices, various sensors and mobile technology in every day life have great impact on shaping human activity and interaction patterns [107].

Human dynamics is complex research problem that requires multidisciplinary approach to be completely understood. Our daily activities are overflowing with communication, through virtual and physically space, and with mobility and transport. The way we are communicating or moving in the urban environment is deeply entailed with time and space, an understanding of those two dimensions of human dynamic have unprecedented value.

In this chapter we will deep dive into the complex problem of human dynamic and inspect one aspect of it, related to human telecommunication connectivity and urban space. We will elaborate on how graph properties calculated from telecom connectivity graphs are associated with spatial semantic from the communities also evaluated from those graphs. We will inspect the interference of those two and elaborate on human dynamics that is implied. This chapter summarizes all five of our hypothesis, but at most it shows how properties of connectivity networks are strongly correlated with human dynamics and how those can be used to evaluate it through time.

6.1 Introduction

One of the richest and most detailed data sources about human daily based activities is mobile phone data [18]. Many diverse applications with significant social impact are developed based on mobile phone data, such as urban sensing and planning [15, 28], traffic engineering [7, 60, 27], predicting energy consumption [22], disaster management [75, 90, 127], epidemiology [23, 72, 126], deriving socio-economical indicators [89, 114], land-use detection [83] and many more. Such high number of diverse applications and thought-provoking approaches indicate the great potential of human dynamics research for society and economy. Nevertheless, despite the great impact human dynamics research has on social development its still relatively young research area that engages small research community with increasing number of publications through years.

Challenges of urban development inspired many researches to tackle the problem considering mobile phone data. More specifically, Ratti et al. in their review [97] highlighted the potential of using mobile phone data for urban planning. Soto et al. [112] used *Call Detail Records* to extract the information to automatically identify land use behaviors in urban environments. They used fuzzy c-means to cluster the Radio Base Station signatures and detect the class representatives of the land use in urban environment. Grauwin et al. used mobile phone data to detect land use classes in three different cities, New York, London and Hong Kong [56]. Furno et al. conducted comparative analysis between ten different cities, in which they constructed specific mobile traffic signatures to determine dynamic patterns of human presence in urban areas [51]. Rios and Muñoz used a big mobile phone data set with 880 million records in a case study for Santiago, Chile for land use pattern detection [99]. They used the latent variable clustering technique in detecting clusters of residential, office area, leisure-commerce and rush hour pattern areas. Pei et al. used hourly relative pattern and the total call volume trough semi supervised fuzzy c-means clustering approach in inferring land use types in Singapore, showing that the accuracy decreased with the increase in heterogeneity of land use and density of cell phone towers [91].

Furno et al. combined simultaneously Call Detail Records and vehicle GPS traces for revealing land use context in French and Italian cities [50]. Unveiling complex ties between land use and human dynamics properties derived from mobile phone data is an active area of research. The latest results demonstrate relations between dominant land use for each Voronoi zone and corresponding human activity represented as aggregated CDRs [10], as well as land-use composition of city's neighborhoods and the time series of CDR intensities [16]. Both studies utilized clustering to group similar land uses on one side and human dynamics properties on the other side and finally estimated agreement between clustering results obtained from this two data sources. Another novel study conducted by Liu et al. examines the interaction between urban land use and commuting flows from mobile phone data based on the regression model [73]. Noyman et al. conducted the study that suggests a methodology of "reversed urbanism" to urban planning and decision making. The methodology considers human behaviour patterns extracted from mobile phone data as a key element of urban design and their association with the functionality of urban areas [85]. One recent study conducted by Cottineau et al. showed the relation of mobile phone data indicators such as number of calls, active days, duration of calls, entropy, etc. and socioeconomic organization of cities [37]. They showed how mobile phone data together with descriptive data such as census and administrative data could be used for urban development.

Mobile phone data generated by users communication and interaction are highly dynamic and usually very large in size. Analytical pipelines developed to extract the meaningful knowledge from such data sets could be computationally expensive, while at the same time delivery of the results needs to be efficient since many applications require almost real-time response. Brdar et al. [24] provided a broad overview of the entire workflow starting from raw data access, followed by demands for analytical performance and data fusion, to the final application. The authors highlighted the critical challenges in mobile phone data analysis that need to be addressed in order to disclose the hidden potential of the data. To make large scale telecom data analytics more efficient the work in [119] suggested using the Apache Spark [66] platform for distributed data analytics. Apache Spark is also a common data platform choice in Big Data applications in commercial domains such as retail, cloud based services, software development and telecommunications.

From this brief overview of the state of the art research related to human dynamics and mobile phone data we can conclude that human dynamics is challenging topic that needs to engage many different expertise to tackle exciting research questions which have significant impact on social, urban and economic development.

To extract meaningful knowledge from mobile phone data we need to address few issues. First, the mobile network data needs to be represented in a structure that keeps information about entities and relationships between entities, for which the most suitable structure is graph structure. Further, to analyze the mobile network we applied graph mining techniques [4, 35] such as community detection and centrality measures. Next, to unveil the relationship between land use and mobile network data and to predict spatial and network properties from land use data, we developed a mechanism based on Machine Learning. Last but not least, to be able to analyse large scale mobile network data in efficient manner we implemented the most demanding tasks in Apache Spark engine.

Based on our previous work which is in detail described in Chapters 4 and 5 and paper [24] we were able to propose a strong hypothesis that human dynamics reflected through telecom connectivity data is strongly correlated to spatial semantic, and that we can predict spatial and mobile network properties from land use data.

A core concept used in our proposal is based on community detection, which involves associating the nodes of a network into meaningful groups (also known as clusters or communities) [83]. Given a graph $G(V,E)$, where V is the set of nodes and E is the set of edges, the output of a community detection algorithm, in its simplest form, is a partitioning of V into c groups V_1, V_2, …, V_c where $\forall i,j$ it holds that $V_i \cap V_j = \emptyset$ [83]. Community detection is in detail described in the Section 2.4.1 about graph clustering.

Centrality measures is graph mining are used to quantify the importance of graph nodes. The most commonly used node centrality measures are *degree centrality*, *betweenness centrality*, *page rank* and *core number*. Among these measures, betweenness centrality and page rank are the most computationally intensive, whereas degree centrality and core number require linear time to compute. Detail description and formal definitions of these measures is presented in Section 2.3.1.

In order to guarantee scalability and efficient processing of large scale data our proposal utilize advance Big Data technologies. The most computationally intensive parts of our methodology, such as graph filtering and community detection, are supported by Apache Spark. Detailed overview of common Big Data technologies as well as Apache Spark is provided in Section 3.

6.2 Data

To unveil the relationship between human dynamics and land use in urban spaces, we are using telecom data for the city of Milan, the same corpus of anonymised CDR data that is used in our previous experiments described in Chapters 4 and 5. In this work we are combining both concepts of evolving networks from Chapter 4 and community detection from Chapter 5. In the following Chapter we will describe how we went even one step further and correlated the telecom data with spatial context and land use.

Telecom data is very rich in user behaviour information and contains very sensitive personal information such as someones activities, habits, movement, etc., which can be misused to reveal someones identity. Due to sensitive nature of the data, Telecom operators follow rigorous procedures for data anonymization to preserve the privacy of users, before sharing the data to third parties. Call Detail Records (CDR) can be anonymized by performing temporal and/or spatial aggregation. CDRs can also be temporally aggregated in predefined time slots such that all telecom traffic that occurred between two base stations is summed and presented as one weighted link. Spatial aggregation is performed to hide the exact location of RBSs in a way that the area is divided into spacial units where each unit is assigned the portion of telecom traffic that refers to units spatial area.

Telecom data that we explored is aggregated in both temporal and spatial dimension. CDR data is provided in the form of text files where each line represent aggregated telecom traffic that occurred between two corresponding spatial units in time frame of 10 minutes. Spatial units are squares of size 235 x 235 meters distributed in regular grid containing 10 000 cells that overlay the area of Milan city with surroundings. The data contains columns:

- timestamp in Unix time,
- square id 1 (outgoing square),
- square id 2 (incoming square) and
- weight that represent the aggregated traffic intensity.

Square ids are associated with their spatial coordinated in separate GeoJSON file that contains regular grid over city of Milan.

The data is further analysed following the methodology steps described in the next section.

6.3 Methodology

Telecom data is large scale data that is very demanding for processing. Only with careful design of analytical platform and methodology the full potential of the data can be exploited. To meet these requirements we have chosen the Apache Spark platform for running the most demanding tasks, and also we included the aggregation and filtering step to focus only on data caring the most significant information.

First step in our processing pipeline is to import data into Apache Spark and transform it to DataFrame structure [9]. We decided to use Sparks' DataFrame structure to be able to exploit powerful SQL semantics while keeping the performance at high levels. The next step in our processing pipeline is to perform additional space/time data aggregation to obtain daily based snapshots of connectivity network across the observed area. Connectivity network is in the form of graph and methods for statistical filtering could be applied to remove the noise from the data. We used Disparity filter [105] as a method for extraction of the relevant connections in weighted networks, for distinguishing between strong and weak links in the graph. Disparity filter is a statistical method used to discard statistically less relevant links, while taking care of the nodes local neighbours. More details about the Disparity Filter are presented in Section 2.3.2.

We have used Equation (2.4) which calculates the *link significance*, α_{ij}, where α denotes the significance *threshold*, p_{ij} is the probability of having a link between nodes i and j, and n is the number of nodes in the network [83].

We note that smaller values of α_{ij} denote more significant edges. Therefore, filtering is applied by keeping all edges where $\alpha_{ij} \leq \alpha$ and thus removing all edges where $\alpha_{ij} > \alpha$. In our experiments, we applied filtering with probability 95% as it represent the less strict filtering, allowing us to keep more edges. The α value used is 0.05.

To discover community structure in the network, we perform community detection using modularity based algorithms, such as the one proposed in [21]. The concept of modularity [80] presented by Equation (2.7), is used as a goodness measure for the quality of partitions, where A_{ij} is the weight of the edge connecting the i-th and the j-th node of the graph, $\sum_j A_{ij}$ is the sum of the weights of the edges attached to the i-th node, c_i is the community where the i-th node is assigned to, $m = (1/2)\sum_{i,j} A_{ij}$, and $\delta(i, j)$ is zero if nodes x and y are assigned to the same community and 1 otherwise. Community detection is based on graph clustering, methods for graph clustering are in detail described in Section 2.4.1.

Unfortunately, computing communities based on the maximization of the modularity, is an $\mathcal{NP}$-hard problem. To provide an efficient solution, the algorithm proposed in [21] uses an iterative process that involves shrinking the graph, every time modularity converges. In each phase, each node is assigned to a neighboring community that maximizes the modularity of the graph, while nodes are moving around communities and modularity grows, the algorithm keeps on executing this process. The algorithm stops when there are no more changes, and a shrinking process is applied. Upon shrinking the graph, each community produced during the previous phase, it is assigned to the same *super node* of the new graph. The same process is applied again to the new graph and keeps on that way until it reaches the stopping point. The stopping point is when the modularity detected in the new graph is less than the modularity detected in the previous one. The algorithm returns the set of communities that maximize the modularity. The outline of the technique is depicted in Algorithm 1. More detail and discussion about modularity based technique for community detection is presented in Section 2.4.1.

To explore spatial semantic of areas that are highly connected between each other we decided to evaluate the community structure of telecom connectivity network. For community detection we applied the LOUVAIN [21] modularity-based algorithm implemented in the Scala programming language, and run it in the Apache Spark engine. This algorithm is very efficient for large networks and also it does not require the number of communities as an input. The number of communities is determined by the algorithm during the runtime. The Louvain algorithm is applied for every single day, and generated communities are further evaluate togheter with graph based properties and land use profiles of the area they cover. Detailed description of the Louvain algorithm is presented in Section 1.

6.3.1 Spatial and graph-based properties

Here we introduce the term *"spatial community property"* to highlight the spatial nature of communities evaluated from telecom connectivity network. We used the results of community detection to inspect spatial community properties that reflect human dynamic the most. Other properties used to evaluate human dynamics that we have explored are graph based properties such as: betweenness centrality, weighted degree, PageRank and core number.

Together with telecom data used to evaluate the community structure from connectivity network we used spatial open data that describe the spatial semantics of locations. One of the open spatial data sources is Copernicus Land Monitoring Service from which we have used Urban Atlas 2012 data set [68]. Urban Atlas provides detailed land use and land cover data for 800 Functional Urban Areas across Europe for the year 2012. For spatial area of Milan city with surrounding suburbs there are 21 different land use class present in the data, as shown in Figure 6.1.

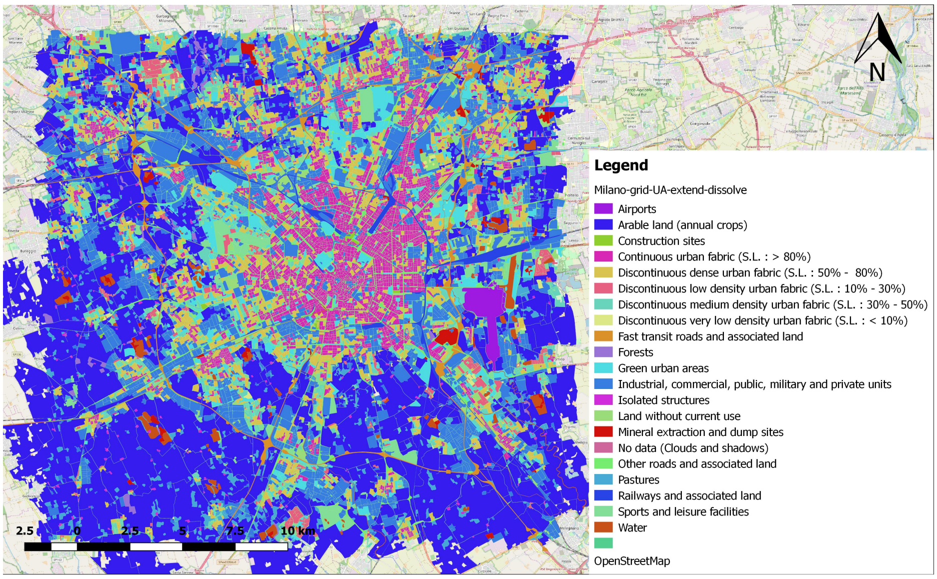

N
Legend
Milano-grid-UA-extend-dissolve
Airports
Arable land (annual crops)
Construction sites
Continuous urban fabric (S.L. : > 80%)
Discontinuous dense urban fabric (S.L. : 50% - 80%)
Discontinuous low density urban fabric (S.L. : 10% - 30%)
Discontinuous medium density urban fabric (S.L. : 30% - 50%)
Discontinuous very low density urban fabric (S.L. : < 10%)
Fast transit roads and associated land
Forests
Green urban areas
Industrial, commercial, public, military and private units
Isolated structures
Land without current use
Mineral extraction and dump sites
No data (Clouds and shadows)
Other roads and associated land
Pastures
Railways and associated land
Sports and leisure facilities
Water
OpenStreetMap
2.5
0
2.5
5
7.5
10 km

Fig. 6.1 Land use classes over the extended spatial area of Milan

Telecom data that we explored in this research is georeferenced as regular grid with 235 x 235 meters cells. Those cells or squares of grid are simulating real RBS, but those are very small in size to evaluate diverse land use profiles within each cell. To explore the relationship between communities evaluated from telecom connectivity network and land use profiles, we generalized real RBS coverage network using Voronoi polygons, Figure 6.2. From the Figure 6.2 we observe that Voronoi polygons represent very good generalization of original RBS coverage. Further, to associate telecom traffic with underlying Voronoi network we performed spatial intersection between Voronoi network polygons and telecom grid coverage. All square ids that fall within each Voronoi polygon are assigned to that polygon together with their associated traffic. For those squares that are between two neighbouring polygons, we calculated the area that falls within each polygon and the square is assigned to that polygon where the intersection area is larger. That way we obtained the connectivity network between overlayed Voronoi network of polygons. Voronoi polygons are large enough in size to evaluate land use profiles for each polygon.

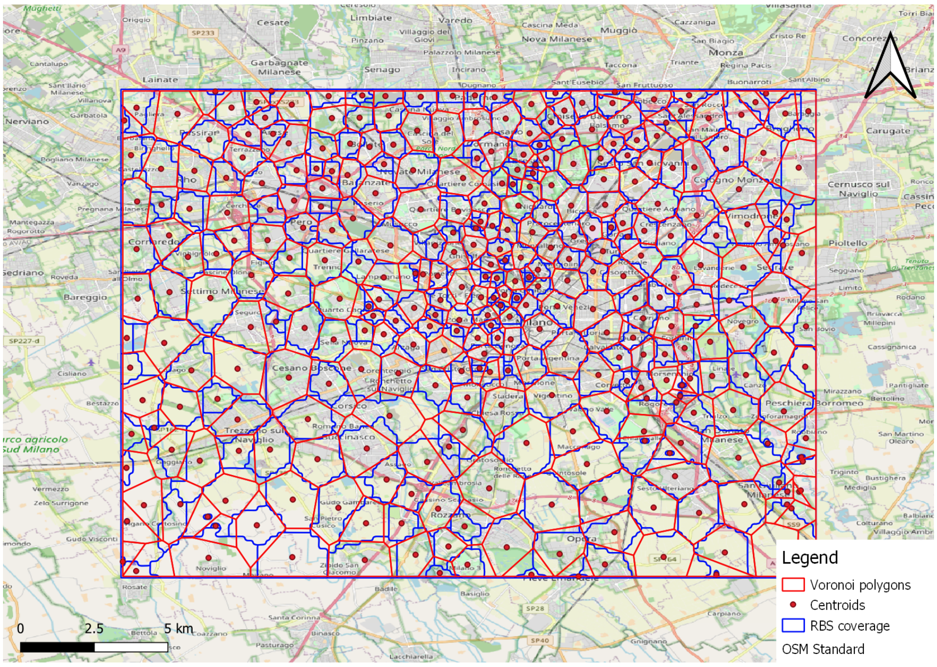
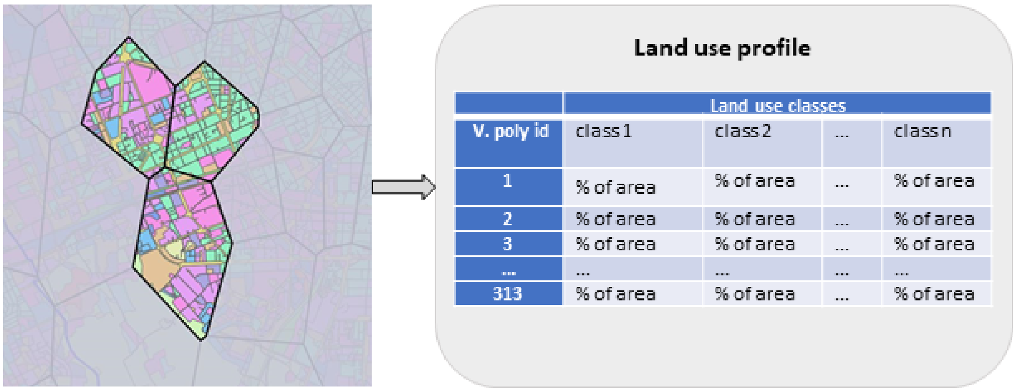


Fig. 6.2 RBS coverage polygons and evaluated Voronoi polygons

To generate a profile vector with land use types for each Voronoi polygon in the network, we first performed spatial intersection between RBS coverage area and Urban Atlas 2012 data layer [68]. The high resolution land use/land cover (LULC) Urban Atlas dataset was produced using Earth Observation data, road network datasets and topographic maps. For each LULC polygon feature resident population data was derived from various available sources, including GEOSTAT grid 2011, based on the 2011 census. The minimum mapping unit for all artificial surfaces is 0.25 hectares and 1 hectare for all the remaining LULC classes [44].

We extracted land use classes present in each Voronoi polygon and calculated the percentage of area covered by the specific class. We further used those vector profiles as polygon based land use features. An example of intersection between three neighbouring Voronoi polygons and land use layers, and extraction of land use profile vectors is illustrated in Figure 6.3.



Land use profile

	Land use classes			
V. poly id	class 1	class 2	...	class n
1	% of area	% of area	...	% of area
2	% of area	% of area	...	% of area
3	% of area	% of area	...	% of area
...	...	...	...	...
313	% of area	% of area	...	% of area

Fig. 6.3 Example of three Voronoi polygon intersection with land use layer and Land use profile extraction.

We performed community detection over graphs where nodes are essentially Voronoi polygon ids, and links are aggregated traffic between those nodes. Therefore, the communities consist of neighbouring Voronoi polygons.

Along with community detection we have performed a deeper analysis of connectivity graphs and quantified local graph properties, which are further used as indicators of human dynamics. Local graph properties uncover localized patterns in the graphs, focusing on adjacent node neighbourhood [82]. We have evaluated different centrality measures, such as betweenness centrality, node degree, PageRank and core number, which are in detail described in Section 2.3.1.

For weighted networks k-core decomposition is modified to some extend because it needs to consider both edge weight and node degree [52]. The core number of the node in our case represents the weighted degree of node of its maximal core. The average of all values for associated node is used as a property. Among other centrality measures we calculated the PageRank of each node and used it as a property. The degree of a node is another important property that represents the number of incoming and/or outgoing edges connected to that node. The weighted degree of a node is calculated considering the edge weights connected to that node. All calculated properties are used as graph based properties associated to Voronoi polygons in the RBS coverage network.

We performed community detection over daily based connectivity graphs. Community formation is a highly dynamic process that depends on the connectivity network structure, and their spatial location, center, geographical spreading fluctuate between days. To quantify the spatial dynamics of community structure we introduced measures that reflect some aspects of the dynamics. Communities over inspected area are formed of single, very often neighbouring Voronoi polygons, and they are overlapping between days to some extent. For each polygon we detected all communities to which that polygon belongs to and we calculated the geographical intersection between sequential communities. Measures that we calculated to evaluate the spatial dynamic of communities are:

- mean intersection area,
- standard deviation of intersection area,
- average coverage area and
- diameter.

Mean intersection area is the property that reflects the extent to which sequential communities are spatially overlapping. Another property related to spatial intersection is standard deviation of intersection area. This property is introduced to evaluate the dispersion of the sequential intersection area values. Daily based communities vary in size and spatial distribution. We have also introduced the coverage area as a measure of overall geographical reach of communities. We calculated the average of all coverage areas, and used it as a property. For each Voronoi polygon, we have computed the distances between the polygon center and the center of communities. The center of community is defined as the geometrical center of the polygon that is created by joining all Voronoi polygons that are part of the community. The maximum of all distances, the diameter, is used as a measure of community center displacement from its starting Voronoi polygon center.

6.3.2 Regression models

Machine learning is a discipline in computer science and artificial intelligence that focuses on algorithms and methodologies that simulate natural learning process in humans and therefore extract meaningful knowledge from very large and complex data sets. Machine learning became very attractive field of science in the past decade due to exponential jump in collected and available data about different human interactions and systems. Such very large and complex data could not be processed using some traditional techniques such as SQL and

Excel analytics. Demand for deeper insight into the data lead to accelerated development of both technical systems and platforms as well as algorithms for data processing.

Regression is a common technique in Machine learning that consist of many mathematical methods used to predict a continuous outcome y based on one or more predictor variables x. Regression analysis are used to unveil the relationship between the variables by estimating how correlated they are and how one variable affects the other. Two regression models are most widely used, Linear Regression [125] and Logistic Regression [129]. Linear Regression is a basic regression model that uses the linear function y = ax + b to estimate the correlation between variables x and y. Aim of the Linear Regression is to fit the strait line in the data point space that best fits the points distribution. By evaluating a and b coefficients, Linear Regression quantifies the relationship between variables x and y, and therefore can be used to predict future values. Different from Linear Regression which predicts continuous values, Logistic Regression predicts the discrete values such as 0 or 1, "yes", or "no", "in" or "out", etc. and therefore is the most suited for binary classification problem. The output of the Linear Regression is the exact value, whereas the output of the Logistic Regression is the probability of the default class, meaning its output is always between 0 and 1. Both techniques fall under the class of supervised Machine learning algorithms, meaning they learn from the labeled data. Despite their relatively simple idea, both techniques are widely used for many different application problems.

Besides Linear and Logistic Regression, there are other techniques used to provide the similar output. One of the very often used is technique based on the ensemble of decision trees, Random Forest Regression [104]. Random Forest is very powerful ensemble learning technique for classification, regression and other tasks that work by constructing a multitude of decision trees during training time. When classification is performed, the output of the Random Forest algorithm is the class selected by majority of the trees. When Random Forest is used to perform regression, the output is the average of the predictions of the individual trees. Ensemble learning methods are algorithms that make very accurate prediction, more than a single model. Specially, Random Forest Regression model is very powerful and accurate, it is known by great performance on many problems, including the problems with non-linear relationships. There are off course few disadvantages of the model, including the lack of interoperability , easily overfitting, the number of trees in the ensemble must be selected prior to the algorithm run time.

To evaluate predictive power of land use profiles to produce predictions on graph based and spatial community properties based on Voronoi polygons we used Random Forest [25] Regression model and Ridge Regression [59] model. From Random Forest model it is possible to derive feature importance [74] measure, which is used as an indicator of more valuable features for prediction. Feature importance values sum to 1 and the higher values indicate higher importance. Alternative approach that we tested is using the Ridge Regression model, because of its property to overcome the problem of multicollinearity amongst regression predictor variables.

6.4 System architecture and analytical pipeline

Diverse data sources that we used in this research demanded high flexibility regarding analytical and technical approaches. Two main data sources that we used are telecom connectivity data from CDR files and land use data from Copernicus Urban Atlas. Due to the size and complexity of telecom connectivity data, we needed to use some advanced Big Data analytics tools to process the data. We decided to use Apache Spark platform because of its rich high level API in Scala which provides tools to analyse both structured and graph data.

First we imported raw CDR data into Spark's DataFrame structure, which is commonly used for "table" like data sources. The "table" analogy is not coincidence, because the DataFrame structure supports SQL operations, as well as functional. The DataFrame is loaded during the runtime, meaning it's kept in the operational memory during the process but it's not saved in the data storage unless it is explicitly specified. The DataFrame in Spark is also distributed data structure like RDD, but it is more convenient to use in the case of structured data. The fact that is kept in operational memory during the runtime enables better performance. Apache Spark supports also the "cashing" operation over the DataFrame, which will give even better performance in the case when we need to access the same DataFrame many times during the runtime. After the raw CDR data is imported into Apache Spark DataFrame, three processing steps are applied:

- additional space/time aggregation is performed to obtain daily based graphs of connectivity,
- graph filtering is applied to keep only statistically most significant edges in the graph while making the graph "lighter" for further processing, and finally
- community detection is applied over the filtered graphs.

Telecom connectivity data makes very dense graph structure, where the number of edges is few times greater than the number of nodes (almost fully connected structure) and the edges are weighted, which makes it even more challenging from the algorithms perspective. Therefore, the filtering is recommended as a measure to make graph structure less dense, while keeping the core backbone of the network. In our previous work, particularly described and presented in Chapter 5 we compared the output of the community detection over filtered and unfiltered graph, to justify the need for filtering. Our results showed that there is not much change in the formed communities from filtered and unfiltered graph, while the runtime performance was significantly enhanced when using filtered graph. To further explore spatial characterises of inspected communities we extracted the results of community detection into csv files which can be further analysed using QGIS software.

Next branch of our analytical pipeline goes from spatial land use data from Copernicus Urban Atlas. We used QGIS software to perform spatial intersection between land use polygons and RBS coverage network. Further, we extracted land use profiles for each RBS coverage polygon which is represented as Voronoi polygon of the coverage network. Next step in our analytical pipeline is performed using programming language Python, which has rich portfolio of libraries for data analytics and machine learning. From filtered graphs we computed graph based properties and from community csv files we computed spatial community properties. To further investigate the relationship between land use and graph based and spatial community properties we applied Random Forest Regression. The whole analytical pipeline is illustrated in Figure 6.4

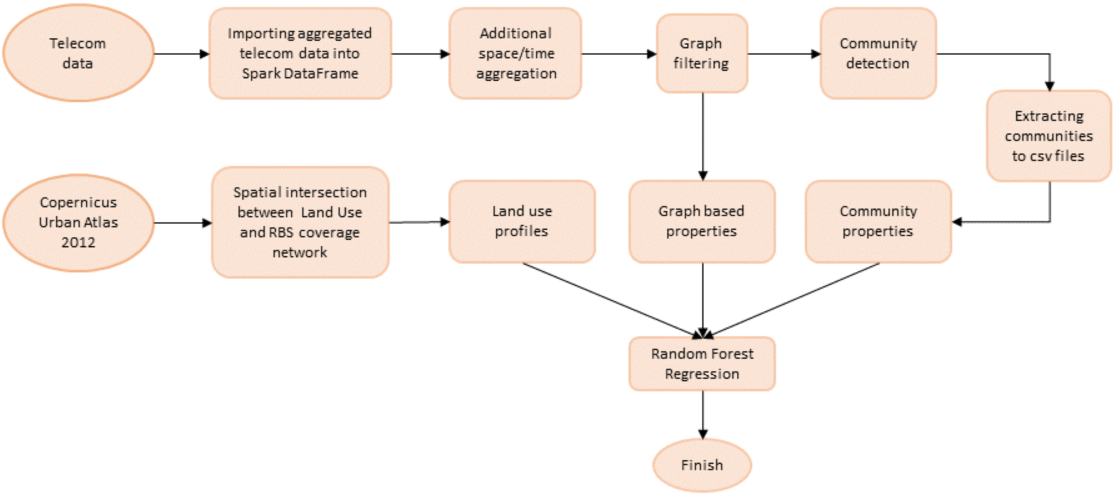

Telecom data
Importing aggregated telecom data into Spark DataFrame
Additional space/time aggregation
Graph filtering
Community detection
Extracting communities to csv files
Copernicus Urban Atlas 2012
Spatial intersection between Land Use and RBS coverage network
Land use profiles
Graph based properties
Community properties
Random Forest Regression
Finish

Fig. 6.4 The pipeline of the proposed methodology.

First steps related to CDR data processing, from importing raw data into Spark's DataFrame to community detection are computationally very challenging since CDR files are very large in size and connectivity graphs created from it are very "heavy" for processing due to high number of weighted edges. Therefore, the most demanding tasks are done in Hadoop Ecosystem using Spark's engine for processing and Hive support for database storage. After raw data is imported, additional space/time aggregation is performed and the results that represent daily based connectivity graphs are stored in HDFS in Apache Hive database storage. We have chosen Apache Hive as a database storage because of it's compatibility with Spark and HDFS, and support for high level SQL API (HiveQL, which is actually common SQL syntax, modified to some extend). Files are stored in distributed manner using binary .parquet format, which saves a lot of storage and makes ETL operations much easier comparing to ETL with common textual files. We further transform the data by applying graph filtering using SQL semantic.

As a final step in CDR data processing we perform community detection. Community detection algorithm is modified to read from Hive database and to iteratively update the clustering result until the algorithm terminates. To visualize the communities in geographical space we extracted the results from Hive databse to common csv files and load it in QGIS software. We have also used those csv files to evaluate spatial community properties using Python. We have used Python and it's rich data analytics libraries to put together land use profiles, graph based properties and spatial community properties, and to evaluate the correlation and predictiveness of graph and spatial properties based on land use using Random Forest and Ridge Regression.

The final result of the processing is detecting the dependence between land use and human dynamics reflected through graph based and spatial properties and evaluating the importance of specific land use class in the predictive model. Our system architecture consist of three main ground stones:

- Hadoop Ecosystem with Apache Spark engine for processing and Hive as a database storage,
- QGIS software used for visualization and spatial data operations, and
- Python used to perform feature extraction and run predictive model.

The system architecture with technology specifics is illustrated in Figure 6.5.

6.5 Results

After careful examination of the problem of community detection in large, very dense graphs and exploring correlation between telecom connectivity and land use, we run experiments using proposed system architecture to evaluate our hypothesis in practice.

To better understand the experimental setup we should recall on some basic characteristics of telecommunication networks. The Telecommunication network consists of spatially distributed Radio Base Stations (RBS) that carry equipment used to provide the signal and

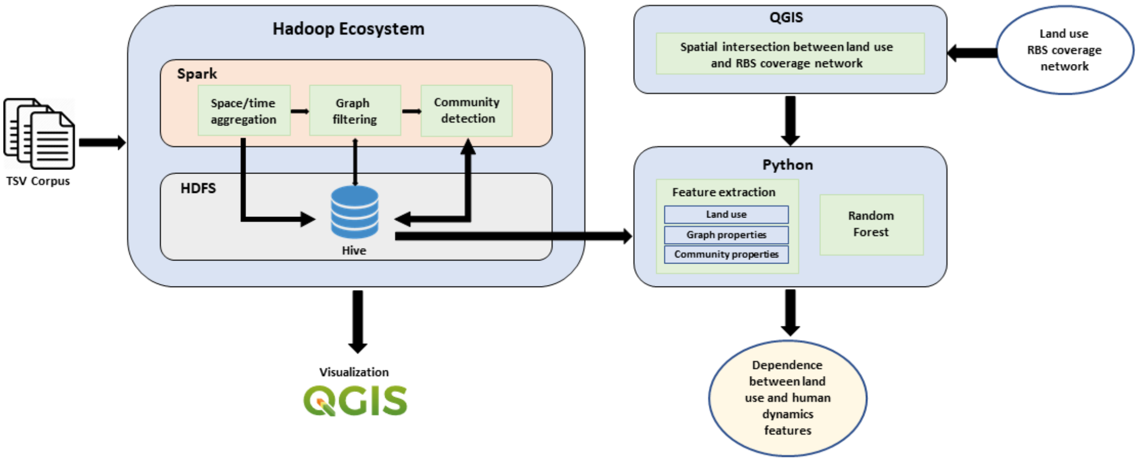


Fig. 6.5 System architecture.

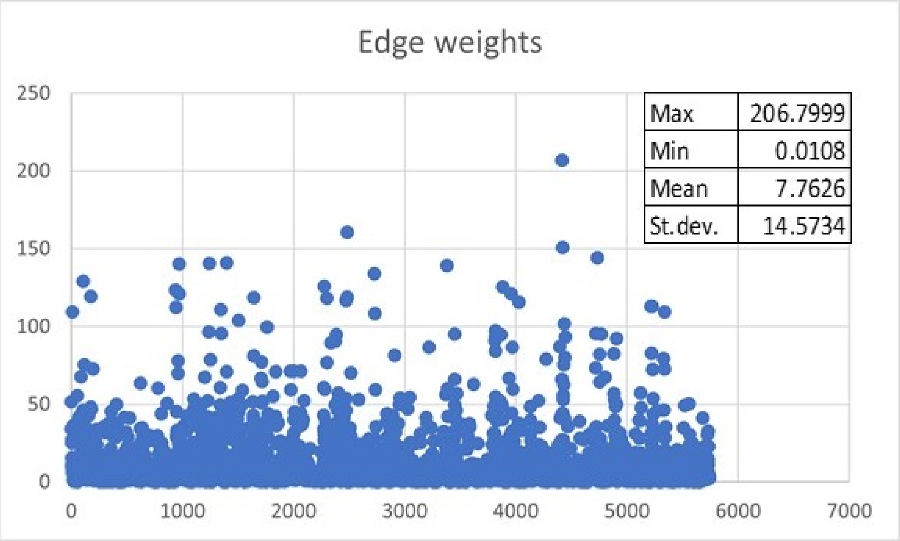
to transfer telecom traffic. Each RBS is characterized with signal coverage, which can vary from RBS to RBS depending on equipment. In urban areas signal coverage is ubiquitous, while rural and less populated areas can be weak in signal coverage leaving some areas out of reach, completely without signal. In telecommunications, signal coverage is often determined with Voronoi polygons, where the antenna is positioned in the polygon center.

In our experiments, we are working with connectivity network where the telecom traffic is aggregated, but still represent the traffic distributed from one RBS to another. The connectivity network is mathematically modeled as graph, where RBSs are aligned with graph nodes and weighted links represent the edges in the graph. Graph created this way is almost fully connected, weighted and directed which makes it very challenging for further analysis, due to high complexity of the algorithmic problem. To decrease the complexity of the problem, we introduced some generalizations in our model. First, we discarded the directions from the graph since it does not carry any significant information for our analysis. We made undirected graph from directed, by summing the weights over the same edges with opposite direction. Second, we inspected the weights distribution in the graphs and concluded that it does not follow any specific distribution such as Normal distribution [1] or Uniform distribution [2], but instead the weights follow some irregular distribution where mean and standard deviation are significantly smaller than maximum value. We can observe the example of statistics for weight distribution in connectivity graph for single day. From Figure 6.6 we observe how most of the weights are concentrated close to the lower values, while higher weight values out stand greatly from the distribution. Edges that carry such high

[1] Normal distribution

[2] Uniform distribution

value weights that out stand greatly from distribution mean could be considered as outliers, but in our case those are edges that carry significant network information and therefore cannot be excluded from the analysis. Similar behaviour of the weight distribution is observed in all other daily based graphs.



Max	206.7999
Min	0.0108
Mean	7.7626
St.dev.	14.5734

Fig. 6.6 Scatter plot of weights of edges in connectivity graph for single day.

Such distribution of weights makes it very difficult to distinguish between edges that carry significant information from those that can be discarded. To overcome this issue, we applied statistical method for graph filtering called the *Disparity filter* 2.3.2, that is designed to consider local neighborhood of the nodes when calculating edges significance therefore keeping the core backbone of the graph undamaged after filtering [105]. The Disparity filter asks for one parameter α to determine the level of filtering that would be applied. The parameter α takes values between zero and 1 and it determines the probability of the edge to be kept in the graph. Higher α values indicate less strict filtering while very low α values would filter out the most of the edges. Finding "just right" α value for particular problem needs a little fine tuning by try and error.

To select the appropriate α value, we compared the results of community detection for the same graph when different filtering levels are applied. The results are presented in Table 6.1. From this table we observe a significant drop in the number of edges when filtering is applied, which is expected when using the Disparity filter. The number of detected communities remains stable when filtering with $\alpha = 0.05$ is applied, while with smaller α values a higher number of communities is observed. To evaluate how similar are the results when different filtering levels are applied, we computed the *Adjusted Random Index* (ARI) which is a common measure for evaluating similarity between clusters [123]. Higher values of ARI indicate clusters with more similar structure. We considered the results of community detection performed over the unfiltered graph as the "ground truth". The ARI between the clustering of the ground truth graph and the graph filtered with $\alpha = 0.05$ is 0.83 which indicates very high similarity. When filtering is applied with $\alpha = 0.01$ threshold, ARI drops slightly to 0.81 which also indicates a high similarity between clustering, while the number of detected communities increases 20% which is significant compared to the increase in communities when filtering with $\alpha = 0.05$ is applied. With more strict filtering, when the α threshold is set to 0.001, ARI drops to 0.71 and the number of detected communities increases for 38% compared to the ground truth. The aim of graph filtering before running community detection is to eliminate a large number of weak edges, in order to make the graph structure less dense and therefore to increase performance, while preserving the network structure. Filtering with $\alpha = 0.05$ threshold showed the best results, eliminating a significant number of edges, while the high value of ARI and the change in number of detected communities of only 4% indicate a very high similarity between clustering results. The results that are part of this thesis are published in the paper [83].

Table 6.1 Number of edges, communities and ARI when different filtering level is applied.

α	**Edges**	**Communities**	**ARI**
1	95 860	53	-
0.05	6 322	55	0.83
0.01	3 809	64	0.81
0.001	2 323	73	0.71

After performing generalizations to discard the direction and filter edges in our graph model, we applied community detection using Louvain algorithm [20] to obtain community structure over inspected area.

We performed community detection over graphs filtered with $\alpha = 0.05$ that kept 95% of the edges. Louvain algorithm starts with assigning each node to its own community and proceeds with moving nodes across communities to maximize the modularity function. Although the initial assignment of nodes might depend on data partitioning and sorting, Louvain algorithm showed very stable results between different iterations for the same graph. We run Louvain algorithm for community detection 10 times for each day and calculated ARI between different iterations. ARI between different iterations for the same input graph is always 1.0 indicating that the algorithm is very stable. The change in community structure between days is caused by change in input graph structure which reflect the dynamics of human connectivity. To evaluate the change in community structure between sequential days we calculated average ARI, which is 0.73. Such value for ARI indicate that cluster structure is changing between days very dynamically.

The path from RBS coverage area network to communities detected based on connectivity network graphs is presented in Figure 6.7. On the first map, far left in Figure 6.7, the network coverage consisting of Voronoi polygons over the inspected area of Milan city with the surroundings is presented. On the middle map in Figure 6.7, the overlaid connectivity graph which is the input for community detection is presented. On the last map from Figure 6.7, the result of community detection for the input graph is presented.

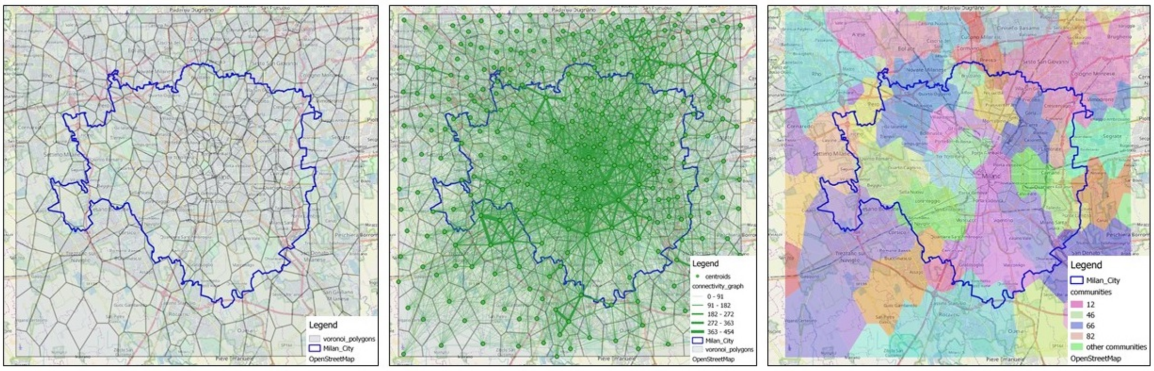


Fig. 6.7 From coverage area to telecom data graphs and communities.

The next step in our processing pipeline is to deeper investigate the space-time evolution of community structure. Since we used multiple different software platforms to run the analysis and examine the results, in this step we needed to to export the data between the platforms. First, we extracted communities from Hive database layer to common csv files and we visually examined the spatial distribution of communities using QGIS software. By visual inspection we noticed that although the form and number of communities differ between days some patterns are repeating. The granularity, size and spatial distribution of communities formed in densely populated, built up areas differ much compared to those of communities formed in peripheral parts of urban zone.

High dynamics of community structure formed over urban core of the city can be explained by high dynamics of connectivity network over same area, reflected in input graph. While exploring the results it is observed that core of the connectivity network, where the strongest links occur, is located over urban core of the city. Communities formed in the urban core of the city tend to be small in size and dynamically distributed in space between days, while communities formed in less urbanized peripheral zones tend to form large, more stable structures. To quantify the observed phenomena, we introduce new features called *spatial community properties* described in Section 6.3.1.

In the Chapter 4 we are discussing how the social event presence could be detected through analysing local and global graph properties of telecom connectivity networks. Those findings motivated us to compute the graph properties described in Section 6.3.1, to evaluate their correlation with land use and communities. To calculate community and graph properties, we used Python and its rich portfolio of libraries such as Pandas, NetworkX, NumPy, SciPy, Matplotlib, GeoPy, GeoPandas, etc. To evaluate the predictiveness of graph and community properties which contain latent footprint of human dynamics, based on land use we applied regression. Regression is applied sequentially over graph and community properties using Random Forest and Ridge regression algorithms.

The results of regression are presented for the graph properties in Table 6.2 and for spatial community properties in Table 6.3. For all properties, Random Forest regression showed better results than Ridge regression. From Table 6.2, we observe that the predicted values of the property avgCN, i.e., *average core number* show the best correlation with real values, followed by property avgWND, i.e., *average weighted node degree* which is expected since both measures are considering weights over nodes edges which is important structural property of the network. This result is significant since it emphasize that the variability in core number and weighted node degree can be explained by the variability in the land use. Another property that shows strong correlation between predicted and real values considering Spearman coefficient is avgPR, i.e., *average PageRank*, while the other metrics, Pearson and R2, indicate weaker correlation. The property that is the least associated with land use is avgBC, i.e., *average betweenness centrality*.

From Table 6.3 we observe that the predicted values of the spatial community property avgCA, i.e., *average coverage area* show the best correlation with real values. Predicted values of spatial community property avgIA, i.e., *average intersection area* show fair correlation with real values considering Spearman coefficient, but show less correlation considering Pearson and R2 metric. Community properties that seem the least associated to land use are st.dev IA, i.e., *standard deviation intersection area* and avgDiam, i.e., *average diameter*.

Overall Spearman correlation test shows the best results for all properties except for avgCN, avgDiam and st.dev IA Ridge regression where the Pearson correlation was higher. Spearman correlation test is expected to show robust results in the most of the cases since it does not carry any assumptions about the distribution of the data. All correlation results are significant at 0.01 (all p-values < 0.006).

Table 6.2 Prediction results for structural properties.

Metric	avg BC		avg WND		avg PR		avg CN	
	RF	Ridge	RF	Ridge	RF	Ridge	RF	Ridge
R2	0.111	−0.0192	0.422	0.424	0.310	0.321	0.628	0.623
Pearson	0.357	0.163	0.659	0.653	0.581	0.574	0.793	0.790
Spearman	0.269	0.289	0.776	0.707	0.715	0.644	0.778	0.753

Table 6.3 Prediction results for spatial community properties.

Metric	avg IA		st.dev IA		avg CA		avg Diam	
	RF	Ridge	RF	Ridge	RF	Ridge	RF	Ridge
R2	0.353	0.196	0.236	0.109	0.521	0.410	0.211	0.140
Pearson	0.596	0.485	0.488	0.385	0.722	0.660	0.474	0.403
Spearman	0.630	0.579	0.490	0.382	0.727	0.692	0.414	0.335

In the next step, we wanted to measure the impact of land use classes to prediction of other properties. To explore how each specific land use class is important for prediction, we calculated feature importances. The importance of land use classes for predicting each graph based and community property is presented in heatmaps in Figures 6.8 and 6.9.

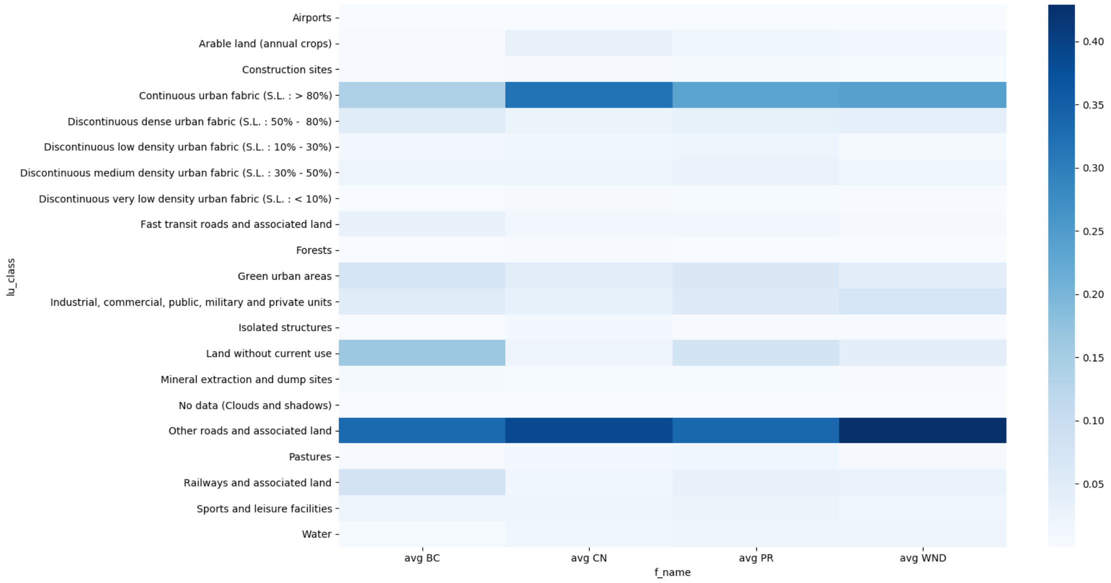


Fig. 6.8 Importance of land use classes for predicting graph based properties.

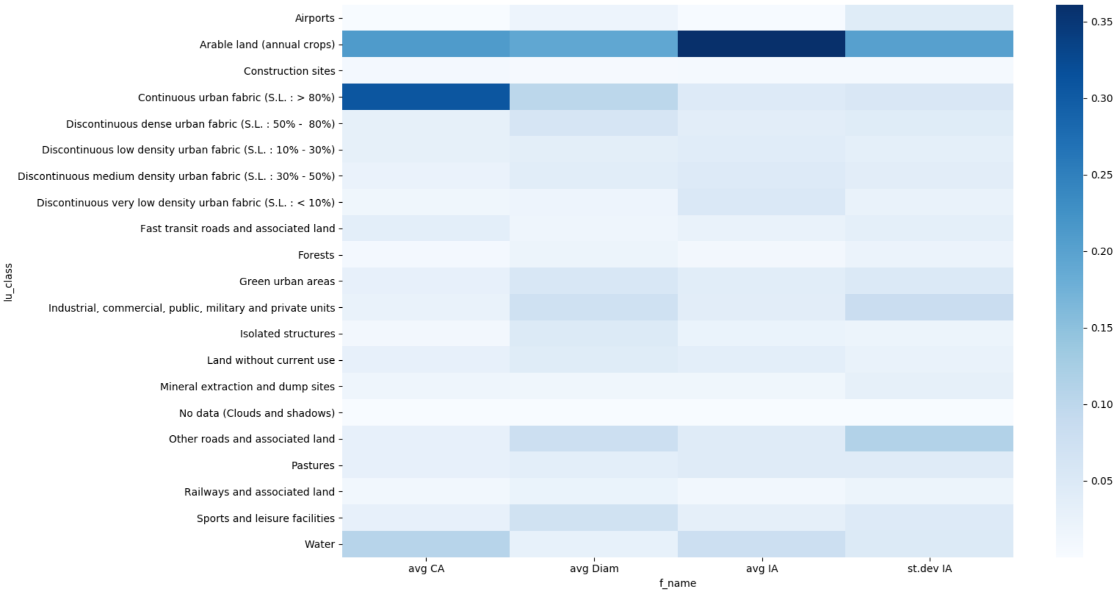


Fig. 6.9 Importance of land use classes for predicting community properties.

For predicting the graph based properties based on land use, the most important classes are *Other roads and associated land, Continuous urban fabric (S.L.: > 80%)*, as shown in Figure 6.8. The highest predictive power has land use class *Other roads and associate land* for the property *average weighted node degree* (avgWND). For predicting the property *average core number* (avgCN) the most important land use classes are *Other roads and associated* *land* and *Continuous urban fabric (S.L.: > 80%)*, which also have the highest impact on predicting the property *average PageRank* (avgPR). Land use has the least predictive power for the property *average betweenness centrality*, which is shown in Table 6.2 with small values of correlation metrices. From the heatmap in Figure 6.8 we observe that only land use class *Other roads and associated land* has impact on predicting the property *average betweenness centrality*, but due to small values of correlation metrices we can conclude that impact is not significant.

For predicting the community properties based on land use, the most important classes are *Arable land, Continuous urban fabric (S.L.: > 80%)*, as shown in Figure 6.9. Land use class *Arable land* has the most impact on predicting the community property *average intersection area* (avg IA), which can be observed from the heatmap in Figure 6.9. The most predictive community property based on land use is *average coverage area* (avg CA), where the most impact have land use classes *Continuous urban fabric (S.L.: > 80%)* and *Arable land*. From the heatmap in Figure 6.9 we observe that only land use class *Arable land* has the impact on predicting the community properties *average diameter* (avg Diam) and *standard deviation intersection area* (st.dev IA), but from the Table 6.3 we can observe small values of correlation metrices which indicate that the impact is not significant.

Next we wanted to deeper explore the urban profiles to detect polygons from network coverage area where the classes of high importance for predicting properties are dominant in area.

In Figure 6.10 is presented network coverage polygon where the dominant land use class is *Other roads and associated land* together with its graph and community spatial properties and its geographical position in the network.

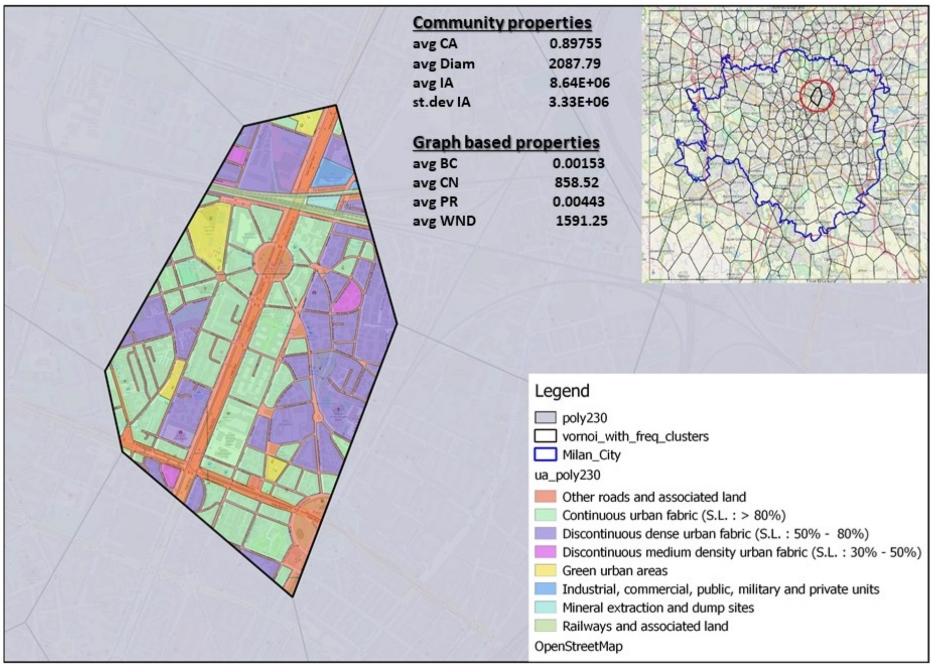

Community properties
avg CA 0.89755
avg Diam 2087.79
avg IA 8.64E+06
st.dev IA 3.33E+06
Graph based properties
avg BC 0.00153
avg CN 858.52
avg PR 0.00443
avg WND 1591.25
Legend
poly230
vornoi_with_freq_clusters
Milan_City
ua_poly230
Other roads and associated land
Continuous urban fabric (S.L. : > 80%)
Discontinuous dense urban fabric (S.L. : 50% - 80%)
Discontinuous medium density urban fabric (S.L. : 30% - 50%)
Green urban areas
Industrial, commercial, public, military and private units
Mineral extraction and dump sites
Railways and associated land
OpenStreetMap

Fig. 6.10 Land use profile of polygon where the dominant class is *Other roads and associated land*.

In Figure 6.11 is presented network coverage polygon where the dominant land use class is *Continuous urban fabric (S.L.: > 80%)* together with its graph and community spatial properties and its geographical position in the network.

In Figure 6.12 is presented network coverage polygon where the dominant land use class is *Arable land* together with its graph and community spatial properties and its geographical position in the network.

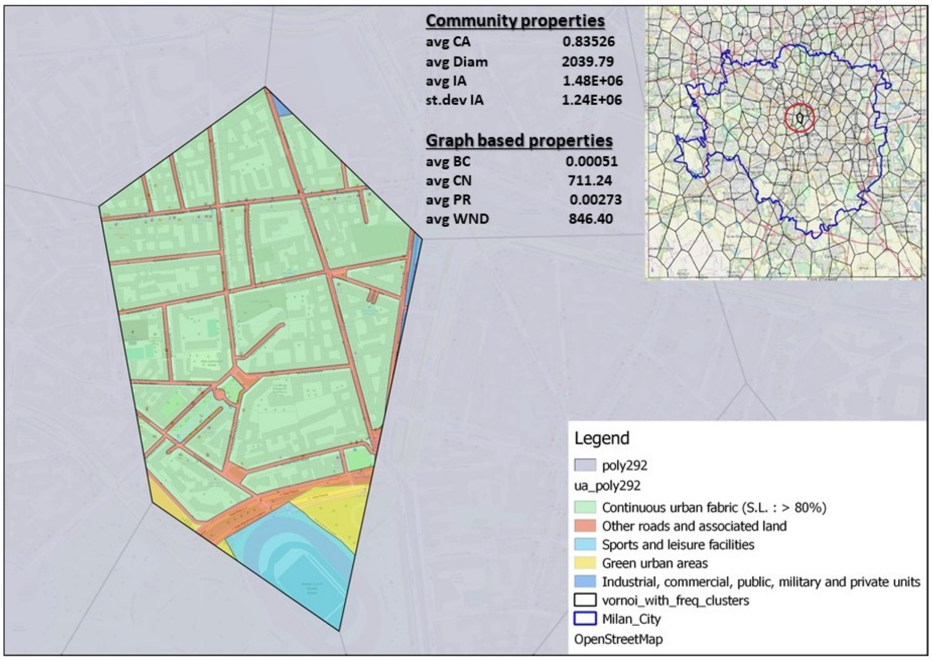


Fig. 6.11 Land use profile of polygon where the dominant class is *Continuous urban fabric (S.L.: > 80%)*.

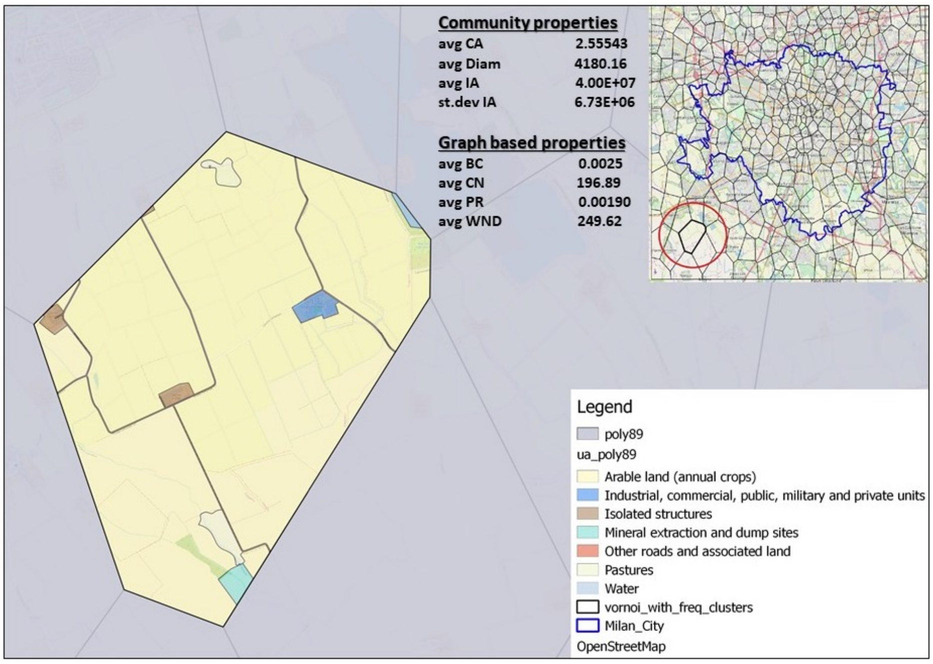


Fig. 6.12 Land use profile of polygon where the dominant class is *Arable land*.

By observing those three special cases presented in Figures 6.10–6.12 with specific dominant land use class we notice some trends related to spatial community and graph properties. Polygon in the highly urban area where the dominant class is *Continuous urban* *fabric (S.L.: > 80%)* is characterized with lower *mean intersection area* than polygons where the most dominant classes are *Other roads and associated land* and *Arable land*. This property indicates that the observed polygon forms highly dynamic communities, which are distributed in space in diverse manner and therefore the spatial intersection between communities is small. On contrary, high value of *mean intersection area* indicates the presence of more stable community structure that is not prone to dynamical change. The property *st.dev. intersection area* shows similar behaviour and indicates the same characteristics. Community properties *average coverage area* and *diameter* indicates the maximal spatial reach of all communities formed including observed polygon. High value of *average coverage area* indicates that formed communities are large in size, covering significant spatial area. This behaviour can occur due to initial size and/or granularity of Voronoi polygons contained in the communities, where we can have fewer very large polygons in the community or many smaller ones. Larger *average coverage area* of the community indicates wider reach of information spread inside community. High value of *diameter* indicate larger and wider reach of the communities, but only as a linear measure and therefore shows not significant correlation with land use which is area feature.

Graph properties of polygons show diverse behaviour depending on the dominant land use class. The property *average betweenness centrality* gains the highest value for polygons with dominant land use class *Arable land*. This can be explained with the geographical position of the polygon in the network and network dynamics. Polygons that are located far from urban core of the city may act as bridges between distinct parts of the network and therefore gain higher value of *betweenness centrality*. Also, if the network around specific polygon is changing fast day by day, it is more likely that the observed polygon will have less impact to network connectivity between days. The property *average PageRank* has the highest value for the polygon where the dominant land use class is *Other roads and associated land*. It is interesting to notice how physical transitivity of a node, i.e., polygon in the network characterized by land use class *Other roads and associated land* is reflected to its communication transitivity defined by *average PageRank* property.

The value of *average PageRank* decreases for the polygons where transit infrastructure defined with land use class *Other roads and associated land* is less present, as shown in Figures 6.11 and 6.12. The property that is the most predictive based on land use is *average core number*, and it has the highest value for the polygon where the dominant land use classes are *Other roads and associated land* and *Continuous urban fabric (S.L. : > 80%)*, as shown in Figure 6.10. Polygons located in the urban core of the city where the dominant land use classes are *Continuous urban fabric (S.L. : > 80%)* and *Other roads and associated land* also show high value for the property *average core number*. In contrast, polygons where the most dominant land use class is *Arable land* and classes *Continuous urban fabric (S.L. : > 80%)* and *Other roads and associated land* are not significantly present show very small values of *average core number*. We can conclude that both classes *Continuous urban fabric (S.L. : > 80%)* and *Other roads and associated land* have a significant impact on the value of the property *average core number*, which is also confirmed by the algorithm when calculating feature importance, as shown in Figure 6.8.

The property *average weighted degree* shows similar trends as *average PageRank* and *average core number*, which is expected since those properties are correlated. Even though those values are highly correlated, they represent different very important graph properties related to connectivity and transitivity and we considered them independently in relation to land use profiles. Correlation between centrality measures is commonly observed in many different networks, but as they represent different property they might have diverse impact on information flow through the network [121].

Polygons that have the highest value of *average weighted degree* are the ones where the dominant land use class is *Other roads and associated land*. Polygons that have the dominant class *Continuous urban fabric (S.L. : > 80%)* also have a high value of *average weighted degree* but significantly less than the polygons with dominant class *Other roads and associated land*, as shown in Figure 6.11. Polygons with the dominant class *Arable land*, Figure 6.12, have very small *average weighted degree*. We can conclude that the land use classes that have the most impact on *average weighted degree* property are *Other roads and associated land* and *Continuous urban fabric (S.L. : > 80%)* which is also detected by the algorithm when calculating feature importance, as shown in Figure 6.8.

6.6 Discussion

When we think about human dynamics in general, first thought is always related to human physical movement and mobility, but in modern time there are other aspects of human dynamics that need to be considered and further investigated. Virtual activity reflected through social media, location-based services, telecommunications, internet usage, even some online games can have severe impact on shaping the patterns of human dynamics. The relationship between physical and virtual space is indisputable, and the impact that digital activity has on overall shaping of human behaviour patterns cannot be neglected and therefore opens new challenging questions in human dynamics research.

For the propose of this research, we have analyzed human dynamics reflected by telecom traffic network through connectivity links. User generated data is usually very large in size, dynamic and it has complex structure. To analyze such data we have designed and developed a processing pipeline based on Apache Spark Big Data platform using programming language Scala. We have introduced graph filtering as a method for performance enhancement by keeping only statistically significant links in the graph.

Since our data is user generated and reflects the real communication through telecom network, we wanted to see the community structure that such activity forms in order to be able to detect sub areas where the internal information spreading is more significant than outer information flow. To investigate community structure of the network we applied modularity-based algorithm for community detection on a daily basis graphs and observed some persistent patterns over time although the number, structure and spatial distribution of communities differ between days. It was very interesting to notice that communities formed over the urban core of the city are smaller in size but highly dynamic while communities formed over peripheral parts of the city and in sub-urban zones are larger in size and more stable. This observation motivated us to deeper investigate the correlation between land use and community structure. We have created land use profiles for each polygon of the RBS coverage network and quantified graph based and spatial community properties, which contain latent information about human dynamics. We further used the properties to learn the predictive Machine Learning model based on regression. We used the output of the model to evaluate the correlation and predictiveness of properties based on land use.

Our results have shown strong correlation between land use and the properties *average core number*, *average coverage area*, fair correlation for the properties *average weighted degree, average page rank* and *average intersection area*, weak correlation for properties *average diameter, standard deviation of intersection area* and almost no correlation for property *average betweenness centrality*. To deeper investigate the impact of specific land use classes on properties predictiveness we have calculated feature importance.

Based on the results, for predicting all properties, the land use classes *Other roads and associated land, Continuous urban fabric (S.L. : > 80%)* and *Arable land* have the highest impact. The land use class *Other roads and associated land* has the highest impact on predicting all graph based properties. For predicting the property *average core number*, which shows the highest correlation to land use, the land use classes *Other roads and associated land* and *Continuous urban fabric (S.L.: > 80%)* have the highest impact. This can be explained by considering the concept of importance, both in physical space and in networks. In physical geographical space, the important urban zone is the one with many amenities, densely built up zone with developed transit infrastructure and good connections. Such zones would be characterized with high presence of land use classes *Continuous urban fabric (S.L. > 80%)* and *Other roads and associated land.*

In network science, graph centrality measures are used to quantify the importance of a node. Therefore a node with high importance would participate in the high degree core, which means its core number would be high. The results of feature importance calculation indicate that highly important zones in physical geographical space would also have an important role in the virtual telecom traffic network. Similar behaviour could refer to the property *average weighted degree*, its expected that highly urban zones would reflect to network nodes with high weighted degree. For predicting the property *average page rank* the most important land use class is *Other roads and associated land*. This can be explained by considering the concept of transitivity. Transitivity of the urban zone is reflected in the prevalence of important roads, public transport and other transport infrastructure, while the transitivity of a node in the network is associated with its page rank property. Our results indicate that transitivity in physical space is linked with transitivity in the telecom connectivity network.

The results of this study lead us to the conclusion that physical geographical space and virtual communication space need to be considered together as one entity.

Besides spatial component, telecom data carries significant information with time component as well. Our communities can be described as space–time phenomena, since both components have important role in forming the communities. To evaluate space–time dynamics of communities we have introduced some new measures called - spatial community properties.

Spatial community property that shows the highest correlation with land use is *average coverage area*. To predict the property *average coverage area* the most impact has the land use class *Continuous urban fabric (S.L. : > 80%)* and *Arable land*. Based on the feature importance result both classes have significant impact, although they predict opposite values. Polygons where dominant land use class is *Arable land* tend to have very high coverage area, while those where the dominant land use class is *Continuous urban fabric (S.L. : > 80%)* form smaller coverage area. This can be explained by observing the size and granularity of the neighbouring Voronoi polygons in the highly urban and non urban area. In highly urban areas, communities are formed by many small polygons, while in non urban zones communities are formed by few very large polygons, affecting the area of the community. The property that is the most related to spatial dynamic of community is *mean intersection area*. If a polygon is forming a stable community structure that does not differ much between days, its *mean intersection area* would be high comparing to polygons which form dynamic community structure with little or non spatial overlapping between days. The land use class *Arable land* has the most impact on predicting *mean intersection area*, but it is interesting to notice that the land use classes associated with highly urban zones such as *Continuous urban fabric (S.L. : > 80%)* and *Other roads and associated land* do not show significant predictive power. Based on the results, the land use class *Arable land* is very predictive for the property *mean intersection area* and it indicates high values of the property, but the opposite pattern is not observed. Land use classes that are associated with highly urban zones, *Continuous urban fabric (S.L. : > 80%)* and *Other roads and associated land* do not seem to have any significant impact on predicting the value of the property *mean intersection area*, neither high or low. This imply that even in the highly urban zones it is possible to form stable community structure that would persist in size and spatial distribution over time.

The results of this study could be beneficial for urban planning and city policy making since it emphasize the correlation between land use and human dynamics reflected through properties evaluated from telecom connectivity network. Although human dynamics is often considered through the lenses of movement and mobility patterns, communication and digital activity shouldn't be neglected when analysing human behaviour patterns. Telco providers are facing many challenges related to network infrastructure management, demands of new services which needs to be supported by the infrastructure, etc. With knowing the importance of the place, its transitivity and dynamics telco providers could enhance and optimize the network infrastructure and services they are providing.

The topic of human dynamics has been studied by many researchers from diverse disciplines over years. In the era where technology is present in our every day life more then ever, when data records are generated by almost every action we take, it is a huge challenge for the research community to design and develop methods to explore human dynamics which will help us answer essential questions about urban development and society in general.

In this chapter we presented how connectivity graph properties can be used to evaluate human dynamics. Results of our experiments show that the properties of connectivity network are strongly correlated to location semantic. Depending on the dominant land use type in formed community we can obtain the different properties of the community which indicate how that community evolves through time. It is very challenging task to combine the space and time evolving features into one measure, and spatial community properties that we introduced in this chapters provide exactly that, therefore we find it is a significant contribution from this research work.

Chapter 7

Mobility networks in urban spaces

Mobility has always been one of the key aspects of human life. In modern time when most people life in densely populated urban areas, daily commuting and mobility in general became crucial. For the long time mobility networks in urban spaces are considered to be traffic networks, but with the massive expansion of Internet and GPS technologies, new concept of mobility networks started to arise - virtual mobility network. For the purpose of our research we explored the mobility network formed using location data from Foursquare [2], which is described further in this Chapter. Some of the results from our research that are presented in this Chapter are published in the paper [84].

Location data became ubiquitous due to global adoption of smartphones, the worldwide availability of the GPS and advanced location-based applications. The value of such data is immense as they contain spatial-temporal patterns of massive number of people. Among popular location-based applications is Foursquare, a location based social network founded in 2009. Foursquare provides cloud-based location technology platform together with mobile app where users are able to notify their friends about their current location through check-ins for which they can receive virtual rewards. Apart from that, it allows users to a leave note about their experience in the specific venue, which can be utilized for building a recommendation system. As commercial company that works with privacy sensitive personal data Foursquare is obliged to follow strict rules to preserve data privacy, despite that they wanted to explore further the value of their data and opened one restricted data set for Future Cities Challenge [1]. With its initiatives to open some of data they collect, Foursquare attracted researches to explore their rich source of information and evaluate its potential for understanding social behaviour, mobility, and propose location intelligence services.

Mobility if one aspect of human dynamics that drives many processes in the urban environment. Although there is massive amount of data collected about human mobility through mobile phones, the data is essentially very private, and companies that collect such data are not very eager to share it. Similar like for telecom connectivity data, prior to releasing the data it must be aggregated and anonymized. Such data reflect the collective mobility from one place to another within the city through some timespan, which reveals a lot about overall peoples habits and behaviour in the city. In this Chapter are presented the results of such case study using aggregated Foursquare check-in data. Our results show that location semantic is essential to human dynamics together with time component and that graph properties evaluated from the network can be used as an indicator of human dynamic.

7.1 Introduction

Many research efforts were dedicated to the analysis of Foursquare data and interesting patterns were discovered. Preoţiuc-Pietro applied k-means clustering on users and used the result for the prediction of user future movements [93]. Joseph et al. clustered users via topic modeling, an approach that is usually used in the classification of text documents according to the latent themes [63]. On the opposite, Cranshaw et al. performed a clustering algorithm on venues regarding the spatial and social characteristics of venues [38]. Jun Pang et al. applied algorithms PageRank and HITS on Foursquare data for the purpose of performing friendship prediction and location recommendation [88]. D'Silva et al. used Foursquare data and machine learning in order to predict the crime [43] and Noulas et al. used machine learning on the data with the aim of predicting the next venue that user will visit [81]. Yang et al. explored the tourist-functional relations between different POI types present in Foursquare data in the city of Barcelona [134]. Moreover, researchers made a comparison of Foursquare data and data from additional location-based services (LBS) in order to check the similarity of patterns, the validity of check-ins, etc [110, 138]. Foursquare data were also utilized to characterize competition between new and existing venues [39] by measuring change in throughput of a venue before and after the opening of a new nearby venue.

Clusters formed from mobility network are essentially the same as communities formed from telecom connectivity network. Therefore, we have community from connectivity network and cluster from mobility network. The underlying algorithm and methodology in the same, just the terminology is a bit different due to the way we understand what is a community vs. cluster. Similar like communities, clusters formed from mobility networks in physical space represent the group of nodes that are strongly connected between each other within the group comparing to their connectivity outside of the group. In the context of Foursquare data, clusters represent the group of venues that are frequently visited together by the users. Such clusters reveal a lot about human dynamics, because they gather the location types with the similar behaviour.

7.2 Data

Foursquare provided data for Future Cities Challenge [1] for ten different world wide cities (Chicago, Istanbul, Jakarta, London, Los Angeles, New York, Paris, Seoul, Singapore and Tokyo).

For each of the ten cities included in the data set there are two corresponding files:

- a *venue information* file where you can see the id, name, coordinates and category of a venue in each line,
- a *movements file* where each line corresponds to an edge between a pair of venues, the month and year that movements were aggregated for the given venue pair, the “period of the day” that arrival to the destination venue took place and the "weight" of an edge.

The period of the day is divided in five categories: overnight (between 00:00:00 and 05:59:59), morning (between 06:00:00 and 09:59:59), midday (between 10:00:00 and 14:59:59), afternoon (between 15:00:00 and 18:59:59), and night (between 19:00:00 and 23:59:59). The "weight" corresponds to the number of check-ins that took place by any user for the given venue pair.

Number of venues differ for each city, where the city with highest number of venues is Istanbul, followed by Tokyo and New York (Figure 7.1).

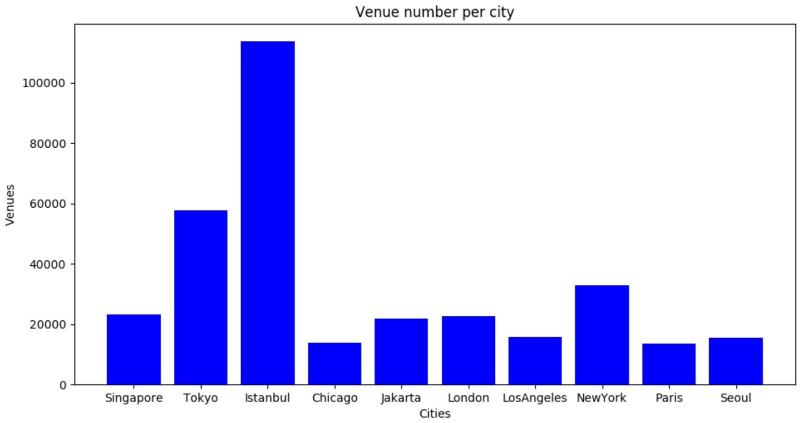


Fig. 7.1 Number of venues per city.

Number of movements per city also differ, but with the smaller overall difference. Approximately, there is around 7 million movements between venues per city, which makes this data very complex to analyse using graph theory. Therefore, the need for advanced Big Data technologies for data processing is indisputable. In the table 7.1 is presented full number of movements per city, average number of movements per month per city and coefficient of variation between monthly values per each city. Coefficient of variation (CV) is a measure of dispersion, it presents the relationship between standard deviation and the mean, which is very important in geosciences [26]. The coefficient of variation is very convenient statistical tool because it permits the comparison of values from the distribution free from scale effects; i.e., it is dimensionless.

From the table 7.1 we can observe that number of movements per city and average number of movements per month are more or less the same for each city which yields that our data sample is carefully selected to present the equal amount of data per each city and therefore keep hidden the true diversity between the global markets of Foursquare services. For our research is more important actually the coefficient of variation, presented in the column "CV". CV has values close to zero for each city which indicates that there is no high variation between number of movements per month. Such unified distribution enables us to treat equally each monthly data sample, which we further explored during performance testing.

Besides monthly distributions, we observed the distribution of the average number of edges per each city per day period, which is presented in Figure 7.2. From the Figure 7.2 we can see the rise in the mobility from morning to midday and fall in the mobility from night

Table 7.1 Number of movements per city in Foursquare data.

city	**# movements**	**monthly mean**	**CV**
Chicago	7,775,376	323,973.96	0.0032
Istanbul	7,372,799	307,199.92	0.0151
Jakarta	7,801,368	325,056.96	0.002
London	7,650,994	318,791.38	0.0071
Los Angeles	7,721,731	321,738.75	0.007
New York	7,805,871	325,244.58	0.0042
Paris	7,574,139	315,589.08	0.0062
Seoul	7,768,926	323,705.21	0.0027
Singapore	7,723,757	321,823.17	0.0034
Tokyo	7,798,240	324,926.62	0.0019

to overnight period for each city, which is expected. The period between midday and night, which covers the timespan between 10AM and midnight is where the most variety happens. We can see that Istanbul has almost constant number of movements between midday and night, while Los Angeles has much higher peak in the midday but drops rapidly in the night. Tokyo follows the same pattern as Los Angeles with almost constant down fall from midday to night but with smaller slope. Paris shows the similar pattern as Istanbul but with a little bit higher peak in the midday and slight downfall in the night. London, Chicago and New York follow the same pattern of high peak in the midday that slightly drops to the afternoon and significantly drop in the night. Singapore comes as the most surprising, having unique pattern with high peak in the midday, then downfall in the afternoon and then again peak in the night.

From descriptive analysis we can conclude that even with the limited and somewhat normalised data sample we can observe different patterns within peoples movements world wide which further emphasize the diversity of habits, behaviour patterns and lifestyle culture between people.

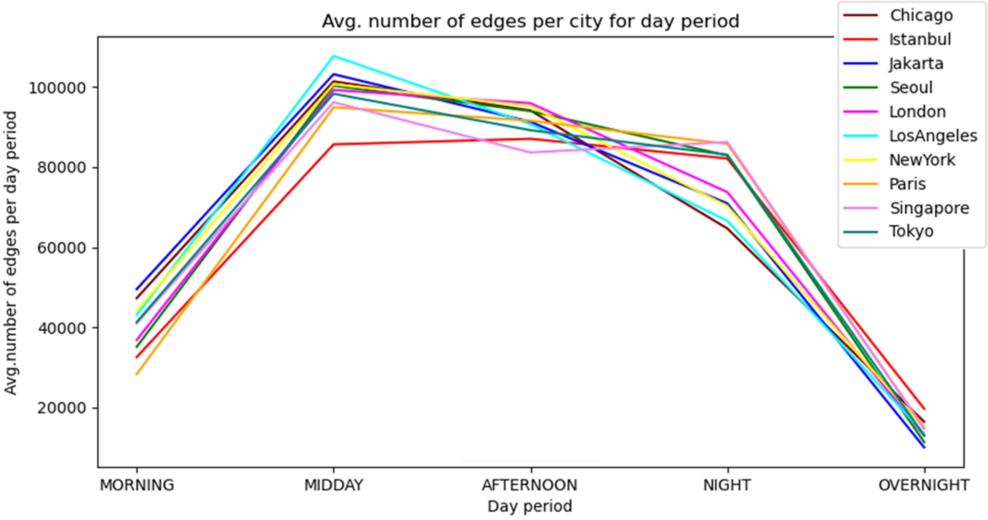


Fig. 7.2 Avg. number of movements per city for day period.

Anonymized and aggregated mobility data provide opportunity to study cities as complex systems, in the way that before massive usage of GPS connected devices wasn't possible. Based on our extensive research that is presented in this thesis, specially in Chapters 4, 5 and 6, we were motivated to explore the cluster structure of venues based on mobility flows, the semantic content of clusters and how cities compare to each other in terms of semantic content of detected clusters. All of those topics are relevant for building different service applications based on recommendation systems and therefore can open many new research questions in the domain of human dynamics research.

7.3 Methodology

To explore the dynamic structure of the user generated movements across the urban spaces we applied clustering technique. Clustering grouped venues based on the statistical similarity and gave us the insight of how venue communities evolve through time. Due to the size and complexity of data we decided to use Apache Spark platform for distributed analysis to enhance the performance of graph clustering analysis, as that could be a large bottleneck in computation.

The first step in our analysis is to generate graphs from input Foursquare data on a monthly basis. In these graphs, nodes correspond to venues, and edges are created from aggregated movements between two consecutive venues. If the movement occurred more than once during different days or day time periods, the weights are aggregated, so in the final graph we have unique edges over one month. Second step in our analysis is to cluster the venues based on movements across the city. To do so, we used Louvain algorithm [21] which proved to be very efficient when working with large, complex graphs as previously described in the Chapter 5.

Clustering in complex networks is a computationally challenging problem. To keep the processing performance high it is common to use approximation algorithms which are used for optimization problems [47]. When optimizing the solution the cost function is defined with the aim to find maximum or minimum value of given function. The Louvain algorithm proposed in [21] uses the modularity 2.5 as a cost function. Louvain algorithm starts by assigning each node in the graph to its own community, and then it begins to move the nodes around and assign them to neighbouring communities, and calculate the modularity based on equation 2.7 at each level. As long as the modularity grows, it continue to move the nodes around, and when the modularity reaches it's local maximum the algorithm terminates and returns the community sets from the previous level as a result.

The result of community detection in Foursquare data is the set of venues assigned to different clusters (or communities) which can be presented on geographical maps, as the data is spatially referenced. The example of clustering analysis for city of Chicago for the month April 2017 is presented in Figure 7.3. From the Figure 7.3 we can notice how venues tend to cluster in spatial proximity. Frequent movements are occurring between places that are spatially close, which is common behaviour described with the Tobler's first law of geography: *"Everything is related to everything else, but near things are more related than distant things"*. Similar behaviour is observed in clustering of telecom traffic data, Chapter 5.

Formed clusters within the city differ in size and spatial distribution. It is observed that in a closer proximity to the city center clusters are formed with smaller number of venues that are densely grouped together, compared to the peripherally located clusters which have many venues widely distributed in space.

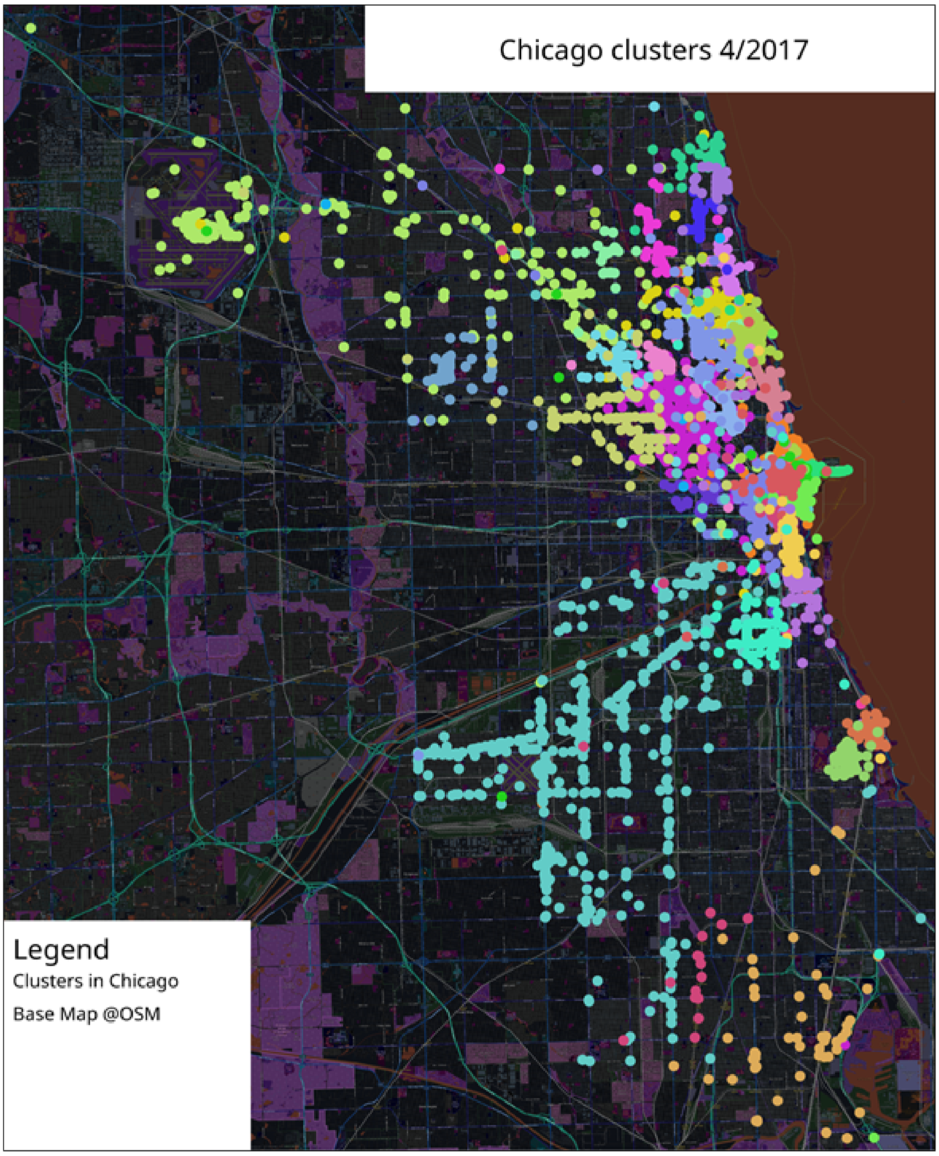


Fig. 7.3 Venue clusters for city Chicago, 2017-04

To gain better understanding of urban dynamics and how location semantic is affecting cluster creation within the city, we further investigated the categories of venues inside each cluster. Semantic categories are described at Foursquare web site, in the section for developers[1], we focused on nine super categories: *Art & Entertainment, College & University, Event, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other, Shop & Service, Travel & Transport, Other*. Category *Other* is used for those venues that do not fit in any of the main categories.

The presence of each category within the cluster is estimated by calculating percentage among all venues in the cluster. Each city has unique digital footprint of the categories that are dominant across clusters. There are similar patterns between some cities, while significant difference emerged between some other.

7.4 System architecture and performance

Mobility data about peoples movements within densely populated urban areas is usually very large in size and graphs produced from such data are also very complex to analyse due to high number of weighted edges. Therefore working with mobility data demands usage of Big Data tools and algorithms that have good scaling properties, which isn't always simple to achieve. To overcome this issue, we decided to perform all intensive calculations using Apache Spark. To justify how well Spark really improves processing performance we conducted the performance testing using Foursquare mobility data.

In this Chapter we will present the results of the performance evaluation. We developed the processing pipeline based on Hadoop Ecosystem, where Spark engine is used as a distributed computing layer and HDFS together with Hive is used as a storage layer. Raw movements data in csv format is loaded into Spark in the DataFrame structure. From that point on the data is distributed and before any processing it is kept in the main memory. First processing step is making graphs from raw data which is computationally very intensive task when there is high number of edges. When graphs are made they are saved in HDFS using Hive and it's orc format. Graphs kept that way are still distributed and could be accessed only by loading back to distributed data structure. Next step is performing community detection using Louvain algorithm 1 which is in detail explained in Chapter 5 of this thesis. Results of community detection are stored in Hive. Last computationally intensive step is joining each venue from the communities with its geographical coordinates so we could spatially visualize and analyse communities. In that step we performed additional operation which was not necessary in the previous two steps, that is saving the results to common csv file outside of HDFS. Since our data is distributed, after joining with coordinates in the process of saving it to csv file we need to apply "coalesce" function with the parameter 1 in order to obtain just one csv file with all data as a result. That operation is memory intensive since it requires the

[1] Foursquare web

Driver node to pull all data from its Slave nodes and merge it to one file. When we have csv file with communities containing geographical coordinates we further used QGIS to analyse it spatially. Analytical pipeline is presented in the Figure 7.4.

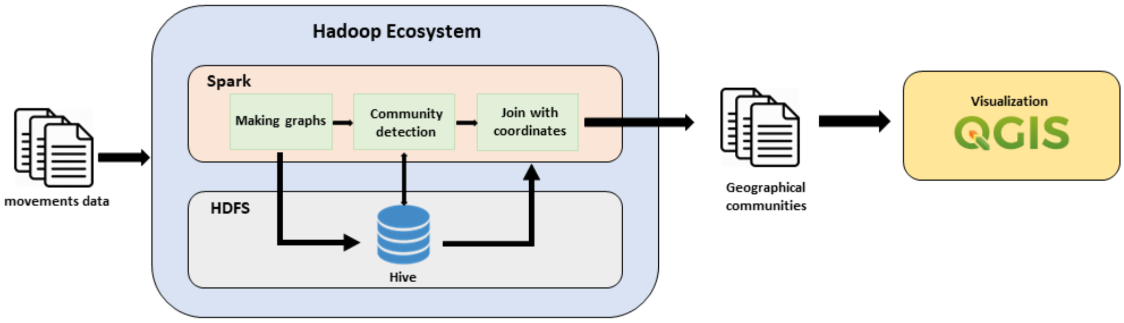


Fig. 7.4 Analytical pipeline

For performance testing we randomly selected three different monthly graphs, Seoul 2017-08, New York 2017-12 and Istanbul 2017-09. As described in Section 7.2, average number of edges per city are similar and also coefficient of variation between months for each city is close to 0 meaning we could select graphs randomly with confidence that our choice of the graphs wouldn't effect the performance significantly. Experiments are performed using single machine with processor AMD Ryzen 7 5800U 1.90 GHz and available RAM memory of 16GB. The processor AMD Ryzen 7 5800U 1.90 GHz allow us to test three levels of parallelism, when using 2, 4 and 8 cores. In such set up slave nodes aren't physically distributed nodes, but logically distributed cores of the CPU. The parallelism is achieved, with the saving of processing time for the network communication that would be otherwise lost in the cluster environment. We were able to run our processing in such limited set up as our data is finite and not that big in size, which is rarely the case in real life Big Data applications where data sources are often streaming data with infinite data flow.

Let's focus on the runtime for community detection step for three selected graphs, Figure 7.5. Starting point in our experiments is when we set the level of parallelism to 1 meaning there's no parallel computing at all. We can see that there is different runtime for each graph which is expected as graphs differ in size and complexity. As we increase the level of parallelism to 2, 4 and 8 we can observe significant down fall in processing time for all three graphs.

If we take closer look to the exact runtime values in the table 7.2, we can see that by gradually increasing the level of parallelism two times, adding 2, 4 and 8 cores we get the

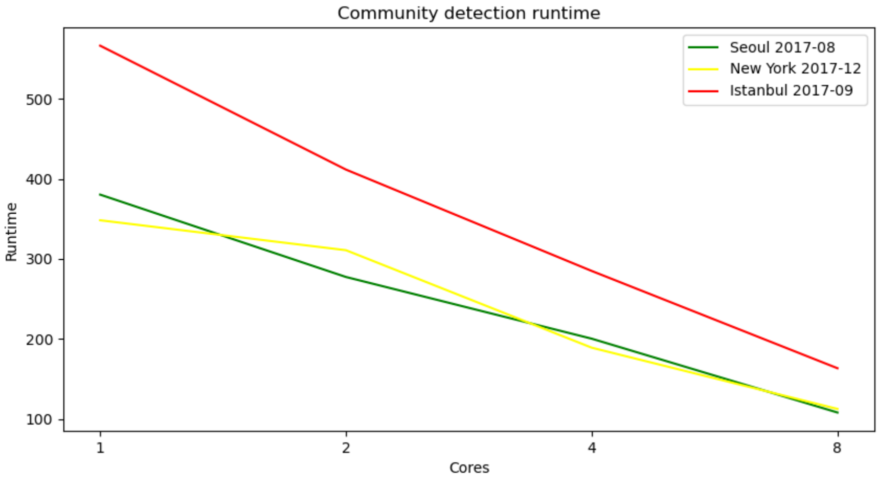


Fig. 7.5 Community detection runtime

improvement in runtime of roughly 20%, 50% and 70%, respectively. With just 8 nodes, or cores in our case, we get speed up in processing time of 70%.

For the operation of making graphs the improvement of runtime is roughly 30%, 50% and 60% when adding 2, 4 and 8 nodes, while for the last step joining with coordinates the improvement between runs is observed but it is not significant. From this experiment we can conclude that the more complex the task is, more improvement in the processing runtime would be seen when adding more nodes and increasing the level of parallelism.

We can notice that there is variation between runtimes for each tested graph and each operation even when we observe values referred to the same level of parallelism. That is expected since our graphs are not same in size or complexity. Graph for Seoul for month 2017-08 has 323259 edges, similar as the graph for New York for 2017-12 which has 326154 edges, while graph for Istanbul for 2017-09 has 301498 edges. These graphs for Seoul and New York have less edges then the graph for Istanbul, but Istanbul is the city with the highest number of venues 7.1. If we take a look at the runtime results for community detection in the table 7.2 we will see that Istanbul has the highest runtime at any parallelism level which comes as no surprise if we consider the nature or the Louvain clustering algorithm 1 that is hierarchical technique starting from assigning each node of the graph to its own community. Therefore, graph for Istanbul although it doesn't have the highest number of edges, but has the highest number of nodes is the most computationally complex.

From the results of this experiment we can conclude that by increasing the level of parallelism in Apache Spark we can significantly improve the processing time for the most complex tasks.

Table 7.2 Spark performance testing

Seoul 2017-08			
cores	making_graphs [s]	community_detection [s]	join_coord_graph [s]
1	89.53	380.3	25.51
2	54.74	277.51	24.44
4	43.81	200.49	23.07
8	35.1	108.34	22.73
New York 2017-12			
cores	making_graphs [s]	community_detection [s]	join_coord_graph [s]
1	91.4	348.29	28.84
2	65.94	310.92	25.44
4	45.13	189.25	23.77
8	32.88	112.81	23.12
Istanbul 2017-09			
cores	making_graphs [s]	community_detection [s]	join_coord_graph [s]
1	87.92	566.18	33.98
2	63.48	411.55	27.59
4	46.97	285.03	25.69
8	36.3	163.6	23.48

7.5 Results

Community detection in Foursquare data is performed by applying graph clustering over graphs made from movements data, for each city, for each month between April, 2017 and March, 2019. The results obtained from clustering differ between cities, even between months for the same city. In Figure 7.6 is presented the variability of clusters across the cities. The plot shows the dependency between the average number of clusters and their size in each city per month.

From the Figure 7.6 we can observe different patterns for each city. Although Istanbul has the highest number of clusters, there are relatively small. However, the opposite pattern can be noted for the city of Paris in which clusters are relatively large, but there are fewer compared to Istanbul. Moreover, some similarities between cities can be observed. The cities of Chicago and Los Angeles have relatively similar number and size of clusters.

Furthermore, to explore the impact of location semantic to cluster forming we analysed which categories are present in the clusters. We selected the largest clusters across each city, those that are consisted from more than 50 venues and calculated percentage of occurrences for each category. Presence, variation and distribution of categories inside clusters can give

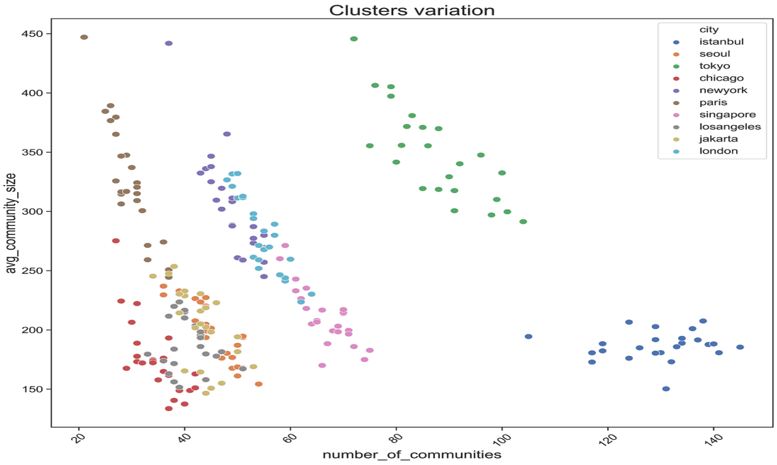


Fig. 7.6 Variation of clusters within the city per month

us valuable input about venues semantic that are strongly connected by users movement. In the Figure 7.7 is presented the largest cluster in the city of Chicago classified by category. From Figure 7.7 we can notice high variety of categories, where the most present category is the *Food*, followed by *Shop&Service*. It implies that people in this cluster generally move between places related to food and shopping.

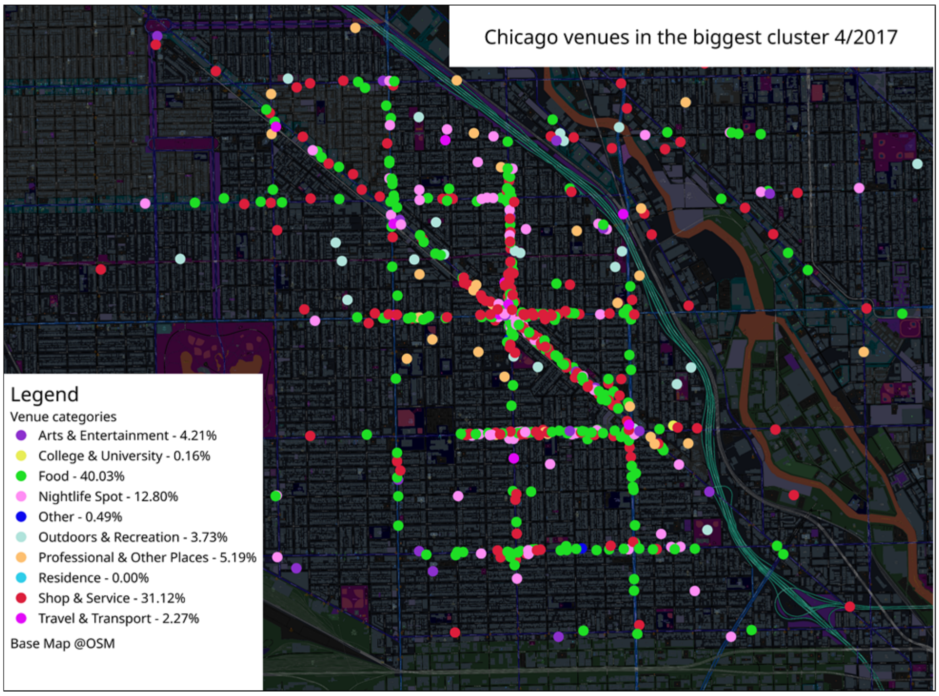

Chicago venues in the biggest cluster 4/2017
Legend
Venue categories
● Arts & Entertainment - 4.21%
● College & University - 0.16%
● Food - 40.03%
● Nightlife Spot - 12.80%
● Other - 0.49%
● Outdoors & Recreation - 3.73%
● Professional & Other Places - 5.19%
● Residence - 0.00%
● Shop & Service - 31.12%
● Travel & Transport - 2.27%
Base Map @OSM

Fig. 7.7 Spatial distribution of categories across biggest cluster in Chicago

In the city of Chicago another large cluster is formed around O'Hare International Airport, in which the most present categories are *Travel&Transport* and *Food*. From Figure 7.8 we can notice how venues are spread in almost regular form following Interstate 90 road, which is one of the main highways in the State of Illinois. As can be seen, clusters are formed around spatially close or well connected places with some categories frequently occurring together. Consequently, we can classify clusters by dominant presence of one, two or even more categories.

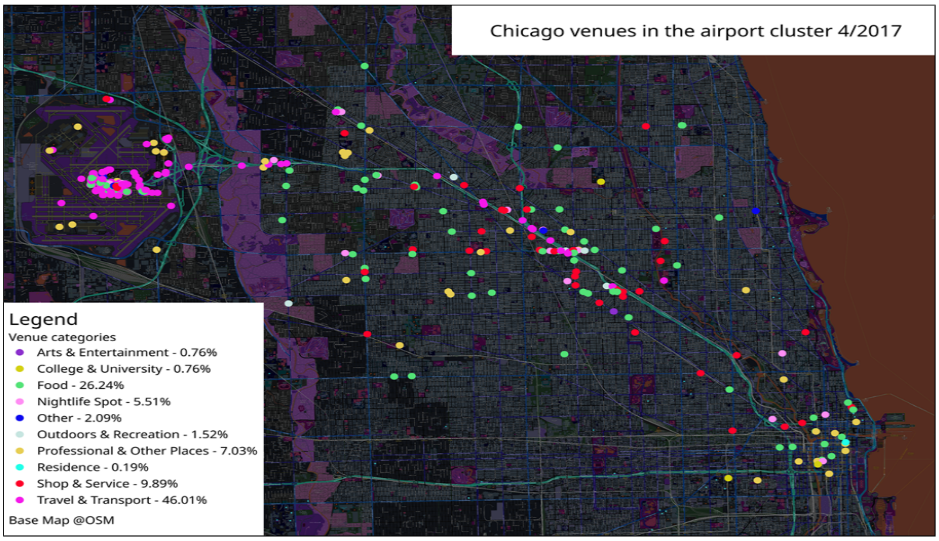

Chicago venues in the airport cluster 4/2017
Legend
Venue categories
● Arts & Entertainment - 0.76%
● College & University - 0.76%
● Food - 26.24%
● Nightlife Spot - 5.51%
● Other - 2.09%
● Outdoors & Recreation - 1.52%
● Professional & Other Places - 7.03%
● Residence - 0.19%
● Shop & Service - 9.89%
● Travel & Transport - 46.01%
Base Map @OSM

Fig. 7.8 Airport cluster Chicago

To obtain global view and compare the cities, we performed hierarchical clustering based on average profile of probability distribution of categories. Result of hierarchical clustering in the form of dendrogram (Figure 7.9) showed which cities are similar, with colors of branches indicating how we could group them.

From Figure 7.9 we can notice high similarity between US cities Chicago, New York and Los Angeles, and European cities Paris and London, while Tokyo with its community profiles stands between US and European cities. Another group of similar cities include Istanbul, Jakarta and Singapore, while Seoul has unique pattern completely different from all cities. To provide more details, we present semantic profiles of venue clusters for four different cities Istanbul, Seoul, Chicago and Tokyo (Figure 7.10).

From visual inspection of the profiles we can notice that category *Food* has high peak in each city, while category *Residence* has low peak. This implies that people tend to "check-in" more frequently at outgoing places such as restaurants then when they come home, which is natural. With comparative analysis between profiles we could notice some general trends related to category variability between clusters. We can conclude that in

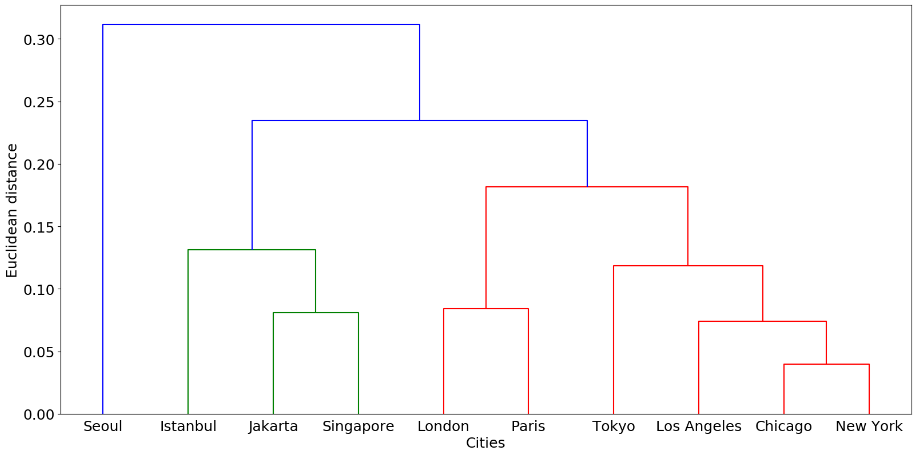


Fig. 7.9 Hierarchical tree presenting the similarity and diversity between cities

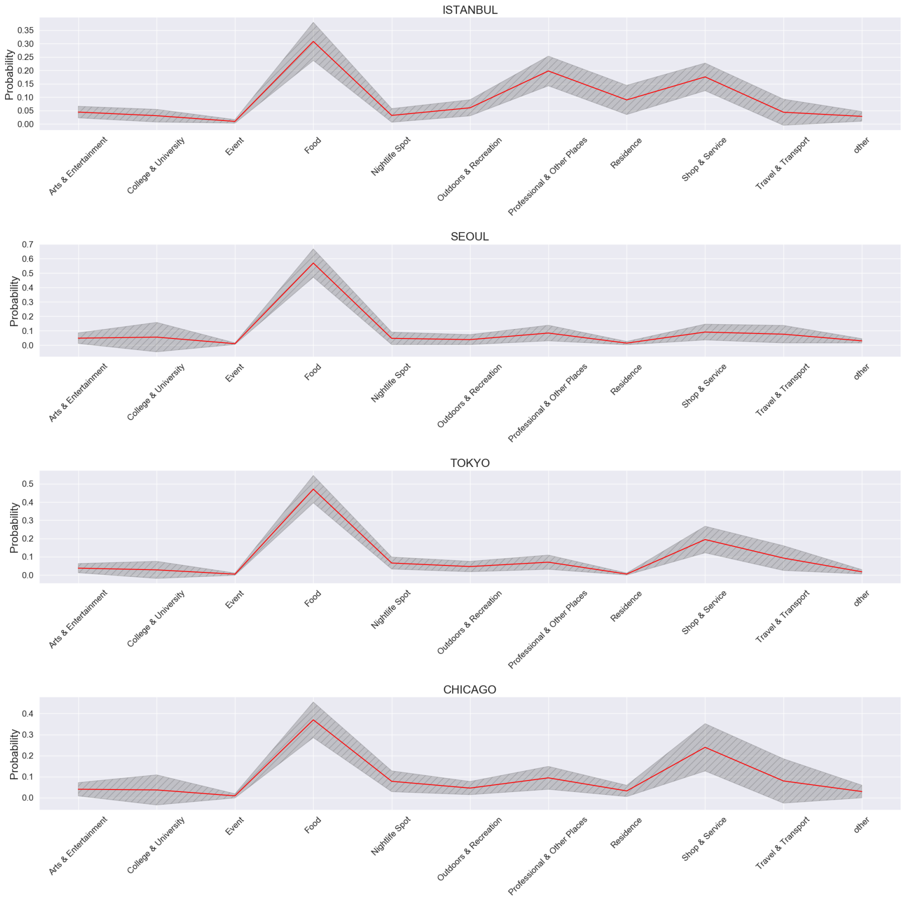


Fig. 7.10 Mean profile and standard deviation of categories probability distribution in detected clusters

Chicago and Tokyo people are very likely to move between venues related to categories *Food* and *Shop&Services*. In Istanbul people are very likely to move between venues related to categories *Food, Shop&Services* and *Professional&Other*, while Seoul has strong dominance of *Food* category. More specific city profiling is possible with exploring more detailed subcategories that are present in the cluster.

7.6 Discussion

Mobility networks generated by users are valuable data source for exploring urban spaces. By performing clustering over graphs made from mobility data we gain deeper knowledge about grouping of venues and mobility flows. With further investigation of venue semantics inside clusters we can detect location types and categories that are frequently visited together by users, which is potentially valuable information for urban planing, retail and services. Detecting relations between clusters and venues inside cluster can help us in building a recommendation application that would serve users who are visiting new cities, based on their preferences. Majority of venues in cluster are either spatially close or they are well connected with transport infrastructure, indicating that users tend to move between locations in limited spatial distance forming in this way urban sub-spaces. Knowledge about urban sub-spaces that stand out as entities could be very valuable input for urban policy making and development, and also for developing new services.

By utilizing user generated location data, knowledge about location semantic and advanced data analytics techniques such as clustering, evolving network analysis and machine learning we can gain insight into human dynamics that was previously impossible to inspect, and therefore support better urban development and service planning.

Our results go beyond the basics of our three main hypothesis, because in this case study we explored the mobility data for ten different world wide cities. During the exploratory phase we have notice that day time matters, and that different patterns occur in different day time, but that was expected with mobility data. In the end, most people are not very active in the night, while some usual work related hours can be applied to the most of the people also. Location semantic also matters, because people tend to move between the places that are logical for their activities. For example, the most visited location category in each city is *Food*, while that category is frequently visited together with the category *Shop&Services*, which means that people tend to eat while shopping, which is common behaviour known to all of us. Such, and similar findings can lead to more efficient development of urban spaces, because the urban planners could plan to put in the close proximity objects and services that people actually use. While exploring the clusters of locations frequently visited together we noticed that different cities have different patterns but some are more a like than others. We have noticed that US and European cities have similar patterns, while Asian cities are similar between each other but different than US and European, Tokyo is somewhere in between US and European while Seoul has completely unique pattern. This finding is very important for analysing cultural differences and how those differences can affect the overall behaviour of the people and their dynamics.

Chapter 8

Conclusions

With the positional technologies constantly evolving and becoming easily accessible to people in their every day life, people around the world are generating more and more spatial data, every minute of the day. Now, more than ever in the history of human evolution is the question of "where" something is happening important. People are interacting between each other without spatial barriers through virtual services and telecommunications leaving behind precious digital footprints that can reveal much about human behaviour, dynamics, habits, lifestyle, cultural differences, etc. The boundary between physical and virtual spaces is becoming more loose with technologically advancements that enable us to search online, navigate, communicate online, shop, pay bills online, educate online etc. With so many things we can do “online”, the definition of human dynamic have to evolve, because people can be very "dynamic" even if they are not moving at all! Nevertheless, there are some connection between physical and virtual space, the same patterns tend to occur in both dimensions. Good example of this peculiar behaviour is described in Chapter6 in the Results section.

Namely, our experiments showed that for predicting the property *average page rank* the most important land use class is *Other roads and associated land*, which can be explained by considering the concept of transitivity. What is “transitivity” in physical space related to transport infrastructure, the same applies for the “transitivity” in virtual telecom connectivity network associated with the page rank property. This result indicate that transitivity in physical space can be linked to the transitivity in virtual space, both evaluated from human dynamic.

Another example from our experiments that show how strong is the correlation between physical and virtual space and human dynamics, is presented in Chapter 4. When we examined local graph properties we observed separate zones of the city, because zone in the city is a node in the graph. Namely, high peak in betweenness centrality of the

Zone 23 is noticed on 15. December. In the same time at that city zone there was a film festival happening, which certainly influenced higher number of people coming to that zone. Betweenness centrality tells us the number of shortest paths that traverse the node, meaning if the betweenness centrality is high, that node is locally important for the information flow in the graph. If higher number of people suddenly come to one location, then that location becomes locally important. That way we assembled the concept of local importance in both physical and virtual space.

With high adoption of smart phones, more and more application and services are becoming "mobile", and while those applications are usually meant to make our lives easier the downside of its high usage is the collection of user personal and location data by commercial companies, which significantly highlighted the question of data privacy in the past few years. More traditional industries, such as telecommunications, are legally regulated and they are obliged to follow strict procedures to keep the users private data safe and secured. On the other hand, more modern companies that are based on internet services are not so strictly regulated, because at the beginning of their presence on the market it was unclear how this would affect user privacy. Today, many governments are trying to introduce the legal regulations related to data privacy, so in Europe we have GDPR that many companies working with users data are obliged to follow.

Even without GDPR regulation, companies are not very eager to share their data for research purpose despite the fact that such research could lead to the development of new services that they could embed in their commercial offering. Large private companies are more likely to form their own research groups and explore the data within the company, but such practice is limiting the potential of data exploration to the areas that are highly valuable to the companies. That usually leave the social good research behind. Despite the obstacles, some companies did recognise the value of scientific, non commercial research and are willing to share their data for research purpose. Besides commercial data, to explore human dynamics we need to utilize large scale spatial data, which puts an extra pressure on computational power, due to the need to extract the results fast. To overcome this issue, we need to use advanced Big Data technologies for data analytics such as Apache Spark and Hadoop, and to move our processing from limiting personal computers to the distributed cloud or cluster computing environment.

In this thesis we presented the novel approach to human dynamics research that uses advanced Big Data technologies to extract meaningful results from user generated data. We presented the methodology how to utilize graph mining, Machine Learning and data analytics to extract meaningful knowledge from diverse data sources such as telecom data, social networks data and spatial data. We have explained the results from the aspect of location semantic and evolving networks, which are key indicators of human dynamics. We plan to extend this research by deeper exploration of time evolving component of networks and by employing new data sources.

References

[1] (2019). Future cities challenge website. https://www.futurecitieschallenge.com/.

[2] (2022). Foursquare company website. https://foursquare.com/.

[3] Acs, G. and Castelluccia, C. (2014). A case study: Privacy preserving release of spatio-temporal density in Paris. In *20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining–KDD '14*, pages 1679–1688, New York. ACM.

[4] Aggarwal, C. C. and Wang, H. (2010). *Managing and Mining Graph Data*. Springer.

[5] Aji, A., Wang, F., Vo, H., Lee, R., Liu, Q., Zhang, X., and Saltz, J. (2013). Hadoop-gis: A high performance spatial data warehousing system over mapreduce. In *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, volume 6. NIH Public Access.

[6] Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97.

[7] Alexander, L., Jiang, S., Murga, M., and González, M. C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58:240–250.

[8] Alsheikh, M. A., Niyato, D., Lin, S., Tan, H., and Han, Z. (2016). Mobile big data analytics using deep learning and apache spark. *IEEE Network*, 30:22–29.

[9] Armbrust, M., Xin, R. S., Lian, C., Huai, Y., Liu, D., Bradley, J. K., Meng, X., Kaftan, T., Franklin, M. J., Ghodsi, A., et al. (2015). Spark sql: Relational data processing in spark. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pages 1383–1394. ACM.

[10] Aung, T., Lwin, K. K., and Sekimoto, Y. (2019). Identification and classification of land use types in yangon city by using mobile call detail records (cdrs) data. *Journal of the Eastern Asia Society for Transportation Studies*, 13:1114–1133.

[11] Bajardi, P., Delfino, M., Panisson, A., Petri, G., and Tizzoni, M. (2015). Unveiling patterns of international communities in a global city using mobile phone data. *EPJ Data Science*, 4:1–17.

[12] Barabási, A.-L. et al. (2016). *Network science*. Cambridge university press.

[13] Barlacchi, G., De Nadai, M., Larcher, R., Casella, A., Chitic, C., Torrisi, G., Antonelli, F., Vespignani, A., Pentland, A., and Lepri, B. (2015). A multi-source dataset of urban life in the city of milan and the province of trentino. *Scientific data*, 2:150055.

[14] Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101(11):3747–3752.

[15] Becker, R. A., Caceres, R., Hanson, K., Loh, J. M., Urbanek, S., Varshavsky, A., and Volinsky, C. (2011). A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4):18–26.

[16] Bernini, A., Toure, A. L., and Casagrandi, R. (2019). The time varying network of urban space uses in milan. *Applied Network Science*, 4(1):1–16.

[17] Bianconi, G., Darst, R. K., Iacovacci, J., and Fortunato, S. (2014). Triadic closure as a basic generating mechanism of communities in complex networks. *Physical Review E*, 90(4):042806.

[18] Blondel, V. D., Decuyper, A., and Krings, G. (2015). A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):10.

[19] Blondel, V. D., Esch, M., Chan, C., Clérot, F., Deville, P., Huens, E., Morlot, F., Smoreda, Z., and Ziemlicki, C. (2012). Data for development: the d4d challenge on mobile phone data. *arXiv preprint arXiv:1210.0137*.

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

[21] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008b). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

[22] Bogomolov, A., Lepri, B., Larcher, R., Antonelli, F., Pianesi, F., and Pentland, A. (2016). Energy consumption prediction using people dynamics derived from cellular network data. *EPJ Data Science*, 5(1):13.

[23] Brdar, S., Gavrić, K., Ćulibrk, D., and Crnojević, V. (2016). Unveiling spatial epidemiology of hiv with mobile phone data. *Scientific reports*, 6.

[24] Brdar, S., Novović, O., Grujić, N., González-Vélez, H., Truică, C.-O., Benkner, S., Bajrovic, E., and Papadopoulos, A. (2019). *Big Data Processing, Analysis and Applications in Mobile Cellular Networks*, pages 163–185. Springer International Publishing, Cham.

[25] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[26] Brown, C. E. (1998). Coefficient of variation. In *Applied multivariate statistics in geohydrology and related sciences*, pages 155–157. Springer.

[27] Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., and Ratti, C. (2010). Real-time urban monitoring using cell phones: A case study in rome. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):141–151.

[28] Calabrese, F., Ferrari, L., and Blondel, V. D. (2014). Urban sensing using mobile phone network data: A survey of research. *ACM Comput. Surv.*, 47(2):25:1–25:20.

[29] Capriolo E., Wampler D., R. J. (2012). *Programming Hive*. O'Reilly Media, Inc.

[30] Carlucci, M., Chelli, F. M., and Salvati, L. (2018). Toward a new cycle: Short-term population dynamics, gentrification, and re-urbanization of milan (italy). *Sustainability*, 10(9).

[31] Chambers, B. and Zaharia, M. (2018). *"Spark: The Definitive Guide"*. "O'Reilly Media, Inc.", "Sebastopol, CA, USA".

[32] Chambers B., Z. M. (2018). *Spark: The Definitive Guide*. O'Reilly Media, Inc.

[33] Chen, C. P. and Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information sciences*, 275:314–347.

[34] Cici, B., Gjoka, M., Markopoulou, A., and Butts, C. T. (2015). On the decomposition of cell phone activity patterns and their connection with urban ecology. In *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pages 317–326.

[35] Cook, D. J. and Holder, L. B. (2006). *Mining Graph Data*. John Wiley & Sons.

[36] Costa, L. d. F., Rodrigues, F. A., Travieso, G., and Villas Boas, P. R. (2007). Characterization of complex networks: A survey of measurements. *Advances in physics*, 56(1):167–242.

[37] Cottineau, C. and Vanhoof, M. (2019). Mobile phone indicators and their relation to the socioeconomic organisation of cities. *ISPRS International Journal of Geo-Information*, 8(1):19.

[38] Cranshaw, J., Schwartz, R., Hong, J., and Sadeh, N. (2012). The livehoods project: Utilizing social media to understand the dynamics of a city. In *Sixth International AAAI Conference on Weblogs and Social Media*.

[39] Daggitt, M. L., Noulas, A., Shaw, B., and Mascolo, C. (2016). Tracking urban activity growth globally with big location data. *Royal Society open science*, 3(4):150688.

[40] Dang, T. A., Deepak, J., Wang, J., Luo, S., Jin, Y., Ng, Y., Lim, A., and Li, Y. (2017). Mobility genome™-a framework for mobility intelligence from large-scale spatio-temporal data. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 449–458. IEEE.

[41] de Montjoye, Y.-A., Smoreda, Z., Trinquart, R., Ziemlicki, C., and Blondel, V. D. (2014). D4d-senegal: the second mobile phone data for development challenge. *arXiv preprint arXiv:1407.4885*.

[42] Dean, J. and Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113.

[43] D'Silva, K., Noulas, A., Musolesi, M., Mascolo, C., and Sklar, M. (2017). If i build it, will they come?: Predicting new venue visitation patterns through mobility data. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 54. ACM.

[44] e Silva, F. B. and Poelman, H. (2016). Mapping population density in functional urban areas. Technical report.

[45] Eldawy, A. and Mokbel, M. F. (2015). Spatialhadoop: A mapreduce framework for spatial data. In *2015 IEEE 31st international conference on Data Engineering*, pages 1352–1363. IEEE.

[46] Fortunato, S. (2010a). Community detection in graphs. *Physics Reports*, 483(3):75–174.

[47] Fortunato, S. (2010b). Community detection in graphs. *Physics reports*, 486(3-5):75–174.

[48] Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics reports*, 659:1–44.

[49] Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569.

[50] Furno, A., El Faouzi, N.-E., Fiore, M., and Stanica, R. (2017). Fusing gps probe and mobile phone data for enhanced land-use detection. In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pages 693–698. IEEE.

[51] Furno, A., Fiore, M., Stanica, R., Ziemlicki, C., and Smoreda, Z. (2016). A tale of ten cities: Characterizing signatures of mobile traffic in urban areas. *IEEE Transactions on Mobile Computing*, 16(10):2682–2696.

[52] Garas, A., Schweitzer, F., and Havlin, S. (2012). Ak-shell decomposition method for weighted networks. *New Journal of Physics*, 14(8):083030.

[53] Ghazi, M. R. and Gangodkar, D. (2015). Hadoop, mapreduce and hdfs: a developers perspective. *Procedia Computer Science*, 48(C):45–50.

[54] Ghemawat, S., Gobioff, H., and Leung, S.-T. (2003). The google file system. In *Proceedings of the nineteenth ACM symposium on Operating systems principles*, pages 29–43.

[55] Girardin, F., Vaccari, A., Gerber, A., Biderman, A., and Ratti, C. (2009). Quantifying urban attractiveness from the distribution and density of digital footprints. *Int. J. Spatial Data Infrastructures Res.*, 4:175–200.

[56] Grauwin, S., Sobolevsky, S., Moritz, S., Gódor, I., and Ratti, C. (2015). Towards a comparative science of cities: Using mobile traffic records in new york, london, and hong kong. In *Computational approaches for urban environments*, pages 363–387. Springer.

[57] Gundlegård, D., Rydergren, C., Breyer, N., and Rajna, B. (2016). Travel demand estimation and network assignment based on cellular network data. *Computer Communications*, 95:29–42.

[58] Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., and Ratti, C. (2014). Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271.

[59] Hoerl, A. E. and Kennard, R. W. (2000). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86.

[60] Järv, O., Ahas, R., Saluveer, E., Derudder, B., and Witlox, F. (2012). Mobile phones in a traffic flow: a geographical perspective to evening rush hour traffic analysis using call detail records. *PloS one*, 7(11):1–12.

[61] Jiang, S., Fiore, G. A., Yang, Y., Ferreira Jr, J., Frazzoli, E., and González, M. C. (2013). A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing*, page 2. ACM.

[62] Jolliffe, I. T. (1986). Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer.

[63] Joseph, K., Tan, C. H., and Carley, K. M. (2012). Beyond "local", "categories" and "friends": clustering foursquare users with latent "topics". In *UbiComp*.

[64] Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., and Newth, D. (2015). Understanding human mobility from twitter. *PloS one*, 10(7):e0131469.

[65] Kandt, J. and Batty, M. (2021). Smart cities, big data and urban policy: Towards urban analytics for the long run. *Cities*, 109:102992.

[66] Karau, H., Konwinski, A., Wendell, P., and Zaharia, M. (2015). *Learning Spark: Lightning-Fast Big Data Analytics*. O'Reilly Media, Inc., 1st edition.

[67] Lam, C. (2011). Introducing hadoop. *Hadoop in Action, MANNING*.

[68] Land Copernicus (2012). Copernicus land monitoring service urban atlas. [Online; accessed 13-November-2019].

[69] Laurila, J. K., Gatica-Perez, D., Aad, I., Bornet, O., Do, T.-M.-T., Dousse, O., Eberle, J., Miettinen, M., et al. (2012). The mobile data challenge: Big data for mobile computing research. Technical report.

[70] Lee, Y. (2019). Modern technology and sustainable future. *International Journal of Applied Engineering Research*, 14(7):1647–1651.

[71] Lepri, B., Oliver, N., Letouzé, E., Pentland, A., and Vinck, P. (2017). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, pages 1–17.

[72] Lima, A., De Domenico, M., Pejovic, V., and Musolesi, M. (2015). Disease containment strategies based on mobility and information dissemination. *Scientific reports*, 5.

[73] Liu, Y., Fang, F., and Jing, Y. (2020). How urban land use influences commuting flows in wuhan, central china: A mobile phone signaling data perspective. *Sustainable Cities and Society*, 53:101914.

[74] Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P. (2013). Understanding variable importances in forests of randomized trees. In *Advances in neural information processing systems*, pages 431–439.

[75] Lu, X., Wrathall, D. J., Sundsøy, P. R., Nadiruzzaman, M., Wetter, E., Iqbal, A., Qureshi, T., Tatem, A. J., Canright, G. S., Engø-Monsen, K., and Bengtsson, L. (2016). Detecting climate adaptation with mobile network data in bangladesh: anomalies in communication, mobility and consumption patterns during cyclone mahasen. *Climatic Change*, 138(3-4):505–519.

[76] Malliaros, F. D., Giatsidis, C., Papadopoulos, A. N., and Vazirgiannis, M. (2019). The core decomposition of networks: theory, algorithms and applications. *The VLDB Journal*, pages 1 – 32.

[77] Mann, A. (2016). Core concepts: Computational social science. *Proceedings of the National Academy of Sciences of the United States of America*, 113 3:468–70.

[78] Naboulsi, D., Fiore, M., Ribot, S., and Stanica, R. (2015). Large-scale mobile traffic analysis: a survey. *IEEE Communications Surveys & Tutorials*, 18(1):124–161.

[79] Newman, M. E. and Girvan, M. (2004a). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

[80] Newman, M. E. J. and Girvan, M. (2004b). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113.

[81] Noulas, A., Scellato, S., Lathia, N., and Mascolo, C. (2012). Mining user mobility features for next place prediction in location-based services. In *2012 IEEE 12th international conference on data mining*, pages 1038–1043. IEEE.

[82] Novović, O., Brdar, S., and Crnojević, V. (2015). Evolving connectivity graphs in mobile phone data. In *NetMob, The main conference on the scientific analysis of mobile phone datasets*, pages 73–75. Vodafone.

[83] Novović, O., Brdar, S., Mesaroš, M., Crnojević, V., and N Papadopoulos, A. (2020). Uncovering the relationship between human connectivity dynamics and land use. *ISPRS International Journal of Geo-Information*, 9(3):140.

[84] Novović, O., Grujić, N., Brdar, S., Govedarica, M., and Crnojević, V. (2020). Clustering foursquare mobility networks to explore urban spaces. In Rocha, Á., Adeli, H., Reis, L. P., Costanzo, S., Orovic, I., and Moreira, F., editors, *Trends and Innovations in Information Systems and Technologies*, pages 544–553, Cham. Springer International Publishing.

[85] Noyman, A., Doorley, R., Xiong, Z., Alonso, L., Grignard, A., and Larson, K. (2019). Reversed urbanism: Inferring urban performance through behavioral patterns in temporal telecom data. *Environment and Planning B: Urban Analytics and City Science*, 46(8):1480–1498.

[86] Oldham, S., Fulcher, B., Parkes, L., Arnatkeviciute, A., Suo, C., and Fornito, A. (2018). Consistency and differences between centrality measures across distinct classes of networks.

[87] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

[88] Pang, J. and Zhang, Y. (2017). Quantifying location sociality. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 145–154. ACM.

[89] Pappalardo, L., Pedreschi, D., Smoreda, Z., and Giannotti, F. (2015). Using big data to study the link between human mobility and socio-economic development. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 871–878.

[90] Pastor-Escuredo, D., Morales-Guzmán, A., Torres-Fernández, Y., Bauer, J.-M., Wadhwa, A., Castro-Correa, C., Romanoff, L., Lee, J. G., Rutherford, A., Frias-Martinez, V., Oliver, N., Frias-Martinez, E., and Luengo-Oroz, M. (2014). Flooding through the lens of mobile phone activity. In *Global Humanitarian Technology Conference (GHTC), 2014 IEEE*, pages 279–286. IEEE.

[91] Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T., and Zhou, C. (2014). A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, 28(9):1988–2007.

[92] Ponting, C. (1991). A new green history of the world: The environment and the collapse of great civilizations.

[93] Preoţiuc-Pietro, D. and Cohn, T. (2013). Mining user behaviours: A study of check-in patterns in location based social networks. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 306–315, New York, NY, USA. ACM.

[94] Qin, S., Man, J., Wang, X., Li, C., Dong, H., and Ge, X. (2019). Applying big data analytics to monitor tourist flow for the scenic area operation management. *Discrete Dynamics in Nature and Society*.

[95] Rahman, M. S. et al. (2017). *Basic graph theory*, volume 9. Springer.

[96] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

[97] Ratti, C., Frenchman, D., Pulselli, R. M., and Williams, S. (2006). Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748.

[98] Rendón, E., Abundez, I. M., Gutierrez, C., Zagal, S. D., Arizmendi, A., Quiroz, E. M., and Arzate, H. E. (2011). A comparison of internal and external cluster validation indexes. In *Proceedings of the 2011 American Conference, San Francisco, CA, USA*, volume 29, pages 1–10.

[99] Ríos, S. A. and Muñoz, R. (2017). Land use detection with cell phone data using topic models: Case santiago, chile. *Computers, Environment and Urban Systems*, 61:39–48.

[100] Ritchie, H. and Roser, M. (2018). Urbanization. *Our World in Data*. https://ourworldindata.org/urbanization.

[101] Rowe, F. (2021). Big data and human geography.

[102] Salah, A. A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.-A., Dong, X., and Dagdelen, O. (2018). Data for refugees: the d4r challenge on mobility of syrian refugees in turkey. *arXiv preprint arXiv:1807.00523*.

[103] Salloum, S., Dautov, R., Chen, X., Peng, P. X., and Huang, J. Z. (2016). Big data analytics on apache spark. *International Journal of Data Science and Analytics*, 1(3):145–164.

[104] Segal, M. R. (2004). Machine learning benchmarks and random forest regression.

[105] Serrano, M. Á., Boguná, M., and Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences*, 106(16):6483–6488.

[106] Shangguan, B., Yue, P., Wu, Z., and Jiang, L. (2017). Big spatial data processing with apache spark. *2017 6th International Conference on Agro-Geoinformatics*, pages 1–4.

[107] Shaw, S.-L. and Sui, D. (2018). *Human dynamics research in smart and connected communities*. Springer, New York, USA.

[108] Shaw, S.-L. and Sui, D. (2020). Understanding the new human dynamics in smart spaces and places: Toward a splatial framework. *Annals of the American Association of Geographers*, 110(2):339–348.

[109] Shvachko, K., Kuang, H., Radia, S., and Chansler, R. (2010). The hadoop distributed file system. In *Symposium on Mass Storage Systems and Technologies*, pages 1–10.

[110] Silva, T. H., Vaz de Melo, P. O., Almeida, J. M., Salles, J., and Loureiro, A. A. (2013). A comparison of foursquare and instagram to the study of city dynamics and urban social behavior. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, page 4. ACM.

[111] Singleton, A. and Arribas-Bel, D. (2021). Geographic data science. *Geographical Analysis*, 53(1):61–75.

[112] Soto, V. and Frias-Martinez, E. (2011). Robust land use characterization of urban landscapes using cell phone data. In *Proceedings of the 1st Workshop on Pervasive Urban Applications, in conjunction with 9th Int. Conf. Pervasive Computing*, volume 9.

[113] Steele, J. E., Pezzulo, C., Albert, M., Brooks, C. J., zu Erbach-Schoenberg, E., O'Connor, S. B., Sundsøy, P. R., Engø-Monsen, K., Nilsen, K., Graupe, B., et al. (2021). Mobility and phone call behavior explain patterns in poverty at high-resolution across multiple settings. *Humanities and Social Sciences Communications*, 8(1):1–12.

[114] Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.-A., Iqbal, A. M., Hadiuzzaman, K. N., Lu, X., Wetter, E., Tatem, A. J., and Bengtsson, L. (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127).

[115] Stoica, I. (2014 (accessed December 7, 2016)). Apache spark and hadoop: Working together. https://www.databricks.com/blog/2014/01/21/spark-and-hadoop.html.

[116] Taylor, R. C. (2010). An overview of the hadoop/mapreduce/hbase framework and its current applications in bioinformatics. *BMC bioinformatics*, 11(12):1–6.

[117] Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., Liu, H., Wyckoff, P., and Murthy, R. (2009). Hive: A warehousing solution over a map-reduce framework. *Proceedings VLDB Endowment*, 2(2):1626–1629.

[118] Truică, C.-O., Novović, O., Brdar, S., and Papadopoulos, A. N. (2018a). Community detection in who-calls-whom social networks. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 19–33. Springer.

[119] Truică, C.-O., Novović, O., Brdar, S., and Papadopoulos, A. N. (2018b). Community detection in who-calls-whom social networks. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 19–33. Springer.

[120] Truică, C.-O., Rădulescu, F., and Boicea, A. (2016). Comparing different term weighting schemas for topic modeling. In *International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC2016)*, pages 307–310.

[121] Valente, T. W., Coronges, K., Lakon, C., and Costenbader, E. (2008). How correlated are network centrality measures? *Connections (Toronto, Ont.)*, 28(1):16.

[122] Vavilapalli, V. K., Murthy, A. C., Douglas, C., Agarwal, S., Konar, M., Evans, R., Graves, T., Lowe, J., Shah, H., Seth, S., Saha, B., Curino, C., O'Malley, O., Radia, S., Reed, B., and Baldeschwieler, E. (2013). Apache hadoop yarn: Yet another resource negotiator. In *Annual Symposium on Cloud Computing*, pages 5:1–5:16.

[123] Wagner, S. and Wagner, D. (2007). Comparing clusterings: an overview. Technical report.

[124] Wampler D., P. A. (2014). *Programming Scala*. O'Reilly Media, Inc.

[125] Weisberg, S. (2005). *Applied linear regression*, volume 528. John Wiley & Sons.

[126] Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., Engø-Monsen, K., and Buckee, C. O. (2015). Impact of human mobility on the emergence of dengue epidemics in pakistan. *Proceedings of the National Academy of Sciences*, 112(38):11887–11892.

[127] Wilson, R., zu Erbach-Schoenberg, E., Albert, M., Power, D., Tudge, S., Gonzalez, M., Guthrie, S., Chamberlain, H., Brooks, C., Hughes, C., Pitonakova, L., Buckee, C., Lu, X., Wetter, E., Tatem, A., and Bengtsson, L. (2016). Rapid and near real-time assessments of population displacement using mobile phone data following disasters: the 2015 nepal earthquake. *PLoS Currents*, 8.

[128] Witayangkurn, A., Arai, A., and Shibasaki, R. (2022). Development of big data-analysis pipeline for mobile phone data with mobipack and spatial enhancement. *ISPRS International Journal of Geo-Information*, 11(3).

[129] Wright, R. E. (1995). Logistic regression.

[130] Xin, R. S., Gonzalez, J. E., Franklin, M. J., and Stoica, I. (2013). Graphx: A resilient distributed graph system on spark. New York, NY, USA. Association for Computing Machinery.

[131] Xing, W. and Ghorbani, A. (2004). Weighted pagerank algorithm. In *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, pages 305–314. IEEE.

[132] Yang, C.-T., Chen, S.-T., and Yan, Y.-Z. (2017). The implementation of a cloud city traffic state assessment system using a novel big data architecture. *Cluster Computing*, 20(2):1101–1121.

[133] Yang, J., Dash, M., and Teo, S. G. (2021). Pptpf: Privacy-preserving trajectory publication framework for cdr mobile trajectories. *ISPRS International Journal of Geo-Information*, 10(4).

[134] Yang, L. and Durarte, C. M. (2019). Identifying tourist-functional relations of urban places through foursquare from barcelona. *GeoJournal*.

[135] Yu, J., Wu, J., and Sarwat, M. (2015). Geospark: A cluster computing framework for processing large-scale spatial data. In *Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems*, pages 1–4.

[136] Zagheni, E., Garimella, V. R. K., Weber, I., and State, B. (2014). Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 439–444.

[137] Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauly, M., Franklin, M. J., Shenker, S., and Stoica, I. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Presented as part of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, pages 15–28. USENIX.

[138] Zhang, Z., Zhou, L., Zhao, X., Wang, G., Su, Y., Metzger, M., Zheng, H., and Zhao, B. Y. (2013). On the validity of geosocial mobility traces. In *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*, page 11. ACM.

[139] Zhao, Y. and Karypis, G. (2002). Criterion functions for document clustering: Experiments and analysis. Technical report.

[140] Zhao, Z.-D. and Shang, M.-S. (2010). User-based collaborative-filtering recommendation algorithms on hadoop. In *2010 third international conference on knowledge discovery and data mining*, pages 478–481. IEEE.

Appendix A

Biography

Olivera Mulić (maiden name Novović) was born on 19.05.1991. in Novi Sad, Serbia. She finished elementary school "Đorđe Natošević" in Novi Sad in 2006. After that she started high school education in "Gimnazija Jovan Jovanović Zmaj" in special class for highly gifted students in mathematics which shaped her interests in engineering and technology. She finished high school and started University studies in 2010. She enrolled at academic studies at the Faculty of Technical Sciences, University of Novi Sad in the program Geodesy and Geomatics, and graduated with Honours in 2014 with GPA of 9.47. The same year, she started her Master studies at the same University and program and earned the MSc degree in Geodesy, in 2015 with GPA 9.11. In 2016 she started working as a Research Assistant at BioSense Institute in Novi Sad, where she first started her scientific and research work. Later in 2016 she enrolled at PhD studies at the Faculty of Sciences in the program Geosciences. During her academic career she engaged in many fruitful collaborations with foreign Universities, among which the most successful ones are with the Aristotle University of Thessaloniki, Greece and Strathclyde University of Glasgow, UK. Collaboration with Aristotle University of Thessaloniki, specially with Data and Engineering Laboratory at the Department of Informatics (DELAB) produced two of her major publications: 1) "Community detection in who-calls-whom social networks" [118] and 2)"Uncovering the relationship between human connectivity dynamics and land use" [83]. At the time she runs her own consultancy company and work on more commercial projects in the area of Machine Learning and Data Science, and explores her research interests in industry.

Appendix B

Extended abstract in Serbian language

Appendix B contains the extended abstract of this thesis written in Serbian language using Cyrillic letters. We briefly summarized the subject and aim of this research with the overview of applied Graph theory in human dynamics research and Big Data analysis. In the end of the summary, main results are presented followed by the conclusion. The extended abstract is divided into five sections each of which describes one aspect of the thesis:

- A.1 - describes subject and scope of the research,
- A.2 - describes fundamentals of Network Science and Graph Theory applied in Human Dynamics research,
- A.3 - describes Big Data analysis and technologies that are used to investigate very large and complex data sets, with the focus on methodologies applied in this research,
- A.4 - describes main results of this research,
- A.5 - describes conclusion of this research and thesis.

With extended abstract written in Serbian language we aim to enlarge the bibliography of scientific notations in Serbian language and to support open science in Serbia.

Appendix B

Продужени апстракт на српском језику

Анализа великих података је нова област у компјутерским наукама која је настала као потреба да се искористе и обраде велике количине података који су по структури хетерогени, динамични, а могу бити и просторно и временски зависни. Наука о подацима и машинско учење има кључну улогу у анализи оваквих података јер омогућава да се подаци групишу, да се релације повежу и да се из њих извуче знање које би другачије остало скривено. Са развојем интернет технологија, ГПС-а и паметних телефона, масивне количине података о људском понашању, кретању и активностима су постале доступне и омогућавају истраживачима да дају нови увид у људско понашање. Динамика људског понашања је просторно и временски условљена и нови типови података као и нове аналитичке могућности нам дају прилику да боље разумемо интеракције које постоје између простора и људи, а које је раније било немогуће сагледати.

А.1 Предмет и циљеви истраживања

Предмет и циљеви истраживања докторске дисертације су различити облици примене аналитике великих података у истраживањима динамике људске популације која је просторно и временски условљена. Када говоримо о динамици људске популације, она се може посматрати са више аспеката, који су систематизовани као **активност, конективност** и **мобилност.** Сваки од ових аспеката динамике људске популације је временски и просторно одређен. Тако активност представља „тачкасто“ понашање, односно одређену акцију која је просторно дефинисана тачком и временски одређена тренутком или суму акција које су просторно дефинисане једном тачком а временски интервалом. Конективност представља акцију повезивања која је просторно дефинисана са две тачке а временски тренутком реализације конекције или интервалом. Мобилност представља акцију кретања која је просторно дефинисана са две или више тачака а временски тренутком реализације или интервалом. Важно је да се напомене да активност и конективност могу да се реализују без физичког кретања, коришћењем телекомуникација и интернета, док мобилност подразумева и физичко кретање.

У оквиру ове докторске тезе истраживали смо сва три аспекта динамике људске популације кроз отворене податке телекомуникационог оператера Телеком Италија и податке са друштвене мреже *„Foursquare“*. Користили смо модерне технологије Аналитике великих података да бисмо ефикасно извукли скривене обрасце понашања из података. Наших пет главних хипотеза су:

1. Обрасци у људској динамици и понашању су у снажној корелацији са временом у току дана и типом дана, где се разликују обрасци из сата у сат током дана, али и општи обрасци током радних и нерадних дана.
2. Обрасци у људској динамици су географски дефинисани и груписани унутар просторних јединице са сличном семантиком локације.
3. Људска динамика и интеракције, представљене кроз конективност и мобилност мреже, могу се даље анализирати помоћу теорије графова и кластеризације за издвајања корисних својства (информација) на нивоу градова и његових целина.
4. Начин коришћења и функционалност локације снажно утичу на предвиђање својстава конективности и мобилности која су у корелацији са људском динамиком.
5. Технологије аналитике великих података могу се користити за дизајнирање и развој ефикасних токова обраде података.

Главне хипотезе у оквиру ове дисертације смо дефинисали посматрањем међусобних веза између људске динамике и просторне семантике у урбаним срединама кроз анализу података, затим смо проширили и надоградили знање и закључке других аутора из ове области и учврстили смо своје тврдње експерименталним резултатима.

Други аутори су се такође бавили истраживањем динамике људске популације, посматрајући ту динамику кроз различите аспекте. Однос директне корелације између људске активности у урбаним срединама и времена у току дана уочен је и у ранијим истраживањима приликом анализе података о мобилности саобраћаја [59]. *Cottineau* и др. [36] и *Calabrese* и др. [28] су истраживали како би се подаци мобилних телефона могли користити за процену друштвено-економске организације градова и као индикатор за урбано опажање (енг. *„urban sensing“).* Брдар и др. [23] су истраживали телекомуникационе податке о конективности и мобилности на подручју целе земље да би извукли скривене обрасце из података користећи теорију графова. Отварање новог, веома детаљног и богатог сета података о конективности и активности од телекомуникационог оператера Телеком Италија [13], подстакло нас је да се упустимо дубље у област аналитике мобилних података са циљем да откријемо ново знање и развијемо нове методе за анализу динамике људске популације.

У оквиру ове дисертације посматрали смо обрасце у динамици људске популације зависно од доба дана, типа дана, сати са појачаном активношћу, радних и нерадних сати у току дана, ноћних сати, викенда и празника. Просторне локације са различитом семантиком су повезиване са различитим облицима људском понашања и динамике и остављале су различите „дигиталне отиске“ у подацима. Како се људска динамика развија кроз време, мрежа конективности и мобилности се такође развија па се и својства мреже могу користити за процену људске динамике. Слободно можемо рећи да се мрежа конективности и мобилности може користити као прокси за квантификацију динамике људске популације.

Извори и опис података

У току истраживања које је представљено у оквиру ове дисертације кориштени су подаци из три извора:

1. Мобилни подаци од телекомуникационог оператера Телеком Италија
2. Подаци са друштвене мреже *„Foursquare“*
3. Географски отворени подаци о употреби земљишта и земљишном покривачу

Мобилни оператери прикупљају велике количине података о активности и интеракцијама корисника. Сваки пут када корисник користи мобилни телефон један *„CDR“* (енг. *„Call Detail Record“*) запис се креира у бази података оператера. Такви записи о активности су примарно креирани у сврху наплате услуге оператера, али њихова вредност далеко превазилази иницијалну потребу јер представљају веома детаљан и континуиран извор информација о људском понашању. Кориснички генерисани подаци о употреби телекомуникационих услуга су власништво оператера али се на њих примењују веома строги прописи о заштити података о личности и заштити приватности корисника, због чека оператери нису вољни да деле те податке изван компаније, ни у сврху научног истраживања. Међутим, како је развој дигиталних технологија све више напредовао и све је већа потреба за новим прецизно дизајнираним производима и услугама оператери полако почињу да уносе флексибилност у своје процедуре и спремни су да деле податке под одређеним условима. Први од услова је да подаци буду анонимизовани и ослобођени сваке могућности да се кроз податке идентификује појединац. Други услов је да подаци буду просторно предефинисани, како би се сакрила права и прецизна локација базних станица. Први и други услов могу да се испуне кроз процес анонимизације и агрегације података који се врши код самог оператера пре него што се подаци објаве на интернету за слободан приступ. Један од начина да се постигне анонимизација јесте да се подаци агрегирају на ниво базне станице па се као резултат добија проток саобраћаја између две базне станице, а не између два корисника. Агрегација се потом понавља како би се сакрила права локација базне станице тако што се саобраћај дистрибуира на ниво претходно дефинисаних просторних јединица. На тај начин се на крају добија агрегирана конективност између просторних јединица у времену. Линкови који представљају ту конективност садрже и тежински коефицијент који је добијен агрегацијом стварног саобраћаја.

Друштвене мреже су постале незаобилазан елемент модерног свакодневног живота, већ дуго година уназад. Највећи део времена друштвене мреже се користе на мобилним уређајима путем апликација, и већина тих апликација захтева информацију о локацији са мобилног уређаја да би пружила, персонализовала и унапредила услуге које нуде. Последично, компаније које развијају друштвене мреже су постале кључни фактор у пружању услуга базираним на локацији и при томе прикупљају огромне количине података о локацији, мобилности, навикама и понашању корисника. Због заштите података о личности, као и заштите компаније од конкуренције, компаније друштвених мрежа строго чувају своје базе података и не дозвољавају приступ подацима. Ипак, у неким случајевима када препознају посебну корист, макар и само маркетиншког карактера компаније друштвених мрежа отварају ограничене сетове података за јавни приступ. Такву иницијативу су примениле компаније *„Twitter“* и *„Foursquare“* кроз дељење *API*-ја отвореног кода. Главне предности у коришћењу података са друштвених мрежа и интернет базираних апликација су:

1. нема бојазни везано за приватност јер су корисници или дали сагласност за коришћење података пре употребе апликације или користе апликацију у јавном режиму рада и
2. подаци имају покривеност целог света, мада зависно од популарности саме апликације у одређеним крајевима.

Упркос великим предностима за употребу података са друштвених мрежа, постоје и неки недостаци. Први је тај што су подаци у власништву компаније и чак и када отворе податке за јавни приступ они су углавном доста ограничени, јер компаније данас тргују међусобно подацима и отварање већег скупа података би умањило профит. Друго, подаци су углавном више померени према млађој популацији која користи друштвене мреже и далеко је активнија у њиховој употреби од старије популације. С обзиром на убрзан раст тржишта за локацијски базиране друштвене мреже може се очекивати да ће у будућности ово постати још релевантнији извор података о људској динамици.

Подаци о географском простору су увек били изазовни за прикупљање и визуализацију. Људи су одувек имали потребу да упознају и разумеју географски простор око себе зато што су топологија, клима и природни ресурси имали велик значај за опстанак живог света, укључујући и људе. У модерно време када већина људске популације живи у великим урбаним срединама постоји још већа потреба за ефикасним управљањем простором и ресурсима. Даљинска детекција се развила у једну од кључних дисциплина у геонаукама, са фокусом на посматрање површине Земље из даљине, најчешће са сателита или из авиона. Даљинска детекција представља скуп метода и технологија за детекцију и надзор над физичким карактеристикама површине, тако што мери емитован и рефлектован сигнал из даљине. Посебни сензори прикупљају слике снимљене даљинском детекцијом које се потом развијају у различите производе. Данас постоји много сателитских оператера који нуде своје производе бесплатно или уз накнаду. Европска свемирска агенција је развила нову групу сателита који се зову *„Sentinel“* специјално за оперативне потребе програма *„Copernicus“*. *„Sentinel“* сателити су опремљени различитим оптичким и радарским сензорима за надгледање временских и атмосферских услова, океана и обалских подручја и да обезбеђују дневне и ноћне снимке Земљине површине у високој резолуцији. *„Copernicus“* сервиси су подељени у шест главних категорија: управљање земљиштем, мора и океани, атмосфера, реаговање у хитним ситуацијама, безбедност и климатске промене. У оквиру ове тезе кориштени су подаци о управљању земљиштем и земљишном покривачу у векторском облику. Поред података који изворно потичу из сателитских снимака, постоје и глобалне иницијативе да се прикупљају подаци о простору мапирањем. Једна таква иницијатива је *„OpenStreetMap“* која данас пружа веома богат векторски сет података о простору и семантици локације поготово у урбаним срединама, као и базне мапе за ГИС за разне апликације. У оквиру ове тезе кориштене су базне мапе *„OpenStreetMap“*.

A.2 Теорија графова у истраживању динамике људске популације

Наука о мрежама је релативно нов правац у академским истраживањима комплексних мрежа као што су телекомуникационе мреже, компјутерске мреже, интернет мреже, биолошке мреже, друштвене мреже итд. Ова област обухвата мултидисциплинарно знање и ослања се на теоријске концепте и методе које укључују теорију графова из математике, статистичку механику из физике, *„Data Mining"* из компјутерских наука, инференцијална статистика, друштвене структуре из социологије, итд. Сваки комплексни релациони систем може бити представљен у облику мреже, што отвара нове потенцијалне могућности за проучавање структура и односа унутар система на потпуно другачији начин користећи методе из науке о мрежама.

Изучавање комплексних система је тесно повезано са теоријом графова. Могло би се рећи да теорија графова стоји у основи комплексних система и науке о мрежама. Мрежа је у форми математичке структуре исто што и граф. Графови, односно мреже су састављени од чворова и грана којим су повезани. Ако људе посматрамо као чворове, а конекције које они праве међу собом путем телекомуникација и интернета као гране, имамо динамички граф људске популације. Ако локације у граду, или базне станице посматрамо као чворове, а мобилност између локација или агрегиран саобраћај између базних станица као гране, опет имамо динамички граф људски интеракција и кретања. На тај начин кроз теорију графова можемо да моделујемо динамику људске популације и да изводимо закључке и предикције на основу тих модела. Пошто је теорија графова формално математички добро дефинисана и постоје бројни алгоритми и методе за анализу графова они могу ефикасно да се користе за анализу комплексних система као што су друштвене мреже, транспорт, авио саобраћај, биоинформатика, динамика људске популације и др.

Графови имају формално дефинисана својства која пружају додатне информације о самој структури графа и његовим особинама. За изучавање графова и откривање знања које постоји скривено у односима ентитета, одн. чворова унутар графа можемо истражити глобалне и локална својства графа. Глобална својства пружају информацију о глобалној структури целог графа и омогућавају квантитативно и квалитативно поређење између графова. Неки од често коришћених глобалних својстава графа су број чворова и грана (линкова), максимални тежински коефицијент, радијус, диаметар и средња вредност коефицијента кластеризације. За разлику од глобалних својстава који пружају једну вредност за цео граф, локална својства се користе да би се открили локализовани обрасци унутар самог графа. Локална својства графа се рачунају по чвору и они откривају структуру графа у непосредној околини тог чвора. Неки од најчешће коришћених локалних својстава су коефицијент кластеризације, степен чвора и средишњост. У оквиру ове тезе користили смо четири мере средишњости да опишемо локалну структуру графова, то су:

- Степен средишњости (енг. *„Degree centrality”*): број грана којим је чвор повезан са другим чворовима графа,
- Средишњост по блискости (енг. *„Betweennees centrality”*): број најкраћих путања које пролазе кроз дати чвор,
- *Page Rank*: вероватноћа да ће слободно изабрана путања проћи кроз дати чвор,
- Кор број (енг. *„Core number”*): вредност којом се одређује колико чвор добро повезан са непосредном околином у графу.

Важно је напоменути да су Средишњост по блискости и *Page Rank* изузетно захтевни према рачунарским ресурсима, док Степен средишњости и Кор број имају линеарно време израчунавања.

Некада је због величине графа и комплексности грана неопходно да се граф редукује ради унапређења перформанси. Редуковање графа није тривијалан задатак, поготово када је у питању тежински граф, јер неке везе могу да имају мали значај у глобалном смислу за цео граф али могу да имају јако велик значај у непосредној околини чвора. Зато је важно пажљиво одабрати методу за филтрирање графа да би се очувале релевантне структуре и након редукције. У оквиру истраживања које је описано у овој тези користили смо статистичку методу под називом Филтрирање диспаритета за редукцију графа са циљем очувања локалне околине чворова. Филтрирање диспаритета је метода за екстракцију релевантних грана у тежинским графовима, која ради тако што одбацује гране које не носе статистички релевантне информације. Једначина 1 се користи за израчунавање значајности гране α_{ij} где α одређује границу значајности, p_{ij} је вероватноћа постојања гране између чворова *i,j* и *n* је број чворова у графу. Филтрирање се реализује тако што се задржавају све гране где је испуњен услов $\alpha_{ij} \leq \alpha$, а одбацују оне гране где је $\alpha_{ij} > \alpha$. Тиме што мењамо ниво значајности α можемо да филтрирано граф прогресивније и да се фокусирамо на значајније гране. Оваква стратегија филтрирања је оправдана сваки пут када имамо граф који је хетероген и где су тежине локално корелисане.

$$\alpha_{ij} = 1 - (n-1)\int_0^{pij}(1-x)^{n-2}dx < \alpha \qquad (1)$$

Кластеризација графа је још једна метода за редукцију величине графа али такође представља и битно својство графа и открива пуно о његовој структури и начину преношења информација кроз граф. Кластеризација графа је процедура којом се чворови по одређеном критеријуму групишу тако да нова група представља супер чвор у новом редукованом графу. Кластери, који се често зову и заједнице када граф преставља неку врсту друштвене мреже, представљају групу чворова који деле заједничке особине и својства. Још једна карактеристика кластера је да су чворови унутар кластера много јаче повезани међу собом него што су повезани са осталим чворовима других кластера. Кластеризација графа и поред тога што је комплексан и некоректно дефинисан проблем, привукао је велику пажњу последњих година унутар научне заједнице и постао једна од водећих тема у оквиру науке о мрежама. Данас постоје бројни алгоритми за кластеризацију графова, који су прилагођени посебним структурама графа. У оквиру истраживања описаног у овој тези, коришћен је *Louvain* алгоритам за детекцију заједница одн. кластеризацију.

Louvain алгоритам представља ефикасно решење за детекцију заједница у великим, комплексним графовима, базиран је на максимизацији функције модуларности. Модуларност партиције графа је скаларана вредност између -1 и 1, и представља меру густине грана унутар заједнице у поређењу са густином грана између заједница. У случају тежинских графова, модуларност је дефинисана формулом 2, где A_{ij} представља тежину грана између чворова i и j, а $k_i = \sum_j A_{ij}$ је сума свим тежина грана према чвору i , c_i је кластер коме је чвор i додељен, Кронекерова делта је 1 ако је $u = v$ и 0 ако није и $m = \frac{1}{2}\sum_{ij} A_{ij}$. Коначна форма функције модуларности која је коришћена у *Louvain* алгоритму је представљена једначином 3.

$$Q = \frac{1}{2m}\sum_{i,j}(A_{i,j} - \frac{k_i k_j}{2m})\delta(c_i, c_j) \qquad (2)$$

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{i,j} - \frac{\sum_j A_{i,j} \sum_i A_{j,i}}{2m}\right] \qquad (3)$$

Нажалост евалуације кластер структуре базирана на максимизацији модуларности је НП тежак проблем. Да би се обезбедила ефикасност процесирања *Louvain* алгоритам користи итеративни процес који укључује редукцију графа сваки пут када модуларност конвергира. Алгоритам почиње тако што се сваком чвору додели његов сопствени кластер, затим се чворови померају између суседних кластера, рачуна се модуларност, и када модуларност достигне локални максимум и ново померање чворова не условљава промену модуларности примењује се процес редукције где сваки кластер постаје нови супер чвор у новом графу. Потом се исти процес понавља над новим графом и тако се кроз пар итерација долази до структуре која максимизује модуларност и ако се настави процес након тога модуларност креће да опада. То је знак да треба да се стане са поделом и узима се структура где је модуларност максимална као коначни резултат кластеризације. Псеудо код описане методе *Louvain* алгоритма за кластеризацију је представљен у табели А.2.1.

```
Алгоритам Louvain кластеризација графа ($G(V, E)$)
Улаз: граф $G$
Излаз: кластери графа $G$

$n \leftarrow |V|$
крај $\leftarrow$ false
док није крај ради:
    додели сваком $u \in V$ сопствени кластер
    док има промене ради:
        за сваки $u \in V$ ради:
            $C \leftarrow$ кластер који максимизује модуларност

    ако новаМодуларност > стараМодуларност онда:
        $G \leftarrow$ редукован граф базиран на кластерима

    иначе:
        врати кластере
```

Табела А.2.1, *Louvain* алгоритам за кластеризацију

Теорија графова има значајну примену у науци о мрежама јер је математичка репрезентација мреже заправо граф. Све што се односи на графове, њихова својства, алгоритми, може да се примени на мреже и да се на тај начин извлачи скривено знање из мрежа и њихове структуре. Динамика људске популације се огледа кроз временску и просторну еволуцију мрежа које представљају људске интеракције. На тај начин, примењујући теорију графова на мреже које представљају људске интеракције ми можемо да квантификујемо динамику људске популације и да откријемо ново знање и законитости везано за људско понашање и простор.

А.3 Анализе великих података

У претходних пар година сведочили смо великим и веома динамичним променама у компјутерским наукама због непрестаног раста количине података који су доступни за аналитику у различитим сегментима и сталне потребе за бољим перформансама рачунарских система. Како би се испратили убрзани развој и нови захтеви многе апликације се ослањају на дистрибуиране системе. Систем је дистрибуиран ако се процесирање одвија на више различитих, физички одвојених машина и само се крајњи резултат враћа на главну машину, одн. кориснику. Јасно је да овакви системи захтевају стабилну интернет конекцију широког спектра да би функционисали, такође се могу појавити извесни застоји услед комуникације, али су свеукупно предности веће него недостаци.

Живимо у времену великих података где све око нас, уређаји, интернет системи, возила, машине итд. генеришу некакве податке који могу да се анализирају. Да би се анализирали и процесирали велики подаци, и да би се извукло релевантно знање из њих, постоји потреба за посебним апликацијама и додатним меморијским простором. Овакви системи имају наравно и посебне захтеве према толеранцији на отказ, паралелном процесирању, дистрибуцији података, балансирању терета, скалабилности итд. Да би се изборио са таквим захтевима *Google* је представио нови модел програмирања, ткз. *MapReduce*. *MapReduce* је програмски модел за дистрибуирано процесирање који омогуђује паралелну обраду и складишти податке у *HDFS (Hadoop Distributed File System)*. Овај програмски модел се састоји из две логичке јединице *Map* (мапирање) и *Reduce* (умањивање). Главна предност *MapReduce* модела је што омогућава паралелну дистрибуирану обраду података, али није најбоље решење када су комплексне операције у питању. *Apache Hadoop* је имплементација *MapReduce* модела отвореног кода. *Hadoop* је дизајниран да ефикасно процесира екстремно велике количине структуираних, неструктуираних и семи-структуираних података. *Hadoop* је писан у *Java* програмском језику и прилагођен је за дистрибуирано окружење. Суштина *Hadoop* система састоји се од слоја за складиштење, што је *HDFS* и програмског слоја који представља *MapReduce* модел. *Hadoop* је дизајниран да ефикасно управља апликацијама које користе велику количину података а пошто је базиран на отвореном коду могу да га користе разне индустрије као и Универзитети због чега је стекао велику популарност и примену.

Поред *Hadoop* система, *Apache* фондација је развила још једну платформу за дистрибуирано процесирање великих података то је *Apache Spark* . *Spark* је развијен како би надоместио недостатке које је *Hadoop* показао приликом комплексних операција. Такође, *Hadoop* није погодан за стриминг апликације, итеративне процесе и интерактивне апликације. Главни проблем који се показао је велик број У/И операција које захтева писање по диску. Са *Spark* платформом је тај проблем превазиђен јер *Spark* омогућава да се процесирање врши у меморији и да се само крајњи резултат уписује на диск. Осим тога, *Spark* нуди и апстракцију која је зове *RDD* (енг. *„RDD – Resilient Distributed Dataset“*) која ефикасно подржава апликације интензивне подацима, јер је *RDD* у ствари дистрибуирана колекција. *Spark* подржава четири програмска језика *Scala*, *Python*, *Java* и *R*, иако је изворно писан у *Scala*. *Spark* послови су базирани *„master-slave"* архитектури јер постоји један *driver* програм одн. машина, а други елементи кластера су извршиоци. *Spark* је надоградио *Hadoop* у оним сегментима где су постојали недостаци и *Spark* подржава велик број различитих апликација, могуће га је интегрисати са стриминг системима, подржава ефикасно извршавање комплексних операција и ефикасно скалира. За кратко време, *Spark* је постао готово подразумевани избор када су у питању технологије за процесирање великих података.

У оквиру истраживања динамике људске популације користили смо велике сетове података из више различитих извора. За ефикасну аналитику тих података морали смо да се окренемо новим технологијама специјализованим за такве задатке. За компјутерски најзахтевније задатке користили смо *Spark* са програмским језиком *Scala*. Тестови перформанси говоре у прилог томе да су ове технологије неизоставни део у истраживањима динамике људске популације јер су сви подаци који описују људе и њихово понашање, велики подаци за које је карактеристично три „V" – екстремни обим (енг. *„Volume"*), широк спектар (енг. *„Variety"*), брзина промене (енг. *„Velocity"*).

А.4 Резултати

У оквиру ове тезе посматрали смо динамику људске популације на иновативан начин, кроз призму науке о мрежама и теорију графова. Динамику људске популације смо моделовали кроз телеком податке и податке са друштвене мреже, тако што смо људске интеракције посматрали као графове. Графови које смо добили су веома комплексни за анализу јер су по структури веома густи (садрже велик број грана у односу на број чворова) и имају додељене коефицијенте тежине на свакој грани. Користили смо *Spark* за рачунарски захтевне експерименте и потврдили смо хипотезу да напредне технологије аналитике великих података морају да се користе да би процесирање било ефикасно.

У четвртом поглављу ове тезе представили смо како телеком мрежа конективности рефлектује начин на који људи остварују интеракције путем телекомуникационих сервиса и како су те интеракције мапиране у простору. Мрежу конективности смо посматрали као граф и применили смо теорију графова да бисмо открили скривене обрасце у понашању. Користили смо локална и глобална својства графа да бисмо ближе одредили динамику људске популације рефлектовану кроз конективност. Да бисмо боље разумели значај просторне локације у начину како се динамика формира и како еволуира, агрегирали смо телеком податке на ниво просторних јединица административне зоне у граду и на тај начин добили графове конективности између просторних јединица. Такође, посматрали смо како ти графови еволуирају кроз време и уочили смо јасне разлике у обрасцима понашања у зависности од доба дана и типа дана. Графови који се односе на викенде и празнике изгледају слично, док се веома разликују од оних који се односе на радне дане и по структури и по интензитету. На слици 4.2 приказани су примери мреже конективности у току четири различита дана, мапирани у простору. Да бисмо боље разумели структуру графа, а сам тим и динамику људске популације рефлектовану кроз телеком мрежу конективности, анализирали смо глобална и локална својства графа. Глобална својства графа које смо посматрали су број чворова и грана, максималани тежински коефицијент грана, радијус, дијаметар, максимална величина клика, просечан коефицијент кластеризације. Поредили смо глобална својства током различитих типова дана и открили смо да постоје значајне разлике између викенда и радних дана. Број грана је значајно већи током радних дана него током викенда што доприноси изгледу гушће структуре графа током радних дана. Диаметар такође потврђује ову разлику, јер је диаметар током радних дана мањи него током викенда што указује да је током викенда мрежа више разуђена односно да постоји већа дистанца у конективности него током радних дана. Локална својства графа откривају особине графа у непосредној близини сваког чвора. Да бисмо показали еволуцију локалних својстава графа изабрали смо три чвора, Зона 1, 71 и 23, слика 4.4. Промене кроз време својства средишњост по блискости су приказане на слици 4.5, а својства *Page Rank* на слици 4.6. Анализом ових својстава може лако да се уочи разлика у обрасцима током викенда и радних дана, али такође и неки необични догађаји. Мера средишњости по блискости за зону 23 показује необичан скок у вредности 15. децембра када се у тој зони града одвијао филмски фестивал. Локална својства графа указују на то како поједине зоне града „дишу“, одн. како просторни контекст неодвојиво утиче на динамику људске популације.

У петом поглављу ове тезе описали смо оглед детекције заједница односно кластеризације графа у телеком мрежама конективности. У оквиру огледа мерили смо перформансе извршавања *Louvain* алгоритма за кластеризацију, креирање графова из улазних података и филтрирање графа уз помоћ *Spark* платформе у два различита окружења – сервер са више језгара и кластер рачунара. *Spark* је иницијално дизајниран да ради на кластеру машина али је могуће постићи и привидну дистрибуираност тиме што ће се извршавање вршити на више језгара једне машине. Резултати огледа показују да је *Spark* показао углавном боље перформансе у кластер окружењу, мада разлике нису значајне. Тешко је прецизно одредити због чега је настала разлика у перформансама између ових система али се може претпоставити да је у питању разлика у хардверу и подешавањима која не могу бити идентична у оба окружења. Оно што може да се извуче као важан закључак јесте да је серверско окружење стабилније и мање дивергира између итерација док кластер окружење више дивергира али има просечно краће време извршавања (слика 5.4). Још једно важно опажање из описаних огледа је да филтрирање графа значајно утиче на побољшање перформанси у детекцији заједница. У табели 5.1 приказани су резултати кластеризације за три нивоа филтрирања и мерење перформанси. Може се закључити да филтрирање графа од 95% и 99% не утиче значајно на број формираних кластера, а при томе значајно унапређује перформансе, док филтрирање на нивоу значајности 99.9% уноси веће промене у структуру кластера.

У шестом поглављу ове тезе представили смо резултате и закључке из опсежне студије о динамици људске популације квантификоване кроз телеком мреже конективности. Резултати ове студије објављени су у међународном научном часопису из области геонаука *„ISPRS International Journal of Geo-Information“*. Ова студија обједињује наш претходни рад где смо анализирали глобална и локална својства графа, утицај типа локације на динамику понашања и перформансе процесирања. Поред тога додатно смо прошили истраживање у оквиру ове студије тиме што смо показали експериментално како технике машинског учења могу да се користе за предвиђање људске динамике. Користили смо регресиони модел да предвиђамо својства графа на основу просторних особина локације које су дефинисане класама употребе земљишта. Кључна ствар за овакав приступ је моделовање података на начин да су чворови у графу који представља телекомуникациону конективност заправо просторне јединице у географском простору (у овом случају у граду Милано, Италија). Такође, кластери, одн. заједнице који су резултат методе детекције заједница представљају такође просторне јединице. Величина кластера у смислу броја чворова који му припадају диктира и величину просторне јединице коју представља кластер. Овде је величина просторне јединица заправо површина што је географски близак појам, та површина има своју класу из података о употреби земљишта. Тиме смо повезали динамику људске популације која се рефлектује кроз телекомуникационе интеракције, географски простор за који су те интеракције везане и теорију графова коју користимо као математички модел за машинско учење. На сликама 6.8 и 6.9 приказано је које класе употребе земљишта имају колики утицај на предвиђање својстава графа, одн. својстава кластера.

У седмом поглављу описали смо резултате везано за истраживање још једног важног аспекта људске динамике, а то је мобилност. Користили смо податке са друштвене мреже *„Foursquare“* за десет светских градова у периоду од две године. Овде смо кретање корисника између две локације посматрали као грану у графу, док су чворови графа локације које су посећене. Број мобилности између две локације представља тежински коефицијент гране. Поред дескриптивних анализа самих података, применили смо кластеризацију над графовима мобилности да бисмо видели које структуре у граду се групишу заједно и под којим условима. Резултате кластеризације смо анализирали тако што смо посматрали појединачне велике кластере унутар сваког града и гледали које семантичке класе се групишу заједно. Тако на пример, у Чикагу кластер који се формира око аеродрома има доминанту класу *„Travel and Transport“,* потом *„Food“* и *„Shop and Services“.* Ако размишљамо о људском понашању сасвим је логично да ове три класе буду тесно повезане. Такође је интересантна просторна дистрибуција тачака које припадају том кластеру, слика 7.8. Са слике се види како се тачке „расипају“ дуж главног пута *„Interstate 90“* до самог центра Чикага где се у мањој мери групишу поново. Додатно, анализирали смо сличност између градова користећи методу хијерархијске кластеризације. Тако амерички градови Чикаго и Њујорк личе међусобно и слични су као и Лос Анђелес, док су значајно другачији Лондона и Париза који су опет слични међусобно. Токио је негде између европских и америчких градова по сличности, док су Истанбул, Џакарта и Сингапур слични међусобно али различити од осталих градова. Сеул је другачији од свих осталих. Детаљи хијерархијске кластеризације су приказани на слици 7.9. Ово је веома значајно опажање за студије о људској динамици јер указује на то како културолошке и географске разлике диктирају људско понашање и навике. Такође, резултати ове студије указују на значај и велик потенцијал података са друштвених мрежа у истраживањима људске динамике.

А.5 Закључак

Како се глобалне позиционе технологије убрзано развијају и постају све лакше доступне људима широм света у њиховом свакодневном животу, људи на глобалном нивоу генеришу све више и више гео-референцираних података, сваког минута у току дана. Људи остварују међусобне интеракције без просторних баријера кроз виртуалне сервисе и телекомуникације остављајући при томе дигиталне трагове у подацима који могу много да нам открију о људском понашању, навикама, динамици, културолошким разликама, животном стилу итд. Граница између физичког и виртуелног простора постаје све „тања“ са напретком у технологијама које нам омогућавају да велик део наших уобичајених активности урадимо – *„online“*. Са растућим бројем активности које људи обављају *„online“* дефиниција динамике људске популације се мења и развија јер људи не морају да буду физички активни да би били динамични! Упркос томе, постоје извесне везе између физичког и виртуелног простора, исти обрасци понашања се појављују у оба простора. Добар пример овог нарочитог понашања је описан у резултатима поглавља 6 ове тезе.

Наиме, наши огледи су показали да за предикцију својства просечан *Page Rank*, класа која има највише утицаја је „Други путеви и придружено земљиште“, што можемо објаснити кроз концепт транзитивности. Концепт транзитивности у физичком, географском простору је везан за саобраћајну инфраструктуру. Транзитивност у виртуелном простору изграђеном од телекомуникационе мреже конективности представља својство *Page Rank*. Резултат који смо добили из огледа указује да се транзитивност из виртуелног простора може поистоветити са тразитивношћу из физичког простора и да су оба концепта условљена динамиком људске популације.

Још један пример из наших огледа указује на снажну повезаност физичког и виртуелног простора и људске динамике, описан је у поглављу 4 ове тезе. Када смо испитивали локална својства графа посматрали смо издвојене зоне града. Конкретно, примећен је нагли пораст у вредности својства међусобна централност за зону 23, дана 15. децембра. Такође, тада се дешавао велики филмски фестивал у тој зони града што је свакако условило повећан број људи који су дошли тај дан у ту зону града. Својство средишњост по блискости нам говори о томе колико најкраћих путања пролази преко датог чвора у графу, одн. ако је средишњост по блискости велика то значи да је чвор значајнији за проток информација у графу. Ако се велик број људи изненада појави на некој локацији, онда ће та локација постати значајнија у локалној околини. На тај начин, повезали смо концепт локалне значајности у физичком и у виртуелном свету кроз својство графа.

Подаци који су кориснички генерисани су велики по обиму и динамични јер се брзо мењају. Да бисмо извукли знање из таквих података потребно је да применимо напредне технологије за аналитику великих података као што је *Spark*, јер на тај начин можемо ефикасно доћи до рачунарских резултата. Такође, *Spark* ради у дистрибуираном окружењу, што је погодно за компаније јер на тај начин подаци не морају да „изађу“ из компаније да би били процесирани, довољно је да алгоритми приступе инфраструктури путем интернета.

У овој тези представљен је нови приступ у истраживању динамике људске популације који користи теорију графова за моделовање људских интеракција и напредне технологије аналитике великих података за ефикасно процесирање. Представили смо методологију како да се користи граф аналитика, машинско учење и аналитика података за откривање знања из различитих извора података као што су телеком подаци, подаци друштвених мрежа и просторни подаци. Резултате смо објаснили са аспекта семантике локације и мрежа које се развијају кроз време, које су кључни индикатори људске динамике.

Овај Образац чини саставни део докторске дисертације, односно докторског уметничког пројекта који се брани на Универзитету у Новом Саду. Попуњен Образац укоричити иза текста докторске дисертације, односно докторског уметничког пројекта.

План третмана података

Назив пројекта/истраживања
Примена Big Data аналитике за истраживање просторно-временске динамике људске популације (енг.) Big Data analysis applied in space-time human dynamics research
Назив институције/институција у оквиру којих се спроводи истраживање
а) Природно-математички факултет б) Институт БиоСенс
Назив програма у оквиру ког се реализује истраживање
1. Опис података
1.1 Врста студије У овој студији нису прикупљани подаци.
2. Прикупљање података
3. Третман података и пратећа документација
4. Безбедност података и заштита поверљивих информација
5. Доступност података
6. Улоге и одговорност