DISSERTATIONES
INFORMATICAE
UNIVERSITATIS
TARTUENSIS
**42**

RAHUL GOEL

# Mining Social Well-being Using Mobile Data

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

**42**

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
**42**

**RAHUL GOEL**

# Mining Social Well-being Using Mobile Data

UNIVERSITY OF TARTU
Press
1632

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor Philosophy (PhD) in Computer Science on May 29, 2023 by the Council of the Institute of Computer Science, University of Tartu.

*Supervisor*

Assoc. Prof. Rajesh Sharma
Institute of Computer Science
University of Tartu, Tartu, Estonia

*Opponents*

Prof. Nishanth Sastry
Director of Research
Department of Computer Science
University of Surrey, United Kingdom

Dr. Mirco Nanni
Researcher, Head of the KDD Laboratory
ISTI - CNR, Pisa

The public defense will take place on June 27, 2023 at 11:30 via Zoom.

The publication of this dissertation was financed by the Institute of Computer Science, University of Tartu.



ISSN 2613-5906 (print) ISSN 2806-2345 (PDF)
ISBN 978-9916-27-247-3 (print) ISBN 978-9916-27-248-0 (PDF)

University of Tartu Press
http://www.tyk.ee/

*To my grandparents*

# ABSTRACT

Mobile data, such as call data records (CDR), and digital data can provide lots of valuable information about people's behaviour. CDR data is generated when we make a phone call. On the other hand, the usage of mobile apps (such as Twitter, Facebook, and YouTube) generates digital data. In this thesis, we employ mobile phone data to contribute to three facets of social well-being. Here, social well-being is defined as making healthy interactions with friends and family, not experiencing segregation in society, having equal opportunities to earn, receiving proper education, and so on. All of these characteristics of social well-being are critical because humans generally seek an environment in which they can feel at ease, healthy, and happy.

In the first social well-being facet, we propose two versions of epidemic models called the mobility-based SIR model for (i) fully-mixed and (ii) for complex networks. Both these models take into account real-life interactions from CDR. This work is inspired by the hypothesis that the fundamental cause for epidemics turning into pandemics is population distribution, people's mobility, and social coherence across the globe. Using mathematical proof, we obtained that the reproduction number, which is defined as the average number of secondary infections produced when one infected individual is introduced into the population, directly depends upon *social connectivity*, *number of connected locations* and *individuals mobility* (and *degree of the individual* in complex networks model) which is in line with our simulation's results. This indicates that introducing *isolation* and *quarantine* is effective in fighting a pandemic crisis. Using the proposed models, we also simulated the COVID-19 cases in Estonia and Rhône-Alpes region in France. Simulation reveals that the mobility-based SIR model can be helpful to forecast the expected number of cases after some proportion of *isolation* and *quarantine* is introduced in society.

Second, we study societal segregation in Estonia using CDR data. Our findings suggest that (i) gender segregation exists in Estonia and its traces are visible in connectivity among age-groups, preferred language of communication, and in the county; (ii) The prime working individuals of age between 25 years and 54 years and elderly of age greater than 64 years are more segregated; (iii) Neighborhood and language play a vital role in segregation. Individuals prefer to connect with other individuals that speak the same language and reside in the same locality (city or county).

Third, we investigate digital traces from mobile apps (like Twitter and Facebook) to predict socio-economic conditions (SEC). These SEC include education, gender, poverty, employment, and other factors. In our work, we analyzed nationwide data from France and extracted three patterns: Typical week signature (TWS), Revealed Comparative Advantage (RCA), and Standardized Cumulative Utilization (SCU). We find that our proposed TWS has richer information diversification compared to RCA and SCU. Using the extracted patterns, our best model

is able to estimate economic, educational, and demographic indicators (attaining an R-squared score up to 0.66). This indicates that mobile app usage patterns can reveal socio-economic disparities.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ORIGINAL PUBLICATIONS

## Publications included in the thesis

I **Rahul Goel**, and Rajesh Sharma. "Mobility based sir model for pandemics-with case study of covid-19." In 2020 **IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)**, pp. 110-117. IEEE, 2020.

*Lead author. The author performed the implementation and the analysis of the experiments and contributed substantially to the ideas and the writing.*

II **Rahul Goel**, Loïc Bonnetain, Rajesh Sharma, and Angelo Furno. "Mobility-based SIR model for complex networks: with case study Of COVID-19." **Social Network Analysis and Mining** 11, no. 1 (2021): 1-18.

*Lead author. The author performed the implementation and the analysis of the experiments and contributed substantially to the ideas and the writing.*

III **Rahul Goel**, Rajesh Sharma, and Anto Aasa. "Understanding gender segregation through call data records: an Estonian case study." **Plos one** 16, no. 3 (2021): e0248212.

*Lead author. The author performed the implementation and the analysis of the experiments and contributed substantially to the ideas and the writing.*

IV **Rahul Goel**, Rajesh Sharma, and Anto Aasa. "Studying segregation in Estonia using call data records." **Social Network Analysis and Mining** 11, no. 1 (2021): 1-13.

*Lead author. The author performed the implementation and the analysis of the experiments and contributed substantially to the ideas and the writing.*

V **Rahul Goel**, Angelo Furno, and Rajesh Sharma. "Predicting Socio-Economic Well-being Using Mobile Apps Data: A Case Study of France." ArXiv preprint (arXiv:2301.09986). Under review at TKDD.

*Lead author. The author performed the implementation and the analysis of the experiments and contributed substantially to the ideas and the writing.*

## Publications not included in the thesis

I **Rahul Goel**, and Rajesh Sharma. "Understanding The MeToo Movement Through The Lens Of The Twitter." In **International Conference on Social Informatics**, pp. 67-80. Springer, Cham, 2020.

*Lead author. The author performed the implementation and the analysis of the experiments and contributed substantially to the ideas and the writing.*

II **Rahul Goel**, and Rajesh Sharma. "Studying leaders & their concerns using online social media during the times of crisis-A COVID case study." Social network analysis and mining 11, no. 1 (2021): 1-12.

*Lead author. The author performed the implementation and the analysis of the experiments and contributed substantially to the ideas and the writing.*

III Christian Ritter, **Rahul Goel**, Rajesh Sharma. "Decolonizing Newsmaking: The Case of Climate Change Communication on YouTube during the COP26 Summit." Association of Internet Researchers (2022).
*The author performed the implementation and the analysis of the experiments and contributed substantially in writing.*

# 1. INTRODUCTION

Humans often sought a state of Well-being, which is defined as being at ease, healthy, and happy. There are various dimensions of well-being. For example, physical well-being refers to keeping our bodies healthy, which can be accomplished through adequate exercise and proper nutrition. Similarly, optimism, self-acceptance, and healthy relationships are linked to emotional well-being. Furthermore, *social well-being*, which is at the heart of this thesis, is defined as the ability to interact successfully in local and global communities [Vou+20]. In today's world, humans can not only achieve social well-being by face-to-face interactions but also through the phone, or through social media platforms like Twitter, Facebook, and Instagram.

Mobile phones have become an inseparable part of our lives, and we mostly use them to connect with others not only by making calls but also over social media platforms. The usage of media applications through smartphones often leaves *digital data* (traces). Most people who spend a large chunk of time on mobile leave their unique usage signature with their distinctive patterns. These patterns are widely used by researchers to study well-being worldwide.

According to [dat22], there are 5.31 billion unique mobile phone users around the world in 2022, accounting for 67.1 percent of the global population. Of these mobile users, 4.95 billion people use the Internet – equal to 62.5 percent of the world's total population. As a result, two types of mobile data (i) *call data records (CDR)*, and (ii) *digital data* of mobile apps (such as Twitter, Facebook, and YouTube) generate massive volumes of data containing vital, albeit generally hidden, information about people's behaviour.

CDR in particular has been researched from a variety of perspectives over the last decade, including social network analysis [WF94], sociocultural characteristics of a city [Pon+16], defining the human mobility pattern [Son+10], the effect of different events on calling operations [Hii+19], and segregation [GSA21]. In this thesis, we focus on three dimensions of social well-being research using CDR and digital data. These dimensions are explained in brief in the following sections.

## 1.1. Epidemic Spreading from People's Mobility

The first dimension of work is modeling, in which we propose models that can help in understanding the spread of communicable diseases or epidemic. The connectivity among different regions of the world plays an important role in disease spread, which makes it easier to affect a wider geographical area, often worldwide making any epidemic a pandemic. However, proposing and evaluation of a model to understand the spread of an epidemic is challenging because people's mobility and social connectivity are non-uniform in the different regions of the world. In addition, these models are difficult to evaluate in the absence of a real data set that can capture mobility.

An example of an epidemic is the coronavirus (COVID-19), which started in December 2019 from Wuhan, China has infected 640M (million) individuals and claimed 6M (as of $10^{th}$ November 2022) lives worldwide [CSS20; COV19]. In our research, we extended existing epidemic models by introducing people's mobility and social connectivity. Our proposed method is known as the *Mobility-based SIR model*. We extended the *SIR* (Susceptible - Infected - Recovered) model as it is the most relevant model to understand the spread of coronavirus. In this model, there are three classes in which a person can reside: *Susceptible* which represents people who are not infected, *Infected* which represents infected people, and *Recovered* which represents recovered people from infection. Initially, the majority part of the population is susceptible and only a few people are infected. As time progresses, susceptible people become infected, and eventually infected people recover.

**Contribution.** In our proposed *Mobility-based SIR model*, we propose two versions, (i) fully-mixed model [KM27] and (ii) complex networks model [Pas+15], which takes into account people's mobility and social connectivity from CDR. This work is inspired by the idea that the main reason for some epidemics turning into pandemics is the connectivity among different regions of the world, which makes it easier to affect a wider geographical area, often worldwide. Also, the population distribution, people's mobility, and social coherence in the different regions of the world are non-uniform and play an important role. Our models are explained in detail in Chapter 2.

## 1.2. Societal Segregation

The second dimension of work is descriptive in nature, where we study societal segregation using CDR. Segregation is defined as the separation of people based on gender, language, or some other demographics. It has long been assumed to play a critical role in many developing countries' socio-economic structure and overall stability [BY08]. According to [Ale+03], ill effects of segregation are not limited to developing countries, but can have a detrimental impact in countries with poor political and legal structures. As a result, a great deal of emphasis is required on policies to facilitate integration and interaction, particularly in diverse societies.

Much of the previous research work on segregation relies on conventional government census data [Sil20]. However, census data can capture the precise pattern of the physical settlement but rarely record trends of social interaction, which are necessary to develop a thorough understanding of the essence of social interaction.

**Contribution.** In our research, we study *societal segregation* in Estonia based on four demographics (gender, age-group, language, and location) using CDR data. The dataset is provided by Estonia's major telecom operators to research the complexities of social interaction and human behavior. The primary contribution

of using CDR data to anticipate segregation is to provide an alternative to costly and time-consuming censuses. Refer to Chapter 3 for segregation research details.

## 1.3. Socio-Economic Well-being from Digital Data

The third dimension of work is predictive in nature, where we predict socio-economic conditions (SEC) of a region using mobile data. Alternative data sources such as demographic censuses or surveys which are being used for monitoring SEC have sparse population coverage or are updated infrequently due to their high costs and time-consuming process. As a result, they do not fit well with the evolution that societies are experiencing nowadays.

Researchers have used alternate digital traces, such as Mobile call data records and mobile application usage, for determining real-time SEC information [Uca+21; Zho+16; Nan+08]. Researchers have created prediction models using temporal and geographical information owing to the increasing usage of mobile devices, social media, and the expanding availability of ubiquitous satellite data [Sot+11; GZZ19; DRZ19; BDK15]. Human mobility and social contacts were shown to be linked to higher income in data from mobile phones and social media [Blu16; Pap+16; Ste+17; Llo+15]. Although these methodologies are pretty effective in forecasting SEC in underdeveloped countries, they are only somewhat accurate in developed countries with more nuanced variations in mobile phone adoption [Has+17; Jea+16; AK20].

**Contribution.** In our study, we show that mobile application usage patterns can be used for predicting SEC in a region. The primary goal of using mobile digital data to anticipate SEC is to provide an alternative mechanism to traditionally costly and time-consuming censuses. This leads to our research objective which is about *predicting the socio-economic conditions using digital traces*, which is explored in Chapter 4.

# 2. MOBILITY BASED SIR MODEL FOR PANDEMICS

In the last decade, humanity has faced many different pandemics such as SARS, H1N1, and presently novel coronavirus (COVID-19). On one side, scientists have developed vaccinations, and on the other side, there is a need to propose models that can help in understanding the spread of these pandemics as it can help governmental and other concerned agencies to be well prepared, especially for pandemics, which spreads faster like COVID-19. The main reason for some epidemics turning into pandemics is the connectivity among different regions of the world, which makes it easier to affect a wider geographical area, often worldwide. Also, the population distribution and social coherence in the different regions of the world are non-uniform. Thus, once the epidemic enters a region, then the local population distribution plays an important role.

**Research Question.** Inspired by the above-mentioned points, our research question is as follows: *Can we develop an epidemic model that takes into account people's mobility and real-life connections?* To answer, we extended the widely used epidemic model called SIR (Susceptible-Infected-Recovered) to the mobility-based SIR model. In particular, we propose two versions of our mobility-based SIR model, (i) fully-mixed and (ii) for complex networks, which especially take into account real-life interactions. To the best of our knowledge, this model is the first of its kind, which takes into account the population distribution, connectivity of different geographic locations across the globe, and individuals' network connectivity information. In addition to presenting the mathematical proof of our models, we have performed extensive simulations using synthetic data to demonstrate the practicability of our models. Finally, to demonstrate the wider scope of our model, we applied our model to forecast the COVID-19 cases at the county level (Estonia) and regional level (Rhône-Alpes region in France).

The rest of the chapter is organized as follows. Next, we discuss the background. We then describe preliminaries and derivations of fully-mixed and complex network models in Section 2.2. Section 2.3 and 2.4 present the evaluation of fully-mixed and complex network models respectively. The results of our model on real-world data of Estonia and the Rhône-Alpes region in France are discussed in Section 2.5. Lastly, we summarize our findings in Section 2.6.

## 2.1. Background

This section is going to cover epidemic modelling and mathematical modelling for epidemics, as well as look at different models. In particular, we focus on *SI (Susceptible-Infected)*, and *SIR* models. In 1927, W. O. Kermack and A. G. McKendrick developed the very first epidemic models [KM27]. These models are relatively simple as they are based on differential equations, but they are still crucial to epidemiology today.

These models are known as compartmental models, as they divide the entire population into compartments. A person can only be in one compartment at a time. The first compartment that is *Susceptible* contains people who have not yet been infected. Then there is the second compartment called *Infected*, which contains people who are infected, which means they are sick and can transmit the disease. People who have recovered are housed in the third compartment called *Recovered*. These are the individuals who became infected and then recovered. These individuals cannot be infected again, implying that they are immune to the disease and cannot disseminate it to others. People who died are also included among the recovered because they can no longer participate in disease transmission.

The compartmental models are also known as fully-mixed with closed population. Here, fully-mixed means that every person can interact with every other person. The term "closed population" refers to the fact that the population within the model does not vary over time. This corresponds to real-world disease transmission in the sense that the rate at which infection spreads is significantly higher than the birth rate. *SI* and *SIR* are two examples of fully-mixed models. Various researchers have also proposed other variants of compartmental models. However, given the scope of our work, we discussed the *SI* and *SIR* models in detail.

**SI model.** First, we present the *SI* model. In this model, there are two compartments, *Susceptible* (represented by $S$) and *Infected* (represented by $I$). The population in the model is closed. Initially, the majority part of the population is susceptible and only a few people are infected. As time progresses, susceptible people become infected. In the end, the entire population becomes infected. The simplicity of the SI model makes it an excellent starting point for understanding the fundamental properties of compartmental models. Let's put what we discussed into mathematical equations:

$$S \longrightarrow I \tag{2.1}$$

$$S(t) + I(t) = N \tag{2.2}$$

here, $S$, $I$, $S(t)$, $I(t)$, and $N$ represent susceptible individuals, infected individuals, susceptible individuals at time $t$, infected individuals at time $t$, and total population respectively. Equation 2.1 shows that susceptible people become infected with time. Equation 2.2 shows that the number of susceptible and infected individuals at any given time $t$ equals the total population, $N$.

Next, we introduce the disease transmission or infection rate which is the number of infection-transmitting contacts per person per unit of time. That implies that these are the contacts that lead to the transmission of the disease. Lets $\beta$ represent the transmission rate. The differential equations for infection are as follows:

$$I(t+\delta t) = I(t) + \beta \frac{S(t)}{N} I(t) \delta t \tag{2.3}$$

$$\frac{dI(t)}{dt} = \beta \frac{S(t)}{N} I(t) \tag{2.4}$$

where, $I(t+\delta t)$ represents the number of infected people at time $(t+\delta t)$, which is the sum of the number of infected individuals at time $t$ and the number of susceptible individuals who got infected in time $\delta t$ with transmission rate as $\beta$. We can rewrite Equation 2.3 as Equation 2.4, which also represents a differential equation for infected individuals.

Next, we convert the number of susceptible and infected individuals into fractions. So, we divide the number of susceptible and infected by the total population $N$ to get fractions as follows: i(t) = I(t)/N, and s(i) = S(t)/N, where i(t), and s(t) represents the fraction of infected and susceptible individuals at time $t$. Now, the differential equations for the *SI* model are:

$$\frac{di(t)}{dt} = \beta s(t) i(t) \tag{2.5}$$

$$\frac{ds(t)}{dt} = -\beta s(t) i(t) \tag{2.6}$$

$$s(t) + i(t) = 1 \tag{2.7}$$

The solution for the *SI* model's differential equations for $i(t)$ if the initially infected fraction is $i_0$ is

$$i(t) = \frac{i_0}{i_0 + (1 - i_0) e^{-\beta t}} \tag{2.8}$$

where, if we put time as infinity ($t \longrightarrow \infty$) in Equation 2.8, then all the individuals become infected, $i(t) \longrightarrow 1$ and $s(t) \longrightarrow 0$.

**SIR model.** The *SIR* model is the most relevant model to understand the spread of coronavirus. In this model, there are three compartments: *Susceptible*, *Infected*, and *Recovered*. The population in the model is also closed. Initially, the majority part of the population is susceptible and only a few people are infected. As time progresses, susceptible people become infected, and eventually infected people recover. We require two parameters to understand *SIR* model, one is the infection rate ($\beta$) with which the susceptible people become infected; the second is the recovery rate (represented by $\mu$), which is the rate with which infected people could recover. Let's put what we discussed into mathematical equations:

$$S \longrightarrow I \longrightarrow R \tag{2.9}$$

$$S(t) + I(t) + R(t) = N \tag{2.10}$$

where, $S$, $I$, and $R$ represent susceptible individuals, infected individuals, and recovered individuals respectively. $S(t)$, $I(t)$, $R(t)$ and $N$ represents susceptible

individuals at time $t$, infected individuals at time $t$, recovered individuals at time $t$ and total population respectively. Equation 2.9 shows that susceptible people become infected and infected people recovered with time. Equation 2.10 shows that the number of susceptible, infected, and recovered individuals at any given time $t$ equals the total population, $N$.

Next, we introduce disease transmission and recovery rates. The differential equations for the SIR model are as follows:

$$\frac{ds(t)}{dt} = -\beta s(t)i(t) \tag{2.11}$$

$$\frac{di(t)}{dt} = \beta s(t)i(t) - \mu i(t) \tag{2.12}$$

$$\frac{dr(t)}{dt} = \mu i(t) \tag{2.13}$$

where, $s(t)$, $i(t)$, $r(t)$ are, respectively, the fraction of susceptible, infected and recovered population at time $t$.

In the *SIR* model differential equations, when time goes to infinity ($t \longrightarrow \infty$), $\frac{dr}{dt} = 0$, the total size of the outbreak $r_\infty = const$, that is

$$1 - r_\infty = s_0 e^{-\frac{\beta}{\mu} r_\infty} \tag{2.14}$$

Initially, if we consider most people were susceptible, that means $s_0 \approx 1$, then Equation 2.14 will become

$$r_\infty = 1 - e^{-\frac{\beta}{\mu} r_\infty} \tag{2.15}$$

Now replace $\frac{\beta}{\mu}$ as $R_0$ in Equation 2.15. Here, $R_0$ represents the basic reproduction number of a disease. In simple words, it is the average number of secondary infections produced when one infected individual is introduced into the population where everyone is susceptible. The $R_0$ plays a vital role to decide whether a disease is an epidemic or not. If $R_0 > 1$, that is $\beta > \mu$, then the disease is considered an epidemic. On the other hand, if $R_0 < 1$, that is $\beta < \mu$, then the disease is not considered an epidemic. Therefore, $R_0$ is the threshold that determines when an infection can invade and persist in a new host population.

In the past, different variations of the SIR model have also been proposed to capture various real-world scenarios. For example, introducing a delay in the model to capture the incubation period during the spreading [Zha+10; Xia+12; Liu15; ASS18] or the introduction of interventions such as antiviral drugs [Tow+11]. In a different work to represent the non-linear nature of epidemic spread, a SIR rumor spreading model was proposed in which tie strengths were dependent on nodes' degree [SS12]. Apart from SIR based models, there exist several flavors of compartmental models, which represent different scenarios such as SIS [Nås96],

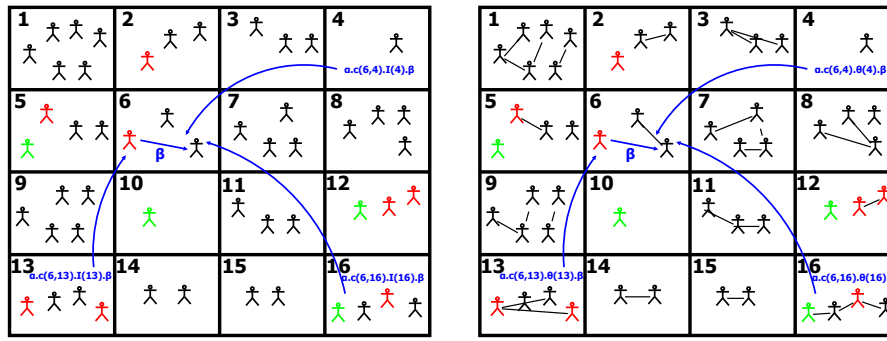
where individuals do not recover and can become susceptible again. This model has also been studied using varying types of underlying topologies [SDC08].

A set of works have also focused on exhibiting the epidemic spreading by using varying types of underlying network structures. For example, authors in [MPV02; Bar+05; Ves12] used a scale-free network and in [LW06] a small-world evolving network for evaluating their epidemiological framework. In their work in [Kis14], researchers combine a discrete, stochastic SEIR (E stands for exposed) model with a three-scale community network model to demonstrate that the different regional trends may be explained by different community mixing rates. A detailed study concerning various epidemic models on varying topologies has been done in [Pas+15].

In another line of work, the authors proposed models to understand epidemics based on the speed of growth. For example, in [VSC16], authors applied their generalized-growth model to characterize the ascending phase of an outbreak on 20 different epidemics. Their findings revealed that sub-exponential growth is a common phenomenon, especially for pathogens that are not airborne. In another work [Hua+16], researchers explained the rapid spread of H1N1 in 2009 around the world by using a flexible Bayesian, space-time, Susceptible Infected Recovered (SIR) modeling approach. [Goj+09] developed a simulation model of a pandemic (H1N1) 2009 outbreak in a structured population using demographic data from a medium-sized city in Ontario and epidemiologic influenza pandemic data.

In another work [Czy], the authors initially utilized limited data and simple mathematical models to understand virus spreading, the requirement of hospitalization, and the fatality ratio. Later, in order to capture patterns of disease transmission, they also used classical epidemic models. They have around 20 workstreams in total running and five key teams that run models and produce outputs that are helping governments around the world. Professor Azra Ghani and her group have built a relatively simple compartmental model to track the epidemic in over 100 Low and Middle-Income Countries (LMICs) around the world. Another group led by Dr. Samir Bhatt is examining publicly available data on COVID-19 deaths to back-calculate in time what that implies for infections in a certain period, then correlate those infection rates with the timing of interventions.

The contribution of our work is two-fold: First, we propose a mobility-based SIR model considering the heterogeneity and mobility of people for a fully-mixed model [GS20]. However, there is a limitation of the fully-mixed model that is it assumes that every person at every location is linked to everyone else at that location. However, in reality, people interact with a limited number of people to form a complex network with non-trivial topological features that do not occur in simple networks such as lattices or random graphs but often occur in networks representing real systems [AB02]. Therefore, in the second model, we propose a mobility-based SIR model for complex networks that captures more realistic interactions than fully-mixed models.




(a) Fully-Mixed Model (b) Complex Networks Model

Figure 1: Local And Global Transmission Of Infection In Fully-Mixed & Complex Networks Model

We model the regions in a 2-dimensional lattice, where each cell represents the mobility parameter (or direct connectivity) from one region to another (see Figure 1). Each cell represents a separate region with a population density (here regions are 1 to 16). Individuals in each cell are color-coded: Black (Susceptible), Red (Infected), and Green (Recovered). The local transmission rate of infection is $\beta$ for all cells. Figure (a) shows the fully-mixed model and infection in this model can transfer as follows: For region 6, its social connectivity is $\alpha$. The mobility of individuals from region 4 to region 6 and the fraction of infected individuals in region 4 is represented as $c(6,4)$ and $I(4)$ respectively. Therefore, an infection can transfer from region 4 to 6 via global transmission rate $\alpha c(6,4)I(4)\beta$. Similarly, $\alpha c(6,13)I(13)\beta$ and $\alpha c(6,16)I(16)\beta$ signifies the global transmission rate from region 13 and 16 respectively to region 6. On the other hand, Figure (b) shows the complex network model: similar to the fully-mixed model, an infection can transfer from region 4 to 6 via global transmission rate $\alpha c(6,4)\Theta(4)\beta$, where $\Theta$ takes care of the degree of the individual (see Section 2.2 for detail).

## 2.2. Model Preliminaries and Derivations

In this section, we present our proposed models for fully-mixed and complex networks. Let '$L$' represent a set containing all locations, and '$c$' denote the connection (or individuals' mobility) between locations. The propagation of infection at each location is explained as follows: each healthy individual can get the infection either from an infected individual located in the same location (local transmission) or from an individual visiting from other connected locations (global transmission). The local transmission rate of infection is represented by $\beta$ and the recovery rate as $\mu$, with $\beta$ and $\mu \in [0,1]$.

### 2.2.1. Non-Linear Dynamical System for Fully-Mixed Model

Next, we discuss the local transmission of infection, the global transmission, and then the dynamical behaviour of the non-linear system of infection for a fully-mixed model. The reader can refer to Table 1 for notations and their meaning.

| Notations | Meaning |
|---|---|
| $L$ | Set of all locations |
| $c$ | Connection between locations |
| $S_i(t)$ | Number of susceptible individual at location $i$ at time $t$ |
| $I_i(t)$ | Number of infected individual at location $i$ at time $t$ |
| $R_i(t)$ | Number of recovered individual at location $i$ at time $t$ |
| $N_i(t)$ | Population at location $i$ at time $t$ |
| $\alpha$ | Social connectivity parameter. In Fully-Mixed Model, this parameter controls the interaction of nodes with other nodes, and its value ranges from 0 to 1 |
| $\beta$ | Infection rate |
| $\mu$ | Recovery rate |
| $J$ | Subset of all locations $L$ |
| $c_{i,j}$ | Individuals mobility from location $j$ to $i$ |

Table 1: Parameters description for non-linear dynamical system

#### Local Transmission

Let $N_i$ be the population at location $i$, where $i \in L$, and the total population are divided into three compartments. The compartments for location $i$ at time $t$ are as follows:

1. $S_i(t)$: the number of individuals susceptible or not yet infected at location $i$ at time $t$. This compartment is referred to as *susceptible compartment*.
2. $I_i(t)$: the number of infected individuals at location $i$ at time $t$ which can further spread the disease to the individuals present in the susceptible compartment. This compartment is referred to as *infected compartment*.
3. $R_i(t)$: the number of individuals at location $i$ at time $t$ who have been recovered from the infected compartment. This compartment is referred to as *recovered compartment*.

Our assumptions regarding the transmission of an individual from one compartment to another compartment are as follows:

1. A healthy individual after becoming infected moves from susceptible to the infected compartment.
2. An individual can recover spontaneously at any time with recovery rate $\mu$. The recovery of an individual is independent of healthy and infected compartments' individuals.
3. Once the individual gets recovered, they will become immune to the disease and, thus, will not transmit the infection to individuals in the susceptible compartment.
4. In addition, this model ignores the demography that is the birth or death of individuals. In other words, the population remains constant.

### Global Transmission

Let $J$ ($J \subset L$) represent a set of locations, which are connected to location $i$. Therefore, $\sum_{j\in J} N_j$ is the maximum possible number of individuals connected to location $i$, from all the locations $J$. The parameter $c_{i,j}$ reflects the mobility of individuals from location $j$ to location $i$. Global transmission depends upon this mobility parameter of individuals from one location to another. Similar to local transmission, $I_j$ is the number of individuals in the infected compartment in location $j$. Hence, the total mobility of infected individuals from all the other connected locations to location $i$ is $\sum_{j\in J} c_{i,j}\frac{I_j}{N_j}$.

Considering the above description, the chances of transmission of infection from all the connected locations to location $i$ is $\sum_{j\in J} c_{i,j}\frac{I_j}{N_j}\beta$. This transmission further depends upon the *social connectivity* ($\alpha$) of all the individuals at location $i$. Therefore, the proportion of healthy individuals at location $i$ which can get infected from infected individuals from location $j$ is $\frac{\alpha\sum_{j\in J} c_{i,j}\frac{I_j}{N_j}\beta}{N_i+\sum_{j\in J} c_{i,j}}$. Thus, the mean-field equations for the dynamics of the pandemic, based on the above discussed interactions are the following:

$$\frac{dS_i(t)}{dt} = -\frac{\beta S_i(t) I_i(t)}{N_i(t)} - \frac{\alpha S_i(t)\sum_{j\in J} c_{i,j}\frac{I_j(t)}{N_j(t)}\beta}{N_i(t)+\sum_{j\in J} c_{i,j}} \tag{2.16}$$

$$\frac{dI_i(t)}{dt} = \frac{\beta S_i(t) I_i(t)}{N_i(t)} + \frac{\alpha S_i(t)\sum_{j\in J} c_{i,j}\frac{I_j(t)}{N_j(t)}\beta}{N_i(t)+\sum_{j\in J} c_{i,j}} - \frac{\mu I_i(t)}{N_i(t)} \tag{2.17}$$

$$\frac{dR_i(t)}{dt} = \frac{\mu I_i(t)}{N_i(t)} \tag{2.18}$$

Where Equation 2.16 describes the rate of change of susceptible individuals at location $i$, Equation 2.17 refers to the rate of change of infected individuals, and Equation 2.18 explains the rate of change of recovered individuals at location $i$.

### Dynamical Behaviour of the Linear System

Equation (2.16-2.18) represents a nonlinear dynamical system of a pandemic spreading, where, at any time $t$,

$$S_i(t) + I_i(t) + R_i(t) = N_i(t) \tag{2.19}$$

In order to solve the mean-field Equation (2.16-2.18), the following assumptions are made (please note that these assumptions are not considered during our experiments):

1. Initially, the population at all locations is equal to *N(t)* at time $t$.
2. Individuals in infected compartments are equal to $I(t)$ at all locations at time $t$ and $\sum_{j \in J} I_j = |j| \cdot I_j = kI_j$, where $k$ is the number of locations connected to location $i$, that is, $k = |j|$.
3. The mobility of individuals from one location to another location is a fraction of the total population $N$. Let $n$ be the sum of the fraction of population mobility from $|k|$ locations. Then, the total individuals' mobility from a set of locations $j$ to $i$ is $n * N$. Therefore, $\sum_{j \in J} c_{i,j} = nN$.

By considering the above assumptions, Equation 2.16 and 2.18 can be written as

$$\frac{dS_i(t)}{dt} = -\frac{\beta S_i(t) I(t)}{N(t)} - \frac{\alpha S_i(t) nN(t) k \frac{I(t)}{N(t)} \beta}{N(t) + nN(t)} \tag{2.20}$$

$$\frac{dR_i(t)}{dt} = \frac{\mu I(t)}{N(t)} \tag{2.21}$$

From Equation 2.20 and 2.21

$$\frac{dS_i(t)}{dR_i(t)} = -\frac{\beta S_i(t)}{\mu} - \frac{\alpha S_i(t) nk\beta}{\mu(1+n)} \tag{2.22}$$

$$= -\frac{\beta S_i(t)}{\mu}\left[1 + \frac{\alpha nk}{1+n}\right] \tag{2.23}$$

$$= -\frac{\beta S_i(t)}{\mu}\left[\frac{1 + (1+\alpha k)n}{1+n}\right] \tag{2.24}$$

For simplicity, Equation 2.24 can be written as:

$$\frac{dS(t)}{dR(t)} = -\frac{\beta S(t)}{\mu}\left[\frac{1 + (1+\alpha k)n}{1+n}\right] \tag{2.25}$$

Equation 2.25 can be rewritten as

$$S = S_0 e^{-\frac{\beta}{\mu} R\left[\frac{1+(1+\alpha k)n}{1+n}\right]} \tag{2.26}$$

$$\frac{dR}{dt} = \mu(N - R - S_0 e^{-\frac{\beta}{\mu}R\left[\frac{1+(1+\alpha k)n}{1+n}\right]}) \tag{2.27}$$

Solving Equation 2.27, we get

$$t = \frac{1}{\mu}\int_0^R \frac{dR}{N - R - S_0 e^{-\frac{\beta}{\mu}R\left[\frac{1+(1+\alpha k)n}{1+n}\right]}} \tag{2.28}$$

As pandemic arrives at steady state when $t \longrightarrow \infty$ hence $\frac{dR}{dt} = 0$ and $R_\infty = C$, where $C$ is a constant:

$$N - R_\infty = S_0 e^{-\frac{\beta}{\mu}R_\infty\left[\frac{1+(1+\alpha k)n}{1+n}\right]} \tag{2.29}$$

Let the initial conditions be: $R(0) = 0$, $I(0) = I$ and $S(0) = N - I \approx N$. Therefore, Equation 2.29 can be written as:

$$R_\infty = N - Ne^{-\frac{\beta}{\mu}R_\infty\left[\frac{1+(1+\alpha k)n}{1+n}\right]} \tag{2.30}$$

Normalizing Equation 2.30 by dividing by total population $N$ gives:

$$r_\infty = 1 - 1e^{-R_0 r_\infty} \tag{2.31}$$

Therefore, the reproduction number $R_0$ is

$$R_0 = \frac{\beta}{\mu}\left[\frac{1+(1+\alpha k)n}{1+n}\right] \tag{2.32}$$

In case there is no social connectivity to other locations ($\alpha = 0$ or $k = 0$ or $n = 0$) then the mobility SIR model will become the standard SIR model and the reproduction number is $R_0 = \frac{\beta}{\mu}$. Therefore, the reproduction number is directly proportional to social connectivity parameter $\alpha$, the number of connected locations $k$, and depends upon individuals' mobility during a pandemic.

The fully-mixed model that we presented in this section has one key limitation, *i.e.,* it assumes that every person at every location is linked to everyone else at that location. However, in reality, people interact with a limited number of people to form a complex network with non-trivial topological features that do not occur in simple networks such as lattices or random graphs but often occur in networks representing real systems [AB02]. Therefore, in the next section, we propose a mobility-based SIR model for complex networks.

### 2.2.2. Non-Linear Dynamical System for Complex Networks

In this section, we discuss the local transmission of infection, the global transmission, and then the dynamical behaviour of the non-linear system of infection by considering complex networks interactions at each location.

### Local Transmission

Let $N_i$ be the population at location $i$, where $i \in L$, and $k$ is the degree of each individual, where $k \in \mathbb{W}$ (whole numbers). The total population is divided into three compartments. The compartments for location $i$ at time $t$ are as follows:

1. $S_i(k,t)$: the number of individuals susceptible or not yet infected at location $i$ at time $t$ having degree $k$. This compartment is referred to as *susceptible compartment*.
2. $I_i(k,t)$: the number of infected individuals at location $i$ at time $t$ having degree $k$, which can further spread the disease to the individuals present in the susceptible compartment. This compartment is referred to as *infected compartment*.
3. $R_i(k,t)$: the number of individuals at location $i$ at time $t$ having degree $k$, who have been recovered from infected compartment. This compartment is referred to as *recovered compartment*.

Our assumptions regarding the transmission of an individual from one compartment to another compartment are the same as discussed in Section 2.2.1.

### Global Transmission

Let $j$ ($j \subset L$) represent a set of locations, which are connected to location $i$. Therefore, $\sum_{j\in J} N_j(k)$ is the maximum possible number of individuals of degree $k$ connected to location $i$, from all the locations $J$. The parameter $c_{i,j,k}$ reflects the mobility of individuals of degree $k$ from locations $j$ to location $i$. Global transmission depends upon this mobility parameter of individuals from one location to another. Similar to local transmission, $I_j$ is the number of individuals in the infected compartment in all the locations $j$. Hence, total mobility of infected individuals of degree $k$ from all the other connected locations to location $i$ is $\sum_{j\in J} c_{i,j,k} \frac{I_j(k)}{N_j(k)}$.

Considering the above description, the chances of transmission of infection from all the connected locations to location $i$ is $\sum_{j\in J} c_{i,j,k} \frac{I_j(k)}{N_j(k)} \beta$. This transmission further depends upon the *social connectivity* ($\alpha$) of all the individuals at location $i$. Therefore, the proportion of healthy individuals at location $i$ which can get infected from infected individuals from location $j$ is $\frac{\alpha \sum_{j\in J} c_{i,j,k} \frac{I_j(k)}{N_j(k)} \beta}{N_i(k)+\sum_{j\in J} c_{i,j,k}}$. Thus, the mean-field equations for the non-linear dynamics of the pandemic, based on the above discussed interactions are the following:

$$\begin{aligned} \frac{dS_i(k,t)}{dt} &= -\frac{\beta S_i(k,t)\Theta_i(t)}{N_i(k,t)} \\ &\quad - \frac{\alpha S_i(k,t) \sum_{j\in J} c_{i,j,k} \frac{\Theta_j(t)}{N_j(k,t)} \beta}{N_i(k,t) + \sum_{j\in J} c_{i,j,k}} \end{aligned} \tag{2.33}$$

$$\begin{aligned}\frac{dI_i(k,t)}{dt} &= \frac{\beta S_i(k,t)\Theta_i(t)}{N_i(k,t)} \\ &\quad + \frac{\alpha S_i(k,t)\sum_{j\in J} c_{i,j,k}\frac{\Theta_j(t)}{N_j(k,t)}\beta}{N_i(k,t)+\sum_{j\in J} c_{i,j,k}} \\ &\quad - \frac{\mu I_i(k,t)}{N_i(k,t)} \end{aligned} \tag{2.34}$$

$$\frac{dR_i(k,t)}{dt} = \frac{\mu I_i(k,t)}{N_i(k,t)} \tag{2.35}$$

where,

$$\Theta_i(t) = \sum_{k'=1}^{k} \frac{\Psi(k')P\left(\frac{k'}{k}\right) I_i(k',t)}{k'} \tag{2.36}$$

$$P\left(\frac{k'}{k}\right) = \frac{k'P(k')}{< k >} \tag{2.37}$$

Where, Equation 2.33 describes the rate of change of susceptible individuals of degree $k$ at location $i$, and Equation 2.34 refers to the rate of change of infected individuals of degree $k$, and Equation 2.35 explains the rate of change of recovered individuals of degree $k$ at location $i$. Please refer to Table 2 for notations and their meaning.

### Dynamical Behaviour of the Non-Linear System for Complex Networks

Equation (2.33-2.35) represents non-linear dynamical system of a pandemic spreading for complex networks, where, at any time $t$:

$$S_i(t) + I_i(t) + R_i(t) = N_i(t) \tag{2.38}$$

where,

$$X(t) = \sum_k X(k,t);\ X \in \{S,I,R,N\} \tag{2.39}$$

In order to solve mean-field Equation (2.33-2.35) similar assumptions as in Section 2.2.1 are made. By considering such assumptions, Equation 2.33, 2.34 and 2.35 can be written as

$$\begin{aligned}\frac{dS_i(k,t)}{dt} &= -\frac{\beta S_i(k,t)\Theta(t)}{N(k,t)} \\ &\quad - \frac{\alpha S_i(k,t) nN(k,t) m\frac{\Theta(t)}{N(k,t)}\beta}{N(k,t)+nN(k,t)} \end{aligned} \tag{2.40}$$

| Notations | Meaning |
|---|---|
| $L$ | Set of all locations |
| $c$ | Connection between locations |
| $S_i(k,t)$ | Number of susceptible individual of degree $k$ at location $i$ at time $t$ |
| $I_i(k,t)$ | Number of infected individual of degree $k$ at location $i$ at time $t$ |
| $R_i(k,t)$ | Number of recovered individual of degree $k$ at location $i$ at time $t$ |
| $N_i(k,t)$ | Population of degree $k$ at location $i$ at time $t$ |
| $\alpha$ | Social connectivity parameter. In complex networks, this parameter restricts the edges of a node with its neighbors, and its value ranges from 0 to 1 |
| $\beta$ | Infection rate |
| $\mu$ | Recovery rate |
| $J$ | Subset of all locations $L$ |
| $c_{i,j,k}$ | Mobility of individuals of degree $k$ from location $j$ to $i$ |
| $\Psi(k')$ | Infectivity strength of a node (with a degree $k'$) to spread the infection to its neighbors |
| $m$ | The number of connected locations |
| $P\left(\frac{k'}{k}\right)$ | Correlation between degree $k'$ and $k$ |
| $P(k')$ | Probability of a node with degree $k'$ |

Table 2: Parameters description for non-linear dynamical system for complex network

$$\begin{aligned}\frac{dI_i(k,t)}{dt} &= \frac{\beta S_i(k,t)\Theta(t)}{N(k,t)} \\ &\quad + \frac{\alpha S_i(k,t) n N(k,t) m \frac{\Theta(t)}{N(k,t)} \beta}{N(k,t) + nN(k,t)} \\ &\quad - \frac{\mu I(k,t)}{N(k,t)} \qquad (2.41)\\ \frac{dR_i(k,t)}{dt} &= \frac{\mu I(k,t)}{N(k,t)} \qquad (2.42)\end{aligned}$$

Note in Equation 2.40 and 2.41, $m$ represents the number of connected locations. We used $m$ instead of $k$ as $k$ represents a node's degree. Now, from Equation 2.40 and 2.42

$$\frac{dS_i(k,t)}{dR_i(k,t)} = -\frac{\beta S_i(k,t)\Theta(t)}{\mu I(k,t)} - \frac{\alpha S_i(k,t) nm\Theta(t)\beta}{\mu(1+n)I(k,t)} \tag{2.43}$$

$$= -\frac{\beta S_i(k,t)\Theta(t)}{\mu I(k,t)}\left[1+\frac{\alpha nm}{1+n}\right] \tag{2.44}$$

$$= -\frac{\beta S_i(k,t)\Theta(t)}{\mu I(k,t)}\left[\frac{1+(1+\alpha m)n}{1+n}\right] \tag{2.45}$$

For simplicity, Equation 2.45 can be written as:

$$\frac{dS(k,t)}{dR(k,t)} = -\frac{\beta S(k,t)\frac{\langle k^2\rangle}{\langle k\rangle} I(k,t)}{\mu I(k,t)}\left[\frac{1+(1+\alpha m)n}{1+n}\right] \tag{2.46}$$

$$\frac{dS(k,t)}{dR(k,t)} = -\frac{\beta S(k,t)\frac{\langle k^2\rangle}{\langle k\rangle}}{\mu}\left[\frac{1+(1+\alpha m)n}{1+n}\right] \tag{2.47}$$

Equation 2.47 can be rewritten as

$$S = S_0 e^{-\frac{\beta\frac{\langle k^2\rangle}{\langle k\rangle}}{\mu} R\left[\frac{1+(1+\alpha m)n}{1+n}\right]} \tag{2.48}$$

$$\frac{dR}{dt} = \mu\left(N-R-S_0 e^{-\frac{\beta\frac{\langle k^2\rangle}{\langle k\rangle}}{\mu} R\left[\frac{1+(1+\alpha m)n}{1+n}\right]}\right) \tag{2.49}$$

Solving Equation 2.49, we get

$$t = \frac{1}{\mu}\int_0^R \frac{dR}{N-R-S_0 e^{-\frac{\beta\frac{\langle k^2\rangle}{\langle k\rangle}}{\mu} R\left[\frac{1+(1+\alpha m)n}{1+n}\right]}} \tag{2.50}$$

As pandemic arrives at steady state, when $t \longrightarrow \infty$, hence $\frac{dR}{dt} = 0$ and $R_\infty = C$, where $C$ is a constant.

$$N - R_\infty = S_0 e^{-\frac{\beta\frac{\langle k^2\rangle}{\langle k\rangle}}{\mu} R_\infty\left[\frac{1+(1+\alpha m)n}{1+n}\right]} \tag{2.51}$$

Let initial conditions are $R(0) = 0$, $I(0) = I$ and $S(0) = N - I \approx N$. Therefore, Equation 2.51 can be written as

$$R_\infty = N - Ne^{-\frac{\beta\frac{\langle k^2\rangle}{\langle k\rangle}}{\mu} R_\infty\left[\frac{1+(1+\alpha m)n}{1+n}\right]} \tag{2.52}$$

Normalizing Equation 2.52 by dividing by total population $N$ gives:

$$r_\infty = 1 - 1e^{-R_0 r_\infty} \tag{2.53}$$

Therefore, the reproduction number $R_0$ is

$$R_0 = \frac{\beta \frac{\langle k^2 \rangle}{\langle k \rangle}}{\mu} \left[ \frac{1 + (1 + \alpha m)n}{1 + n} \right] \tag{2.54}$$

The $R_0$ is called basic reproduction number which determines the spread of infection. When $R_0 > 1$, the propagation occurs at a fast rate. When $R_0 = 1$ , the propagation happens at a slow rate. When $R_0 < 1$, the propagation finishes. In case there is no social connectivity to other locations ($\alpha = 0$ or $m = 0$ or $n = 0$), then the mobility SIR model for complex networks gives the reproduction number as $R_0 = \frac{\beta \frac{\langle k^2 \rangle}{\langle k \rangle}}{\mu}$. Therefore, the basic reproduction number is directly proportional to social connectivity parameter $\alpha$, number of connected locations $m$, depends upon individuals' mobility during a pandemic, and degree of an individual in a complex network.

## 2.3. Evaluation of Fully-Mixed Model

In this section, we first explain our experimental setup, and next, we discuss the results of our simulation conducted using the proposed model for non-complex network on synthetic data. In addition, we also show results of our model when applied for predicting the number of COVID-19 cases at country level (Estonia) and regional level (Rhône-Alpes region in France).

### 2.3.1. Experimental Setup

For the analysis, we created an aggregated flow matrix whose cells represent the number of trips of individuals per day from origins to destinations. We call this matrix *Origin-Destination (OD)* matrix. The synthetic OD matrix considered for our experiment follows a random distribution. Furthermore, three different techniques are considered for selecting the seed infection location:

1. *Pandemics origin from a random location:* In this, a random location is selected as seed infection location, and a small fraction of individuals were infected at that location.
2. *Pandemics origin from a weakly connected location:* Here, the seed location is selected strategically, *i.e.,* in a location which is weakly connected to other locations. This implies the least mobility of individuals from this location to other locations.
3. *Pandemics origin from a strongly connected location:* In this also, the seed location is selected strategically, *i.e.,* in a location which is strongly connected to other locations. This signifies that the highest mobility of individuals from a location is considered for the selection of the infection seed.

Our simulation is oriented towards addressing the following questions:

- How *social connectivity parameter* '$\alpha$' affects the fraction of individuals in different compartments (*susceptible, infected and recovered*) during a pandemic?
- What are the outcomes of restricting the mobility (for top-X percentile) of strongly connected locations?
- What is the relationship between *social connectivity parameter* '$\alpha$' and the mobility restriction (top-X percentile of strongly connected locations)?
- How efficiently this model can perform in real scenarios? We answer this question by projecting the expected COVID-19 cases for Estonia and Rhône-Alpes region in France.

We used Python programming language for the implementation. In particular, we used odeint function available in the scipy package to simulate mean-field equations (MFE). For all the experiments, we use Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz 1.80 GHz with 16 GB RAM.

### 2.3.2. Results

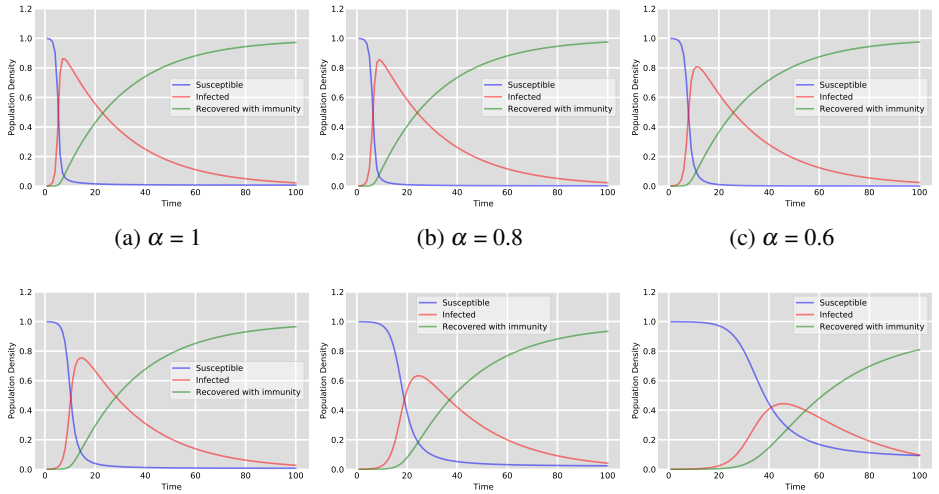
We perform various simulation experiments to explain the proposed model on *OD* matrix by using previously discussed techniques for selecting the seed infection location. It is to be noted that, the model will behave as a standard SIR model in two cases, (i) if $\alpha = 0$, (ii) if the mobility is reduced to 100 percentile (that is no mobility allowed) from connected locations.

#### Pandemic Origins from Random Location

Figure 2 displays the influence of the *social connectivity parameter* '$\alpha$' while keeping the other parameters constant. Figure 2a to 2f shows the pandemic dynamics with different values of $\alpha$ starting with $\alpha = 1$ to $\alpha = 0.1$. We observe that the peak of the infected compartment decreases significantly, as the $\alpha$ decreases, and it also takes longer to reach its peak. This indicates that there is a positive impact of lock-down in controlling a pandemic.

The effect of restricting the mobility from the top-X percentile of highly connected locations with other locations is shown in Figure 3. Figure 3a to 3d displays the pandemic dynamics with different percentile of mobility restrictions of highly connected locations starting with 0% to 30% (keeping $\alpha = 0.5$). We observe that in the case of a pandemic, restricting the mobility from the top-10 percentile of highly connected locations can reduce the number of individuals who can get infected to 27%. Therefore, quarantine plays a vital role during pandemics.

In order to understand the relationship between $\alpha$ and *mobility restriction* from strongly connected locations, we performed the numerical simulation of the proposed mean-field equations (see Figure 4). It shows the fraction of population in various compartments at time $t \to \infty$ for various value combinations of $\alpha$ and *mobility restriction*. From Figure 4, we can infer that the *social connectivity parame-*




(a) $\alpha = 1$ (b) $\alpha = 0.8$ (c) $\alpha = 0.6$

(d) $\alpha = 0.4$ (e) $\alpha = 0.2$ (f) $\alpha = 0.1$

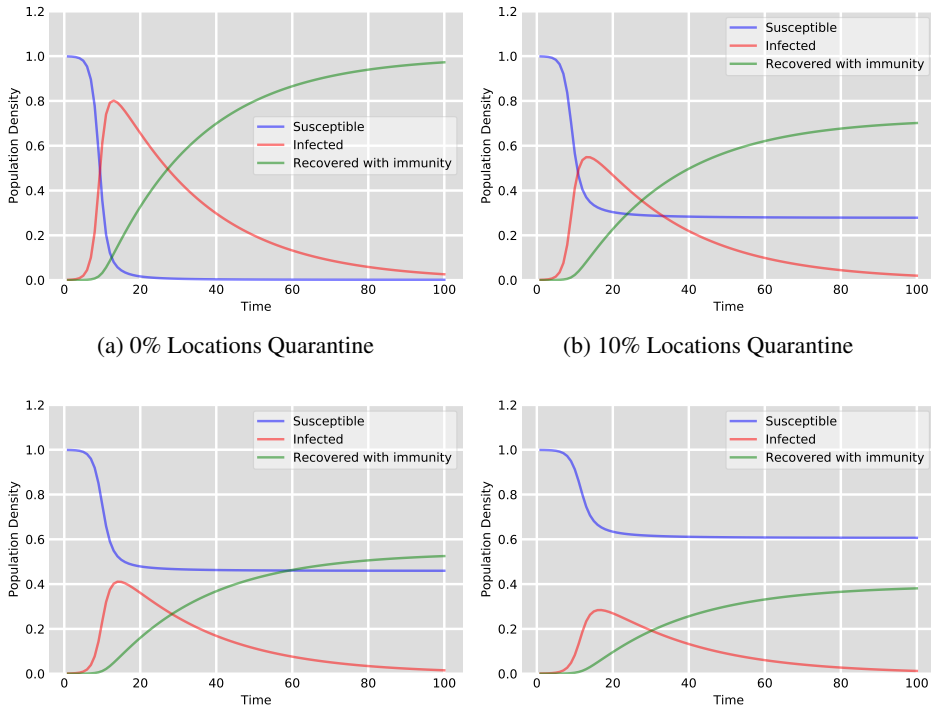
Figure 2: Pandemic Origin From Random Location: Effect of *Social Connectivity Parameter* '$\alpha$'

*ter* '$\alpha$' and *mobility* plays both a fundamental role in determining the dynamics of the pandemics. Therefore, it is advisable to follow a dual strategy approach during a pandemic outbreak as *controlling mobility* reduces the fraction of infected individuals, and $\alpha$ delays the peak. Furthermore, we analyzed the number of days required to reach the point where the highest fraction of individuals get infected (see Figure 5). This indicates that mobility restrictions and minimal social contact will postpone the pandemic's peak and will give sufficient time for preparations, especially for the health sector.

### Pandemic Origins from Weakly and Strongly Connected Locations

Figure 6, and 7 display the influence of the *social communication parameter* '$\alpha$' while keeping the other parameters constant for weakly and strongly connected locations respectively. Figure 6a to 6f shows the pandemic dynamics with different values of $\alpha$ starting with $\alpha = 1$ to $\alpha = 0.1$ for weakly connected locations. Similarly, Figure 7a to 7f shows the pandemic dynamics with different values of $\alpha$ starting with $\alpha = 1$ to $\alpha = 0.1$ for strongly connected locations.

It can be noted that when a pandemic originates from a weakly connected location, it takes longer to reach its peak compared to when it starts from a strongly connected location. This shows that the location of origin also plays an important role during a pandemic. Similar to a random location, reducing mobility from the highly connected locations by 10 percentile can reduce the number of infected individuals between 18% to 27% for weakly and strongly connected locations, respectively.




(a) 0% Locations Quarantine

(b) 10% Locations Quarantine

(c) 20% Locations Quarantine

(d) 30% Locations Quarantine

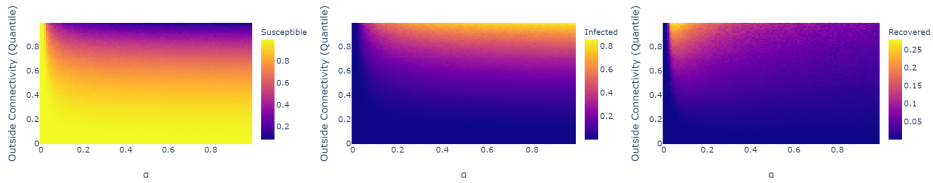
Figure 3: Pandemic Origin From Random Location: Effect of Quarantine Strongly Connected Locations

## 2.4. Evaluation of Complex Networks Model

In this section, we first explain our experimental setup, and next, we discuss the results of our simulation conducted using the proposed model for complex networks on synthetic network.

### 2.4.1. Experimental Setup

For the analysis, we created a synthetic network using configuration model [New03], which follows the power law distribution (see Figure 8) with scale-free exponent (gamma) as 3. In particular, we utilize the *random_powerlaw_tree_sequence* and *configuration_model* function of *networkx* [HSS08]. The configuration model produces a random pseudograph (graph with parallel edges and self loops) by randomly assigning edges to fit the given degree sequence. We removed all parallel edges and self loops from our network. Table 7 provides the statistics of our synthetic network. Based on the definition of configuration model and various properties of our synthetic network, we can infer that this network reflects the real-world contact network [Voi+20]. For the rest of this chapter, we refer to this network as *ConNet*.




(a) Susceptible Compartment (b) Infected Compartment (c) Recovered Compartment

Figure 4: Pandemic Origin From Random Location: Numerical simulation of relationship between $\alpha$ and *quarantine*.

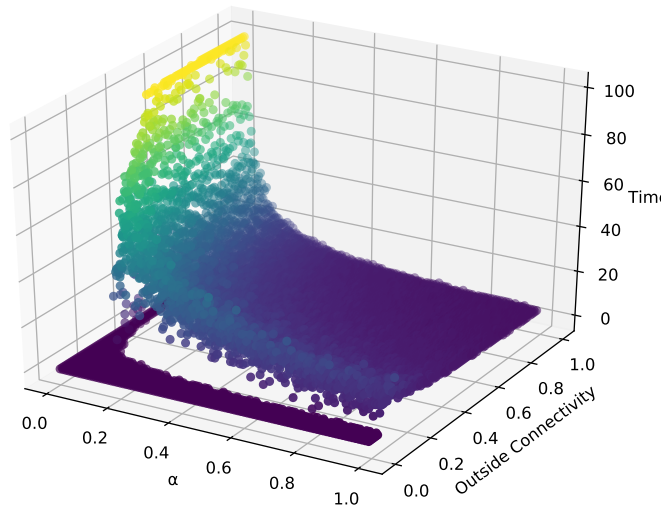


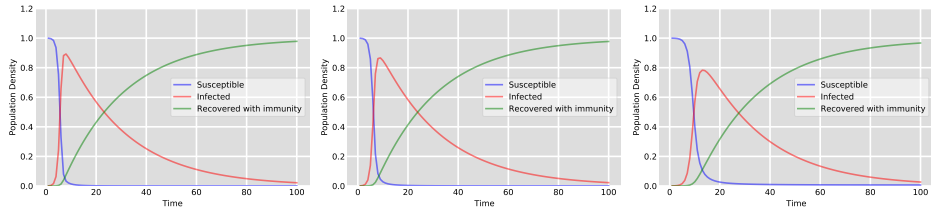

Figure 5: For different combinations of $\alpha$ and *quarantine* percentile, number of days required to reach peak of infected compartment.

Our simulation is oriented towards addressing the following questions:

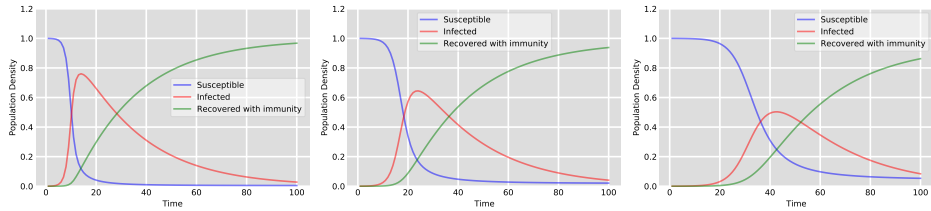
- How *social connectivity parameter* '$\alpha$' affects the fraction of individuals in different compartments (*susceptible, infected and recovered*) for a complex network?
- What are the outcomes of restricting the mobility (for top-X percentile) of strongly connected locations in a complex network?

### 2.4.2. Results

We perform various simulation experiments to explain the proposed model for complex network on *ConNet* by selecting the seed infection location randomly. It is to be noted that, the model will behave as a standard SIR model in two cases, (i) if $\alpha = 0$, (ii) if the mobility is reduced to 100 percentile (that is no mobility allowed) from connected locations.

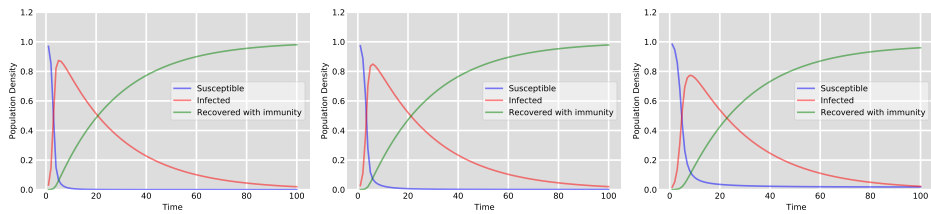



(a) $\alpha = 1$ (b) $\alpha = 0.8$ (c) $\alpha = 0.5$

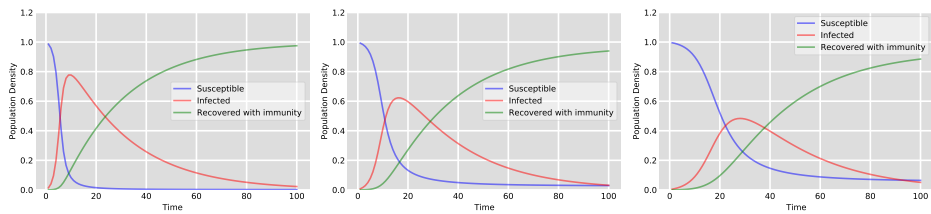



(d) $\alpha = 0.4$ (e) $\alpha = 0.2$ (f) $\alpha = 0.1$

Figure 6: Pandemic Origin From Weakly connected Location: Effect of *Social Connectivity Parameter* '$\alpha$'

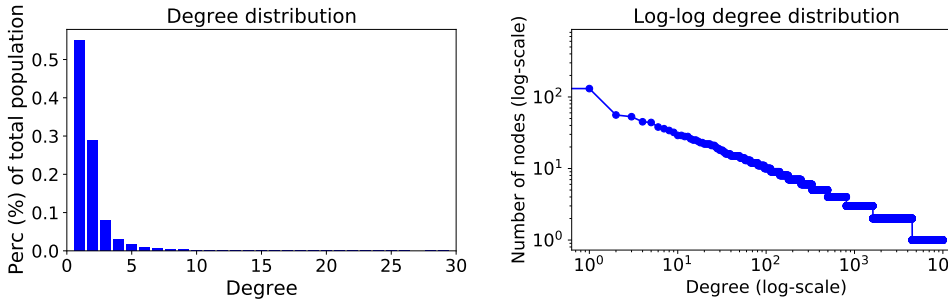



(a) $\alpha = 1$ (b) $\alpha = 0.8$ (c) $\alpha = 0.5$




(d) $\alpha = 0.4$ (e) $\alpha = 0.2$ (f) $\alpha = 0.1$

Figure 7: Pandemic Origin From Strongly connected Location: Effect of *Social Connectivity Parameter* '$\alpha$'




(a) Degree distribution of nodes.

(b) Log-log degree distribution.

Figure 8: The degree distribution of our synthetically generated network shows that it follows power law distribution.

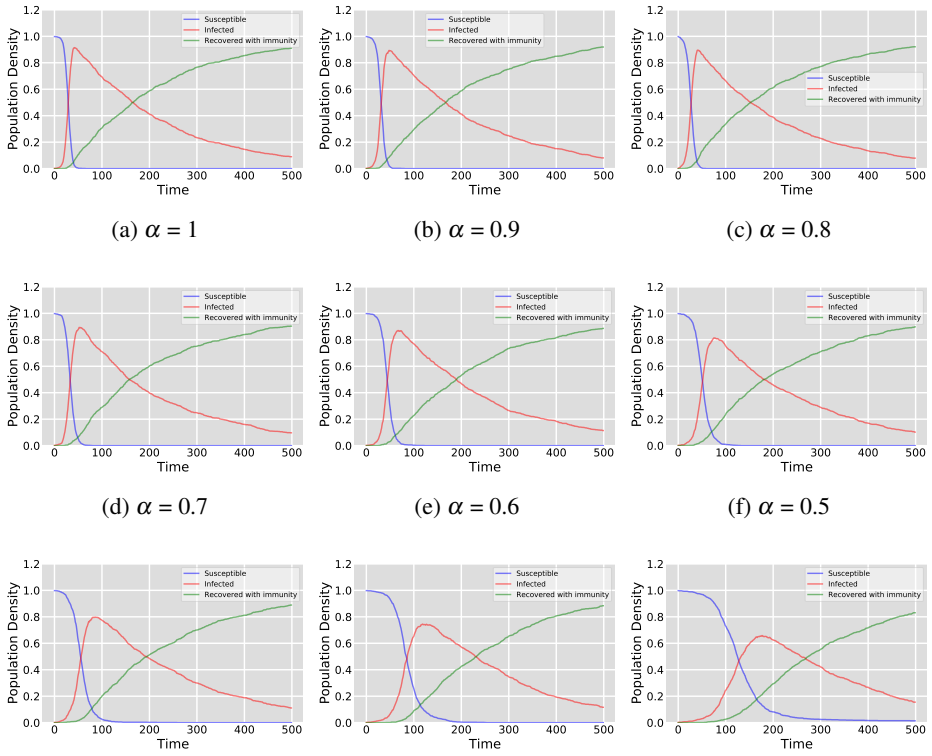
| Network Properties | Value |
|---|---|
| Nodes | 10,000 |
| Edges | 9,960 |
| Average degree | 1.992 |
| Edge density | 0.0002 |
| Number of triangles | 390 |
| Average clustering coefficient | 0.0038 |
| Number of components | 1117 |
| Reciprocity | 0 |

Table 3: Network statistics.

**Pandemic Origins From a Random Location.** Figure 9 displays the influence of the *social connectivity parameter* '$\alpha$' while keeping the other parameters constant. Figure 9a to 9i shows the pandemic dynamics with different values of $\alpha$ starting with $\alpha = 1$ to $\alpha = 0.2$. We observe that the peak of the infected compartment decreases significantly, as the $\alpha$ decreases, and it also takes longer to reach its peak. This indicates that there is a positive impact of lock-down in controlling a pandemic.

The effect of restricting the mobility from the top-X percentile of highly connected locations with other locations is shown in Figure 10. Figure 10a to 10c displays the pandemic dynamics with different percentile of mobility restrictions of highly connected locations starting with 10% to 30% (keeping $\alpha = 0.5$). We observe that in the case of a pandemic, restricting the mobility from the top-10 percentile of highly connected locations can reduce the number of individuals who can get infected to 15% to 21% (see Figure 10). Therefore, quarantine plays a vital role during pandemics.

In order to understand the relationship between $\alpha$ and *mobility restriction* from strongly connected locations, we performed the numerical simulation of the proposed mean-field equations for complex network. The results are similar to the




(a) $\alpha = 1$ (b) $\alpha = 0.9$ (c) $\alpha = 0.8$

(d) $\alpha = 0.7$ (e) $\alpha = 0.6$ (f) $\alpha = 0.5$

(g) $\alpha = 0.4$ (h) $\alpha = 0.3$ (i) $\alpha = 0.2$

Figure 9: Pandemic Origin From Random Location In Complex Network: Effect of *Social Connectivity Parameter* '$\alpha$'

mean-field equations for non-complex network (see Figure 4). Therefore, we can infer that the *social connectivity parameter* '$\alpha$' and *mobility* both plays an important role during pandemics.

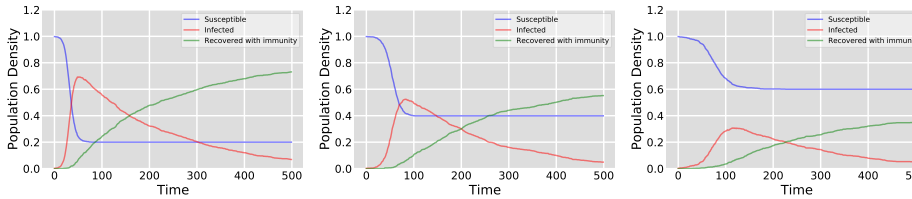
## 2.5. Results on Real-World Data of Estonia and Rhône-Alpes Region in France

Here, we show the results of our model when applied to predict the number of COVID-19 cases at the country level (Estonia) and regional level (Rhône-Alpes region in France).

### 2.5.1. Case Study of Estonia

In this section, we begin with the dataset used to calculate OD matrix for Estonia at the county level. Then, we use the extracted OD matrix to predict coronavirus cases in Estonia.

**Dataset & OD Matrix.** We utilize anonymized call data records (*CDR*) issued by one of Estonia's leading mobile operators. Please note that the dataset is times-




(a) 10% Locations Quarantine (b) 20% Locations Quarantine (c) 30% Locations Quarantine

Figure 10: Pandemic Origin From Random Location In Complex Network: Effect of Quarantine Strongly Connected Locations

tamped and contains the cell phone tower's passive mobile location (or county). The call records span from $8^{th}$ May 2017 to $13^{th}$ May 2017. The data collection consists of 12,317,970 independent call records from 1,175,191 unique individuals.

Each call activity in the CDR dataset includes: the randomly generated (pseudonymous) ID of the user, the timestamp information of the call activity, and the location of the network cell (or network Cell ID). The allocated pseudonym ID ensures user's anonymity and cannot be associated with a particular person or phone number. Please note that the Cell ID accuracy is higher in heavily populated areas (100–500 m in cities) and those with denser road networks, however accuracy is lower (500–5000 m) in sparsely populated areas [Aha+08]. Table 6 summarizes statistics for the dataset. For simulation, we created the *OD* matrix between counties of Estonia using call data records [Nov+13]. Furthermore, these call interactions are converted into population mobility between counties using Estonian population data from census [Est18].

*CDR dataset validation using census dataset.* As *CDR* data contains call records and might not represent the real population of Estonia, we validated it and find that it is indeed an acceptable representative of Estonia's actual population. Please refer next chapter for more details.

***CDR* dataset availability and Ethical concern.** The dataset is owned by our collaborator in the University of Tartu and is accessible for research purposes after signing the NDA. Additionally, the dataset is anonymized at two levels, so that specific persons cannot be identified.

**Results.** To demonstrate the usability of the model, we applied it to real-time data of Estonia's COVID-19 cases. For the local transmission of the virus (within the county), we consider the reproduction number $R_0 = 2.5$ [Org20]. Figure 11 shows the actual number of cases and the cases forecast by the model using different values for $\alpha$ and mobility percentile. For example, $\alpha = 0.95$, indicates that the social connectivity of individuals is reduced by 5% and also top-5 percentile of strongly connected locations are restricted from mobility. Similarly, $\alpha = 0.7$, implies that the social connectivity of individuals is reduced by 30%, and also the top-30 percentile of strongly connected locations have introduced restricted mobility.

| Parameters | Value |
|---|---|
| Time period | 8 May'17 to 13 May'17 |
| Call Records | 12,179,970 |
| Unique Users | 1,175,919 |
| **Locations** | |
| Harju | 507,365 |
| Hiiu | 6,959 |
| Ida-Viru | 87,212 |
| Järva | 18,157 |
| Jõgeva | 23,125 |
| Lääne | 18,377 |
| Lääne-Viru | 44,787 |
| Pärnu | 55,873 |
| Põlva | 21,083 |
| Rapla | 24,108 |
| Saare | 25,374 |
| Tartu | 132,888 |
| Valga | 17,528 |
| Viljandi | 33,767 |
| Võru | 28,405 |

Table 4: CDR data statistics.

Cases reported until $11^{th}$ March, 2020 are considered as an initial condition for the model. The reason behind selecting $11^{th}$ March, 2020 as initial condition is that, till this date no local transmission of the virus was reported [ERR20a]. Till the day of initial condition, the *Estonian Health Board* confirmed 13 cases in *Harju* and two cases in *Tartumaa* and *Saaremaa* each [Ter20]. During the simulation, the number of cases in all other counties are initialized to zero. The infection rate $\beta$ and recovery rate $\mu$ are adjusted according to the value of $R_0$ for COVID-19. The reported cases in Estonia as well as the forecast cases using the model, are shown in Figure 11 until $10^{th}$ April 2020. It can be noticed that the model predicted much higher cases of COVID-19 if no restrictions are introduced ($\alpha = 1$). However, as the restrictions were introduced by the Government, the number of cases got damped (Actual). Thus, the applicability of this model is to forecast a range of predicted number of cases which can help the government and health agencies to understand the impact and introduce proportional interventions to restrict the spread of the epidemic.

Please note that in Figure 11, the predicted cases with different alpha values and the actual cases vary from each other. The reason for this variation can be seen in the early days of the Coronavirus. During the time, the Italian volleyball

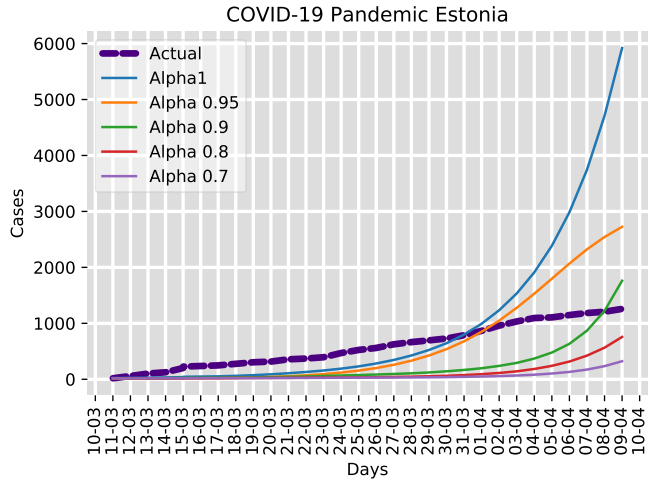



Figure 11: COVID-19 Cases In Estonia

club Power Volley Milano, which competed in the 2019–20 CEV Challenge Cup matches played on Saaremaa island on March 4th and 5th, 2020, is said to have introduced the coronavirus to Saaremaa, according to health authorities. Through a subsequent champagne celebration, the virus quickly spread throughout the area. Health officials believe that at the time, the virus had infected half of the island's inhabitants [ERR20b; Exp20; BBC20]. This leads to the rapid rise of COVID cases in Estonia in the initial days, which is a challenging task to be included in the proposed model.

### 2.5.2. Case Study of Rhône-Alpes Region in France

We also applied our model to real-time data of Rhône-Alpes region's COVID-19 cases. Figure 12 shows the actual number of cases and the cases forecast by the model using different values for $\alpha$ and mobility percentile. The region is divided into 14 sectors.

**Dataset & OD Matrix.** For simulation purposes, we again considered the *OD* matrix between the sectors of the region obtained via network signaling data of Orange (the largest telecommunications provider in France) and census data [ins16]. This OD matrix has been built using the approach proposed by Fekih et al. in [Fek+20]. The latter uses a heuristic-based approach to extract trips between sectors at individual level from mobile phone passive traces. As the signaling data collected by the telecommuting operator covers around 30% of the whole population of the region, the resulting trips have been re-scaled using the census data. Finally, after an aggregation step, we obtain the regional OD matrix used in this study.

For privacy matters, the number of COVID-19 cases is not reported in France at a fine spatio-temporal resolution in publicly available data [fra20b]. Instead, the dataset only reports cumulative values of COVID-19 cases on a 7-days rolling window for each area of the administrative segmentation of the French territory.

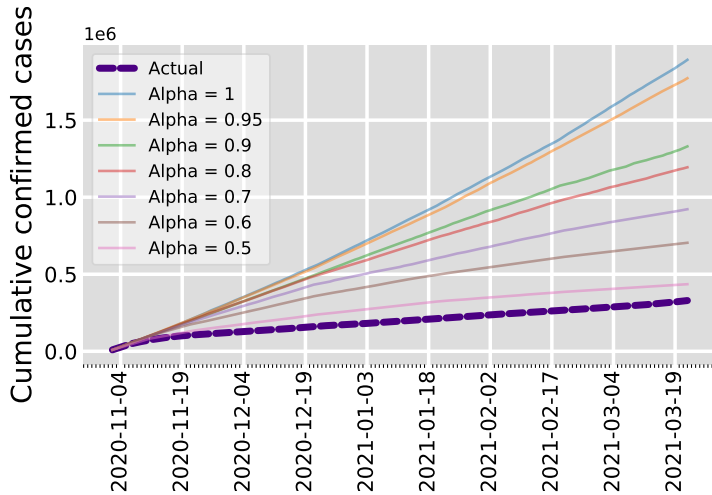



Figure 12: COVID-19 Cases In Rhône-Alpes Region In France.

In addition, the number of cases is reported discretely, *i.e.,* as a range of values between a lower and upper bound containing the real value. Therefore, in order to obtain a daily estimation of the number of COVID-19 cases per each sector of the analyzed Rhone-Alpes region, two main assumptions have been made. On the one hand, we consider the number of cases as the mid value between the lower and upper bound of the reported range for the given area on a specific day. On the other hand, we replaced the 7 days rolling time window by the median day (*i.e.,* the $4^{th}$ day of the time window). As a result, to obtain the daily estimation of the number of cases, the cumulative reported estimation provided on a 7-days rolling time window is divided by 7. After summing this estimation for all the administrative areas belonging to a given sector of the OD matrix, we finally obtain an estimation of the number of COVID-19 cases per sector and per day. COVID-19 data cover the period from $2^{nd}$ November, 2020 to $19^{th}$ March, 2021.

**Results.** For our simulation, the number of cases in all sectors is initialized as on $2^{nd}$ November, 2020. For the local transmission of the virus (within the sector), we consider the reproduction number $R_0 = 2.5$ [Org20]. The infection rate $\beta$ and recovery rate $\mu$ are adjusted according to the value of $R_0$. The reported cases in the Rhône-Alpes region in France as well as the forecast cases using the model, are shown in Figure 12 until $22^{nd}$ March 2021. It can be noticed that the model predicted much higher cases of COVID-19 if no restrictions are introduced ($\alpha = 1$), while we can observe that, for $\alpha = 0.5$, the number of actual cases and forecast ones are quite close to each other. To explain this result, it is worth to remind that strong mobility restrictions were re-introduced in France by the end of October 2020 [fra20a], after the first lockdown ended during summer. The new restrictions contributed to keep low the number of COVID-19 infections (Actual). Moreover, it is reasonable to assume that mobility and social interactions were already sig-

nificantly reduced at the beginning of this second lock-down, with respect to pre-pandemic behaviors, as a consequence of the first COVID-19 wave and previously imposed restrictive measures. In conclusion, this second case study confirms the applicability of the model to forecast a range of predicted number of cases. The latter can thus help the government and health agencies to understand the impact and introduce proportional interventions to restrict the spread of the epidemic.

## 2.6. Summary

Classical compartmental epidemic models are unable to describe the spreading pattern of pandemics such as COVID-19 as they do not take into account the effect of *social connectivity* and *mobility* in the spreading of the virus. Our proposed mobility based SIR models for fully-mixed and complex networks shows the significance of *social connectivity* and *mobility* during pandemics by taking into consideration the local and the global transmission rate of the infection.

From the mathematical proof for our proposed models, we obtained that the reproduction number $R_0$ directly depends upon *social connectivity of individuals*, *number of connected locations* and *individuals mobility between locations* (and *degree of the individual* in complex networks model) which is in line with our simulation's results. This indicates that introducing *isolation* and *quarantine* is effective in fighting a pandemic crisis. Using the proposed model, we also simulated the real-world scenario by considering the COVID-19 cases in Estonia and Rhône-Alpes region in France. Simulation reveals that the mobility-based SIR model can be helpful to forecast the expected number of cases after some proportion of *isolation* and *quarantine* is introduced in society.

# 3. STUDYING SEGREGATION USING CDR

Segregation is defined as the degree of separation between two or more population groups and is considered detrimental for society's social well-being. Over many years, it attracted a lot of interest from the research community. The socio-economic structure and general stability of many developing countries have long been thought to be significantly influenced by segregation [BY08]. According to [Ale+03], the repercussions of segregation are not limited to developing countries, but the detrimental impact of segregation is more severe in countries with poor political and legal structures. As a result, strategies that promote integration and interaction in varied societies must be prioritized.

In the past, research on segregation has faced an issue of the availability of reliable data. Therefore, much of the previous research work relies on conventional government census data [Sil20]. Census data can contain physical settlement information but rarely record trends of social interaction, which are necessary to develop a thorough understanding of the essence of social interaction. In [Dev+14], the authors showed that physical settlement is dynamic and not necessarily coincident with formal status (e.g. students can live in a city that is not their formal residence; during summer, residents move away from main cities). In this sense, CDR data are more correct. Other limitations of using census data are that they are costly to gather and updated infrequently.

**Research Question.** In this work, we utilize CDR provided by one of Estonia's major telecom operators to research the complexities of social interaction and human behavior in order to understand segregation. The primary goal of using CDR data to anticipate segregation is to eliminate or replace costly and time-consuming censuses. In this chapter, we used descriptive techniques on CDR data to address the following research question: *How accurately societal segregation can be investigated using CDR data based on demographics such as gender, age-group, language, and location?*

The rest of the chapter is organized as follows. Next, we discuss the background. We then describe the dataset and its validation using census data in Section 3.2. Sections 3.3 and 3.4 present the results of our descriptive analysis of the dataset and we summarize our findings in Section 3.5.

## 3.1. Background

In this modern age, we have undergone a major transformation in the way people interact. Mobile phones have become one of the most significant and influential assets in our everyday lives. These devices have modified people's approach to communicate and, as a result, have influenced our society. These sensor-equipped devices are revolutionizing our economy, health care, social networking, and travel [GZH18].

Exploring call data records (CDR) generated by mobile phone use can reveal insights about human interaction behavior. Mobile phone usage creates data traces that can be exploited to determine a person's whereabouts. Such datasets can be used in various applications, including mobility patterns, travel demand, and regional dispersion of communications. Disease spread tracking, human travel dynamics and the identification of human congestion are also among other applications incorporating such large-scale data sets and analysis.

In [B P+15; EPL08], the authors showed that CDR data can provide valuable insights into the social structure of societies when analyzed using social network analysis. In [Onn+07], the authors identified the strong and weak ties between individuals. In another work, population density is calculated using CDR [Dev+14]. Some research has also been done using CDR data for identifying mobility patterns. For example, [Son+10], [GHB08] authors demonstrate that the human path is predictable and reproducible. In another work, the authors proposed a human mobility model and validated using a real dataset from New York and Los Angeles metropolitan areas [Isa+12]. In [AM05], the authors analyzed human behavior to find the movement pattern across various age groups.

A set of works also focused on explaining the various forms of segregation within societies using different datasets. In [Sim05] the authors studied racial discrimination using a statistical approach. They find that there is a shortage of statistical metrics to measure the extent of discrimination and that there is a common perception that discrimination is widespread and that both governmental institutions and common people should unite to reduce it. In other work, the authors also studied ethnic [CR07], residential [Qui02], and social segregation [RRT12].

In [AT12; JPF04], the authors claimed that four types of factors appear to contribute to segregation: discrimination, disadvantage, preferences, and social networks. In a different work, segregation is decomposed into two types: social segregation, as observed in interactions among people, and spatial segregation, as determined by the physical locations of people [BF13]. Furthermore, a framework is proposed to model and measure fine-grained patterns of segregation from large-scale digital data.

In another line of work, the authors studied segregation using segregation indices. These indices can be broadly categorized into non-spatial and spatial indices. The non-spatial indices are the indices in which the information is independent of all geographic considerations [BG95], and on the other hand, spatial indices are defined as those which directly or indirectly consider location information [Won93]. In our work, we use non-spatial indices as they are more suitable to answer our research question.

The most well-known non-spatial index is the dissimilarity index D [DD55a], which can be formulated as follows:

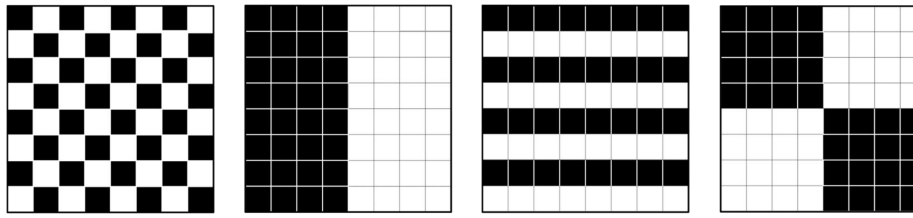
$$D = \frac{1}{2} \sum_i \left| \frac{p_{i,g}}{p_g} - \frac{p_{i,\bar{g}}}{p_{\bar{g}}} \right| \tag{3.1}$$

| S.No. | Index | Citation | Usage |
|---|---|---|---|
| 1 | Dissimilarity index | [DD55a] | Measure to compare the levels of residential segregation. |
| 2 | Freeman's index | [Fre78] | To quantify different levels of segregation in social networks. |
| 3 | Coleman's Homophily index | [Col58] | Coleman's homophily index computes homophily scores for each group defined by a vertex attribute. |
| 4 | Gini index | [DD55a] | Used to measure of statistical dispersion intended to represent the income inequality or the wealth inequality or the consumption inequality. |
| 5 | Centralisation index | [DD55b] | The centralization index has historically been used as a global spatial segregation index that quantifies segregation relative to the urban center of a region. |
| 6 | Exposure index | [Lie81] | It measures a given group's exposure to all other groups, including itself. |
| 7 | Neighbourhood sorting index (NSI) | [Jar96] | It is a unitless scale ranging from 0 to 1, with 1 representing full economic segregation and 0 representing perfect economic integration. |
| 8 | Typology for classifying ethnic residential areas | [PJF01] | TO measure residential ethnic segregation for topology. |
| 9 | Location Quotient (LQ) | [BC06] | Location quotient is useful in demographic studies because it shows what makes the region's demographics unique in comparison to its state and/or the nation. |

Table 5: Segregation indices.

where $i$ is the index of the spatial unit; g, $\bar{g}$ represent two population groups; $p_g$, $p_{\bar{g}}$ are total population of the two groups in the entire study region; $p_{i,g}$, $p_{i,\bar{g}}$ are the population of groups g, $\bar{g}$ in spatial unit $i$, respectively.

There exist a few similar measures like D (Equation 3.1), summarized in Table 5. The main properties of these measures are the following. First, they consider the residential system as consisting of different entities, each area being isolated from neighboring areas. Second, they focus on a global summary of a city or region, assuming that spatial relations are consistent across that area.



(a) (b) (c) (d)

Figure 13: The Checkerboard Problem.

The 'Checkerboard problem' is an example that is typically used to investigate the properties of various segregation measures [Won99; Won02; Daw04; Har16] is shown in Figure 13. In most people's perceptions, the four arrangements reflect varying levels of segregation. However, the values of D are the same, i.e., it does not differ between different spatial arrangements. In fact, the value of D will still be 1 as long as one population group is inhabited exclusively in each spatial unit [Whi83; Mor91; Won93].

Among the mentioned segregation indices, we consider two commonly used indices for calculating segregation [Mel20]. These indices are *Freeman's segregation index (FSI)* and *homophily index (HI)*.

### 3.1.1. Freeman's Segregation Index (FSI)

In [Fre78], the author proposed a segregation index called the Freeman Segregation Index (FSI) that has been used for understanding segregation in social interaction. According to FSI, if a given attribute (group label) is not applicable to social connections, then connections should be randomly distributed with respect to the attribute. Thus, the disparity between the number of cross-group ties expected by chance and the number observed is used for measuring segregation.

**Calculating FSI value:** Let us consider a network with static attribute groups $A$ and $B$ (of relative size $N_A$ and $N_B$ with $N_A + N_B = 1$) distributed among nodes uniformly at random and independently of the network structure, such that there is a fraction $P_{AB} = P_{BA}$ of edges between groups, and fractions $P_{AA}$, $P_{BB}$ within each group ($P_{AA} + P_{AB} + P_{BB} = 1$). The *FSI* can be measured using the following formula:

$$FSI = 1 - \frac{X}{E(X)} \tag{3.2}$$

where, $X$ is the proportion of between group ties and $E(X)$ is the expected proportion of random ties. The $X$ can be calculated using the formula 3.3.

$$X = \frac{P_{AB}}{P_{AA} + P_{AB} + P_{BB}} \tag{3.3}$$

### 3.1.2. Homophily Index (HI)

Homophily is the tendency of individuals to interact and associate with other individuals. In the past, homophily has been studied in great detail in numerous works [Asi+20; MSC01; Kan78; Gil+15]. These studies indicate that the similarity is correlated with the connection among individuals and can be categorized based on age, gender, class, ethnicity [Fu+12; Sho+12; ST17; SJ18], etc. In this work, we use the Coleman homophily index (*HI*) [Col58] for comparison with OBI since *HI* is commonly used to compare the homophily of groups with different sizes by normalizing the excess homophily of groups by its maximal value [Col58].

**Calculating *HI* value:** Considering the notations defined in previous section. In the case of two attribute groups, the probability that a random edge from a node in a group $A$ leads to a node in group $A$ is defined as:

$$T_{AA} = \frac{2P_{AA}}{2P_{AA} + P_{AB}} \tag{3.4}$$

Similarly, we can write an equation for $T_{BB}$. The *HI* value for group $A$ ($HI_A$) and $B$ ($HI_B$) can be calculated using

$$HI_A = \frac{T_{AA} - N_A}{1 - N_A} \tag{3.5}$$

$$HI_B = \frac{T_{BB} - N_B}{1 - N_B} \tag{3.6}$$

The range for both $HI_A$ and $HI_B$ is from -1 to 1, where -1 for $HI_A$ means that group $A$ individuals only connect with group $B$ individuals (only in between groups connections), whereas 1 for $HI_A$ means that group $A$ individuals only connects with group $A$ individuals (only within-group). To conclude, we can say that *FSI* index measure segregation. On the other hand, *HI* index measure both segregation and inclination simultaneously.

The segregation study [SA14] is the closest to our work in which the author studied the temporal variation of ethnic segregation in the city of Tallinn, the capital of Estonia. Their findings revealed that segregation is significantly lower on workdays and during the summer holidays. The contribution of our work is as follows: In our work, we perform an exploratory analysis that uses communication links and social network information of individuals to measure segregation. In particular, this work utilizes segregation indices along with descriptive analysis techniques on CDR data to assess gender, age, language, and county segregation.

## 3.2. Dataset Description

This study utilizes anonymized call data records (*CDR*) shared for research purposes by one of Estonia's leading mobile operators. The overview of the dataset

is provided in Section 2.5.1. However, in this section, we discuss the data in more detail. Please note that each call record contains information about the caller and called (counterpart of the caller) user.

In addition, the *gender* of the user, *year of birth* and preferred language of communication are provided in the dataset. It is assumed that the preferred language (*Estonian*, *Russian*, or *English*) selected as the language of communication by the user with the operator is the user's first language. This is the language that is used for billing, special offers, and technical messages by the network operator. As bilingualism is not very common in Estonia [SA14], we presume that the people who chose the Russian language are members of the Russian-speaking minority in Estonia and similar for other languages.

For preprocessing, we performed quantile filtering to remove users with fewer connections. Here, the upper range quartile is 95% and the lower range quartile is 5% for outlier removal. It is to be noted that not all users have additional details (gender, language, and county) in the dataset. For example, *gender* information is available for 130,988 users with 61,933 males and 69,055 females. Table 6 summarizes statistics for this dataset.

**Encoding users' age.** Centered on the official age-group categorization suggested by *Europe-Bureau* and *Statistics Estonia* [Est], we categorize the age of users into the following five groups: (1) 5-14 years: Children; (2) 15-24 years: Early working age; (3) 25-54 years: Prime working age; (4) 55-64 years: Mature working age; and (5) 65+: Elderly.

This categorization is useful to analyze and report findings for specific age-groups which was difficult using provided continuous value for age in the CDR dataset. In Table 6, the *Age-Groups* row displays the distribution of users in the dataset according to their *age-group*. E.g., for the age-group *(54,64)* with value: *21,427* means that there exist 21,427 individual users in the dataset that belongs to age-group (54,64).

***CDR* dataset validation using census dataset.** As *CDR* data contains call records and might not represent the real population of Estonia, we need to validate that the *CDR* data is indeed an acceptable representative of Estonia's actual population. To do so, we compare the distribution of users in the *CDR* based on four features (*county, language, age-group* and *gender*) with the actual Estonian population in Figure 14, where, the x-axis represents percentage and y-axis represents users' features from top to bottom namely *county* (written in black), *language* (highlighted in purple), *age-group* (highlighted with **green**) and *gender* (highlighted in blue). The sum of all percentages based on each feature is 100 percent for both *CDR* and actual Estonian population separately. For each feature value, pink dots represent the percentage of users in the *CDR* dataset and sea green dots represent the percentage of people in the actual Estonian population. Furthermore, for each feature value, the difference between *CDR* and actual population percentage is calculated using the formula 3.7.

| Parameters | Value |
|---|---|
| Time period | 8 May'17 to 13 May'17 |
| Call Records | 12,179,970 |
| Unique Users | 1,175,919 |
| **Gender** | |
| Male | 61,933 |
| Female | 69,055 |
| **Age-Groups** | |
| (0,14] | 76 |
| (14,24] | 1,196 |
| (24,54] | 83,028 |
| (54,64] | 21,427 |
| (64,100] | 12,323 |
| **Languages** | |
| Estonian | 102,545 |
| Russian | 14,882 |
| English | 236 |
| **Locations** | |
| Harju | 507,365 |
| Hiiu | 6,959 |
| Ida-Viru | 87,212 |
| Järva | 18,157 |
| Jõgeva | 23,125 |
| Lääne | 18,377 |
| Lääne-Viru | 44,787 |
| Pärnu | 55,873 |
| Põlva | 21,083 |
| Rapla | 24,108 |
| Saare | 25,374 |
| Tartu | 132,888 |
| Valga | 17,528 |
| Viljandi | 33,767 |
| Võru | 28,405 |

Table 6: CDR data statistics.

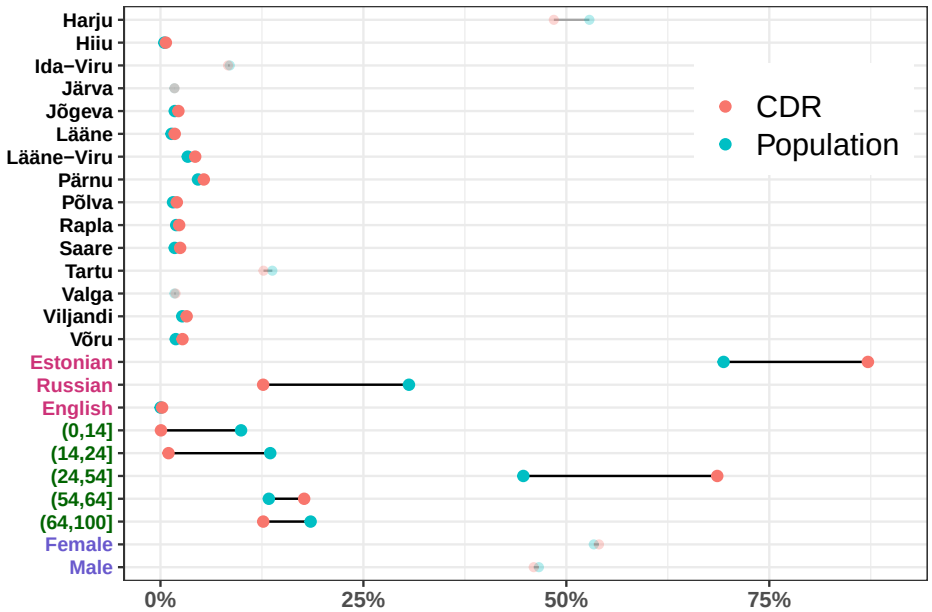



Figure 14: Comparison of the actual population of Estonia and users in *CDR* based on four features.

$$Difference = \frac{|CDR\ \% - Population\ \%|}{min(CDR\ \%, Population\ \%)} * 100 \tag{3.7}$$

The differences greater than 10 percent are written and highlighted with the text color. The features and the difference interpretation from top to bottom on the y-axis are as follows:

***County***: Estonia has 15 counties, with *Harju* county, which includes the capital *Tallinn*, being the most populous, *Tartu* county being the second most populous, and *Hiiu* county being the least populous. Figure 14 shows that the gap between CDR users and the overall population in the top four populous counties (*Harju, Tartu, Ida-Viru* and *Pärnu*), which account for approximately 80% of Estonia's total population [Sta21], is less than 10%. All counties with a population and CDR disparity of more than 10% cover nearly 14% of Estonia's population. As a result, we can infer that CDR data is a fair representation of the real Estonian population in terms of county.

***Language***: As previously mentioned, the preferred languages of interaction choices in the CDR dataset are *Estonian*, *Russian*, or *English*. Figure 14 reveals that the percentage of the Estonian-speaking population in CDR data is higher than the actual Estonian population. The Russian-speaking population, on the other hand, exhibits a distinct pattern of behavior. However, CDR data can be used to calculate language segregation.

***Age-groups***: According to age-group, mobiles are often used by prime working

age users (i.e., (24,54)), mature working age users (i.e., (54,64)), and elderly users (i.e., (64,100)). These three age-groups account for nearly 99 percent of the *CDR* dataset and 76 percent of the total Estonian population. Mobile usage percentages for other age-groups are lower than their actual population. From this, we can infer that in the *CDR* dataset, the representation of the prime working age users is significant as compared to their actual population in Estonia. As a result, we can suggest that the findings of the prime working age users can be considered accurate with reasonable confidence. The same is valid for mature working age and elderly users. On the other hand, the representation of children and early working age-group (i.e., (14,24)) in *CDR* is minimal or non-existent in comparison to the actual population, making it difficult to analyze these age-groups. This shows that mobile phone use is common after a certain age.

***CDR* dataset availability and Ethical concern.** Our dataset is partly location data, and it can not be shared due to privacy concerns. The dataset is owned by our collaborator at the University of Tartu and is accessible for research purposes after signing the NDA. For privacy, the dataset is anonymized at two levels, so that specific persons cannot be identified.

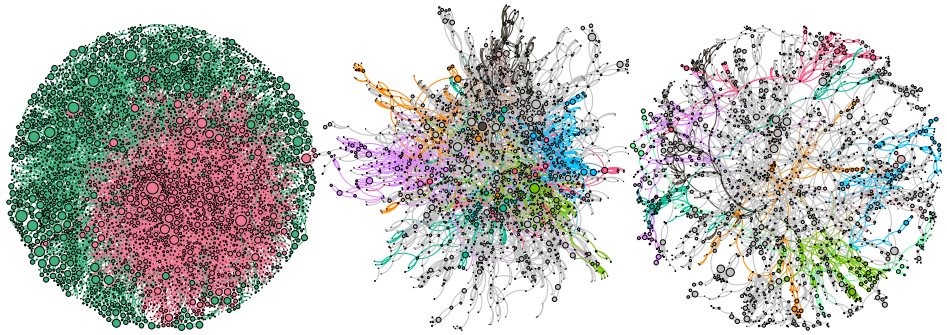
**Implementation.** We used Python programming language for the implementation. For all the experiments, we use Intel(R) Xeon(R) Gold 6246R CPU @ 3.40GHz with 128 GB RAM.

## 3.3. Segregation using Social Interaction

In this section, we understand gender segregation, performed by exploring the social network interactions of the CDR. We create a directed network that represents the call connections among users where an edge ($u \longrightarrow v$) is formed if a user $u$ has called user $v$. Figure 15(a) shows the CDR network, where each node is color-coded based on gender. The red nodes represent male users, and the green nodes represent female users. Links between users are also color-coded. Links that originate from males are colored red (i.e., calls from male to male; and male to female), and similarly links that originate from females are colored green (i.e., calls from female to female; and female to male).

Furthermore, we employ the well-known *PageRank* algorithm [Pag+99] to identify the centrality of nodes in the network. *PageRank* reflects the importance of a node in terms of its influence in the network. For example, an individual with a higher Pagerank could reflect its bigger social influence in propagating a piece of information in the network. In Figure 15(a), the size of the node reflects the Pagerank of the node.

Table 7 provides the statistic of the network. The lower value of the average clustering coefficient and edge density can be used to infer that the network is sparse. The values of these metrics further indicate that the network is spread out and the transmission of information would possibly take longer to transmit throughout the network. From the values of strongly and weakly connected com-



(a) CDR Network (b) Male Network (c) Female Network

Figure 15: Users network formed using CDR data.

ponents' size and the number of components, we can conclude that there are a large number of small communities. The value of reciprocity indicates that only 24.6% individuals have mutual interests with each other.

| Network Properties | Value |
|---|---|
| Nodes | 1,175,919 |
| Edges | 4,664,821 |
| Average in-degree | 3.97 |
| Edge density | 3.37e-6 |
| Number of triangles | 2,909,058 |
| Average clustering coefficient | 0.096 |
| Strongly & weakly connected component size | 1,170,309 |
| Number of components | 14870 |
| Reciprocity | 0.246 |

Table 7: Network statistics of users.

Based on color-coding, we can easily observe clusters of males and females in the CDR network (Figure 15(a)). For further studying the segregation, we study the giant component of the males-only (see Figure 15(b)) and the giant component of the females-only (see Figure 15(c)) networks separately. For creating the males-only network, we drop all the caller and callee ids, which belong to females. Similarly, we drop all the caller and callee ids which are males to create a females-only network. For better comparisons of these networks, we report the properties of each of these networks in Table 8. Although this creation of males-only and females-only networks is synthetic, nevertheless it can provide some significant information about segregation in these networks.

In Figure 15(b) and 15(c) users are grouped into communities based on modularity values (0.803 for the males-only and 0.913 for females-only network). The higher value of modularity indicates that the females-only network has more clusters but these clusters are densely connected within themselves as also supported by the higher average clustering coefficient of the females-only network, which

| | Male | Female |
|---|---|---|
| Nodes | 40,711 | 45,931 |
| Edges | 76,188 | 64,824 |
| Edge density | 4.6e-5 | 2.9e-5 |
| Average clustering coefficient | 0.097 | 0.138 |
| Average path length | 10.46 | 24.65 |
| Diameter | 136 | 203 |
| Modularity | 0.803 | 0.913 |

Table 8: Statistics comparison of the giant component of male and female network.

suggests that females bonds in smaller groups, but these groups are tightly connected compared to their males' counterparts. The higher value of edge density for a males-only network suggests that males, in general, have more connections compared to females. In addition, the smaller diameter and average path length values of the males-only network indicate that the network is compact compared to the females-only network, which is more spread out. These males-only and females-only network metrics point out that the transmission of information is fast in the males-only network compared to the females-only network.

## 3.4. Demographics Segregation using FSI and HI

This section focuses on understanding interaction in the CDR data by exploring various users' demographics, including gender, age, language, and location. We also calculated segregation using FSI and HI by considering the mentioned demographics. In Section 3.4.1 and 3.4.2, we study gender and age-groups segregation. In Section 3.4.3 and 3.4.4, we study language and location segregation.

### 3.4.1. Measuring Gender Segregation in Estonia

First, we explore the individuals interaction based on gender in Estonia using the CDR data. The gender segregation calculated using FSI is 0.267, and HI is 0.158 for males and 0.1 for females (shown in Table 9), indicating its presence in Estonia. Next, we analyze the data in-depth to explore gender segregation based on age group, language, and county.

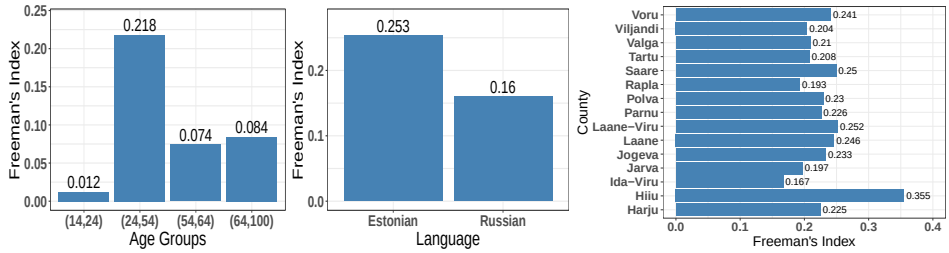
***High Gender Segregation In Prime Working Age***. Here, we analyze gender segregation based on age-groups. Figure 16a compares the FSI value for various age groups based on gender. We find that *prime working age* group, i.e., (24,54) is highly segregated followed by *elderly* group, i.e., (64,100). Please note that we excluded (5,14) age-group for this comparison because of insufficient data with gender information.

The comparison of males and females calling pattern based on age groups also highlights that males of *early working age*, *mature working age* and *elderly*

| Parameters | FSI | HI |
|---|---|---|
| **Gender** | | |
| Male | 0.267 | 0.158 |
| Female | 0.267 | 0.1 |
| **Age-Groups** | | |
| (0,14] | 0.374 | 0.032 |
| (14,24] | 0.206 | 0.172 |
| (24,54] | 0.385 | **0.561** |
| (54,64] | 0.322 | 0.398 |
| (64,100] | **0.432** | 0.332 |
| **Languages** | | |
| Estonian | 0.738 | 0.704 |
| Russian | **0.742** | **0.751** |
| English | 0.267 | 0.094 |
| **Locations** | | |
| Harju | 0.712 | 0.798 |
| Hiiu | 0.812 | 0.724 |
| Ida-Viru | **0.856** | **0.838** |
| Järva | 0.762 | 0.537 |
| Jõgeva | 0.758 | 0.587 |
| Lääne | 0.773 | 0.608 |
| Lääne-Viru | 0.778 | 0.695 |
| Pärnu | 0.796 | 0.705 |
| Põlva | 0.754 | 0.591 |
| Rapla | 0.755 | 0.570 |
| Saare | 0.828 | 0.766 |
| Tartu | 0.727 | 0.714 |
| Valga | 0.772 | 0.635 |
| Viljandi | 0.793 | 0.681 |
| Võru | 0.777 | 0.710 |

Table 9: Demographics segregation using FSI and HI.

call more to females of the same age group, but at the same time, they maintain strong connectivity with males of other age groups as well. On the other hand, most females' calls remain within the same age group females, although their connectivity with females of different age groups is relatively weak compared to males. Based on the calling behavior between age groups, we can conclude that males are more socially connected with other males and females than females who prefer to communicate with females of the same age group. To explore further, next, we examined language-based gender segregation.




(a) Age-Group (b) Language (c) County

Figure 16: Gender segregation using FSI.

***Higher Gender Segregation Among Estonian Speaking Individuals***. In Estonia, Estonian is spoken by the majority population, followed by Russian. Some peoples' preferred language is also English. We analyzed the segregation based on languages in our analysis. In terms of gender segregation, the Estonian-speaking population is more segregated than the Russian-speaking individuals (see Figure 16b). Please note that here we excluded English-speaking population for this comparison because of insufficient users with gender information.

Based on call activity among different language speakers, we observe that Russian-speaking individuals call comparatively higher than Estonian-speaking individuals. Additionally, call patterns comparison based on age groups and language indicates that Estonian speaking females are more inclined to call other Estonian speaking females of the same age group. This shows that Estonian-speaking females are more segregated compared to others. To explore further, we also investigate gender segregation based on counties.

***Gender Segregation In Estonian Counties***. Figure 16c compares the FSI values for various counties based on gender. We observe that *Hiiu* county is the most segregated and *Ida-Viru* is the least segregated county based on gender. On further analysis, we observe that in all counties, both males and females are inclined towards the same gender, which also inclined with previous research [GSA21].

### 3.4.2. Measuring Age-Group Segregation in Estonia

We calculated the *FSI* and *HI* index for age-groups segregation as shown in Table 9. We find that the two age-groups: (24,54), and (64,100) are more segregated than others. Next, we explore the reasons for this segregation by analyzing the CDR data at the county level.

We observe that *FSI* index for all age-groups is less than the average *FSI* index in *Voru*, *Polva*, *Saare* and *Rapla*. We can also observe that in *Jogeva*, *Laane* and *Laane-Viru*, the segregation index for age-group (24-54) is approximately near to the average segregation. The interesting observation to be noted here is that the age-group (24-54) is mostly segregated in county *Parnu*. Another interesting observation can be that in *Hiiu* county, only elderly (i.e., (64,100)) individuals are more segregated than the mean *FSI* value. Therefore, we further analyzed this

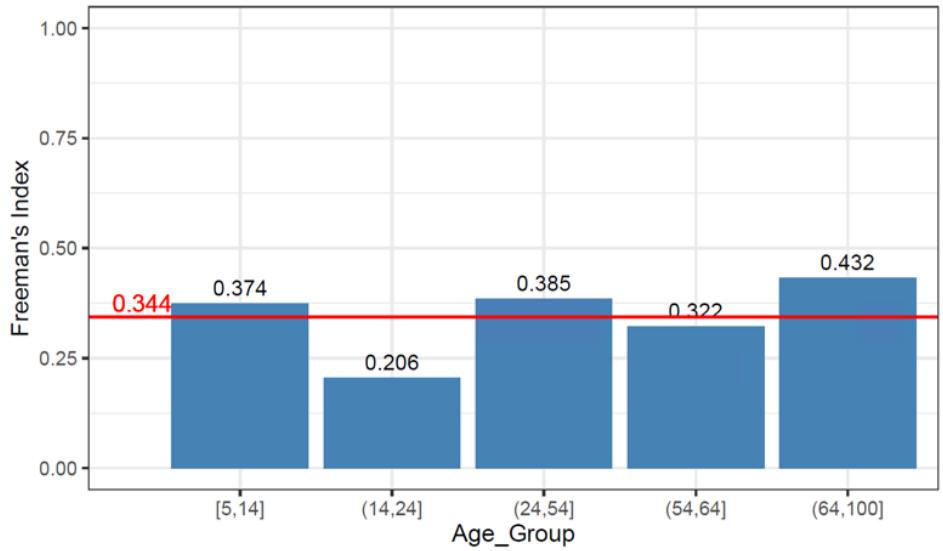



Figure 17: The *FSI* values for age-groups segregation in Estonia.

age-group for Hiiu county using social network analysis and observe that elderly individuals prefer to connect other elderly individuals. We also observe that there are many elderly pairs which are only connected to each other.

Finally, we can conclude that the *prime working* individuals (i.e., (25-54) age-group) and *elderly* (i.e., (64-100) age-group) are more segregated than average segregated individuals in Estonia based on age-group.

### 3.4.3. Measuring Language Segregation in Estonia

Here, we explore the individuals interaction based on their preferred language of communication in Estonia using the CDR data. We calculated the *FSI* and *HI* index for language segregation as shown in Table 9. The two language-speaking individuals (Estonian-speaking and Russian-speaking) are more segregated than the mean segregation (also see Figure 18). As we have already shown in Table 6 that our dataset contains only 256 English-speaking individuals, it is difficult to infer that this language-group is segregated. But, we can conclude for Estonian-speaking and Russian-speaking individuals with reasonable confidence that they are more segregated. In the next section, we explore the language segregation in detail by analyzing the CDR data at county level.

### 3.4.4. Measuring County Segregation in Estonia

This section first explains the segregation in Estonia based on county using CDR data. Further, we explore four different cases by removing large municipalities in different counties to measure their effect on segregation. The reason to remove these large municipalities is explained in the following subsections. The four cases are as follows:

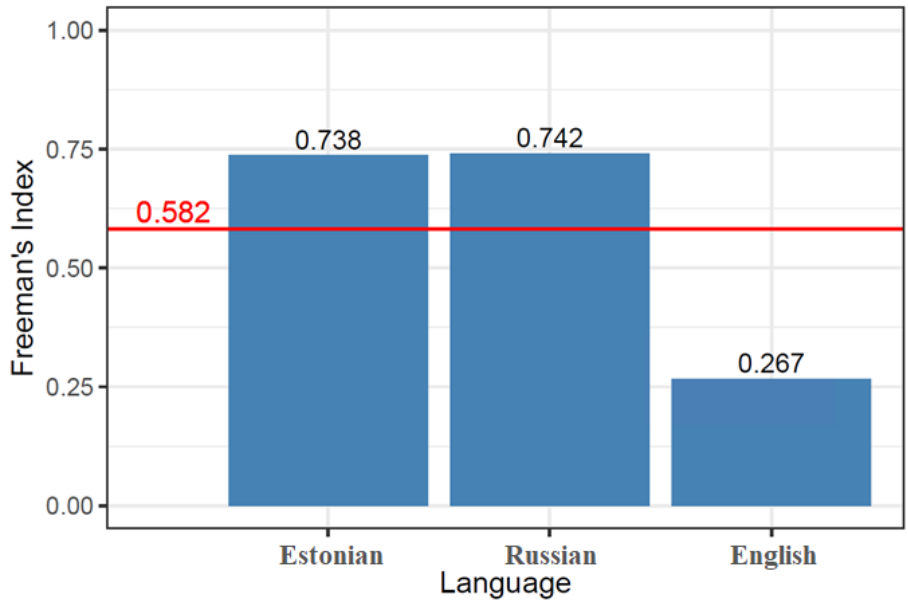



Figure 18: The *FSI* values for language segregation in Estonia.

1. Segregation in Estonian counties after removing **Narva** city. Please note that *Narva* city is a part of *Ida-Viru* county.
2. Segregation in Estonian counties after removing **Pärnu** city. Please note that *Pärnu* city is a part of *Pärnu* county.
3. Segregation in Estonian counties after removing **Tartu** city. Please note that *Tartu* city is a part of *Tartu* county.
4. Segregation in Estonian counties after removing **Tallinn** city. Please note that *Tallinn* city is a part of *Harju* county.

**Segregation in Estonian counties.** We calculated the *FSI* index for segregation based on county, and find that the mean *FSI* value for county segregation is 0.777. This mean segregation value for counties suggests that most individuals prefer to call within same county. Furthermore, the individuals in six counties (*Hiiu, Ida-Viru, Lääne, Pärnu, Saare* and *Viljandi*) are more segregated than the mean segregation (see Figure 19). We also observe that individuals in *Harju* county and *Tartu* county are less segregated than others. One of the reasons for this less segregation can be the diversity of the population in these counties.

We summarized all our *FSI* values based on *gender*, *age-group*, *language* and *county* in Table 10. We also highlighted segregation using red text color. For example, (1) based on *language*, Estonian-speaking and Russian-speaking individuals are more segregated; (2) For county *Ida-Viru*, prime working age-group (i.e., (24,54)) has *FSI* value as 0.48 which is more than average *FSI* value based on *age-group*, therefore it is shown in red text color.

**Segregation in Estonian counties without Narva city.** We measured the *FSI*

| | **Gender** | **Age Group** | | | | | **Language** | | |
|---|---|---|---|---|---|---|---|---|---|
| **County** | | *(5-14)* | (14-24) | *(24-54)* | (54-64) | *(64-100)* | *Estonian* | *Russian* | English |
| Harju | 0.224844 | 0.132 | 0.1793 | *0.414* | *0.379* | *0.408874* | *0.6416* | *0.6428* | 0.1838 |
| *Hiiu* | *0.35499* | 0 | 0 | 0.333 | 0.209 | *0.4699* | 0.239 | 0 | 0.107 |
| *Ida-Viru* | 0.1672 | 0.12 | 0.23 | *0.48* | *0.45* | *0.44* | *0.652* | *0.6501* | 0.2548 |
| Järva | 0.1969 | 0 | 0 | *0.383* | *0.3987* | 0.2934 | *0.6493* | *0.6756* | 0.42 |
| Jõgeva | 0.2333 | 0 | 0.066 | *0.352* | 0.2985 | 0.2766 | 0.48 | 0.5297 | 0 |
| *Lääne* | 0.2455 | 0 | 0.073 | *0.3529* | 0.2895 | 0.32388 | *0.7933* | *0.7933* | 0 |
| Lääne-Viru | 0.2523 | 0.311 | 0.1436 | *0.3704* | *0.34786* | 0.3394 | 0.546 | 0.5571 | 0.1986 |
| *Pärnu* | 0.221536 | *0.51* | 0.1521 | *0.6355* | 0.3392 | *0.3685* | 0.5387 | 0.5617 | 0.1246 |
| Põlva | 0.22918 | 0 | 0.0889 | 0.294 | 0.2139 | 0.337 | 0.4192 | 0.4192 | 0 |
| Rapla | 0.19347 | 0 | 0.048 | 0.3148 | 0.2412 | 0.293 | 0.5262 | 0.5262 | 0 |
| *Saare* | 0.2504 | 0 | 0.044 | 0.315 | 0.254 | 0.2688 | 0.2943 | 0.024 | *0.5663* |
| Tartu | 0.20787 | 0.3389 | 0.2358 | *0.393* | *0.3454* | *0.3727* | 0.4516 | 0.459 | 0.154 |
| Valga | 0.21 | 0 | *0.407* | *0.377* | 0.3168 | *0.3533* | 0.436 | 0.436 | 0 |
| *Viljandi* | 0.20438 | 0.0816 | *0.4868* | 0.3358 | 0.2458 | 0.3193 | 0.22877 | 0.2705 | 0.1502 |
| Võru | 0.2407 | 0 | 0.0511 | 0.3 | 0.286 | 0.2758 | 0.0245 | 0 | 0 |

Table 10: *FSI* values based on different combinations of *gender*, *age-group*, *language* and *county*.

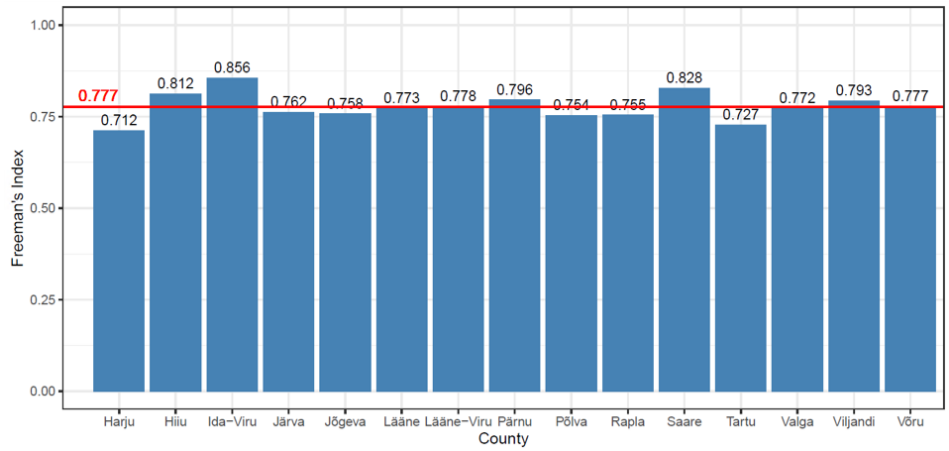



Figure 19: The *FSI* values for counties segregation in Estonia.

values for each county after removing the *Narva* city of *Ida-Viru* county. The reason for removing this particular city is the fact that the majority of its population has Russian-speaking individuals. In particular, 95.7% of the population of *Narva* city are native Russian speakers, and 87.7% are ethnic Russians [Est]. We calculated *FSI* index after removing *Narva* city and find that the segregation of *Ida-Viru* county decreased from 0.856 to 0.843. Hence, we can conclude that the Russian-speaking population in *Narva* city of *Ida-Viru* county is more segregated than the Estonian-speaking population. We can further infer that county segregation and language segregation are correlated with each other.

**Segregation in Estonian counties without Pärnu city.** In this section, we measured the *FSI* values for each county after removing the *Pärnu* city of *Pärnu* county. The reason for removing this particular city is the fact that the majority of its population has Estonian-speaking individuals (83%) [Est]. On the other hand, the Russian-speaking, Ukrainian-speaking, and other language-speaking individuals are 12.8%, 1.7%, and 2.5% respectively [Est]. We calculated *FSI* index after removing *Pärnu* city and find that the segregation of *Pärnu* and *Haju* county increased from 0.712 and 0.796 to 0.718 and 0.805 respectively. Hence, we can conclude that the Estonian-speaking population in *Pärnu* city of *Pärnu* county is less segregated than other language-speaking populations. In other words, we see a similar behavior as earlier that the Russian-speaking population in *Pärnu* county is more segregated. Therefore, we can infer with reasonable confidence that county segregation and language segregation are correlated with each other.

**Segregation in Estonian counties without Tartu city.** Here, we measured the *FSI* values for each county after removing the *Tartu* city of *Tartu* county. The reason for removing this particular city is the fact that it is the second largest city of Estonia, after Estonia's political and financial capital Tallinn. *Tartu* is often considered the intellectual center of the country [inf; Bou11; Che], especially since it is home to the nation's oldest and renowned university, the University of

Tartu. *Tartu* city itself is also the oldest city of Estonia [MR13].

The majority population of *Tartu* city speaks Estonian (80%) [Est]. On the other hand, the Russian-speaking and other language-speaking individuals are 14% and 6% respectively [Est]. We calculated the *FSI* index after removing *Tartu* city and find that it affects many counties. The *FSI* value for six counties (*Harju*, *Jõgeva*, *Põlva*, *Tartu*, *Valga* and *Võru*) have increased. On the other hand, the *FSI* value for four counties (*Hiiu*, *Jõgeva*, *Rapla* and *Saare*) have decreased. Hence, we can conclude that individuals in *Tartu* city of *Tartu* county have a greater impact on other counties either by increasing or decreasing their *FSI* values. Interestingly, we also observe that removing *Tartu* city has no impact on *Ida-Viru* and *Pärnu* counties which shows that their segregation is independent of *Tartu* city. This is also true for other counties which are not affected by the removal of *Tartu* city. Therefore, we can conclude that *Tartu* city is well connected with other counties of Estonia.

**Segregation in Estonian counties without Tallinn city.** Next, we measure the *FSI* values for each county after removing the *Tallinn* city of *Harju* county. The reason for removing this particular city is that it is the capital and the most populous city of Estonia. *Tallinn* is also the main financial, industrial and cultural centre of Estonia [inf]. *Tallinn* city also have approximately equal number of Estonian-speaking (50.1%) and Russian- speaking (46.7%) individuals [Est].

We calculated *FSI* index after removing *Tallinn* city and find that it affects all counties. The *FSI* value for eight counties (*Harju*, *Hiiu*, *Ida-Viru*, *Lääne*, *Lääne-Viru*, *Pärnu*, *Rapla* and *Saare*) have increased. Hence, we can conclude that these eight counties are very well-connected to the *Tallinn* city individuals. On the other hand, the *FSI* value for seven counties (*Järva*, *Jõgeva*, *Põlva*, *Tartu*, *Valga*, *Viljandi* and *Võru*) have decreased. This indicate that these counties are well-connected with other cities and counties other than *Tallinn* city.

Based on our analysis of the county, we can infer that language geography plays a vital role in segregation. Individuals prefer to connect with individuals that speak the same language and reside in the same locality (city or county).

## 3.5. Summary

This work analyzed countrywide CDR data with user demographics to study societal segregation in Estonia. In particular, we explored demographics such as *gender*, *age-group*, *language*, and *county*. We employed two segregation indices, Freeman Segregation Index (FSI) and Homophily Index (HI), to measure segregation. The highlights of our study are the following:

1. We validated our CDR dataset using census datasets and showed that CDR datasets can indeed fairly represent the actual population, and all the segregation study findings on a quality CDR dataset can be considered correct with reasonable certainty. Therefore, findings from CDR datasets can be

used by government agencies to make target policies in particular for the needed segregated group.

2. *Gender segregation:* The individuals are segregated based on gender. The traces of gender segregation can be found in connectivity among age-groups, preferred language of communication, and calling location.
3. *Age-group segregation:* The *prime working* individuals (i.e., (25-54) age-group) and *elderly* (i.e., (64-100) age-group) are more segregated than average segregated individuals in Estonia. The results using network analysis techniques further supported our findings.
4. *Language segregation:* The Estonian-speaking and Russian-speaking individuals are segregated based on language. With reasonable confidence, we can say that they prefer to call the same language individuals.
5. *County segregation:* We find that neighborhood and language play a vital role in the county's segregation. Individuals prefer to connect with other individuals that speak the same language and reside in the same locality (city or county).

# 4. PREDICTING SOCIO-ECONOMIC WELL-BEING USING MOBILE APPLICATIONS DATA

Socio-economic indicators provide context for assessing a country's overall condition. These indicators contain information about education, gender, poverty, employment, and other factors. Therefore, reliable and accurate information is critical for social research and government policing. Most data sources available today, such as censuses, have sparse population coverage or are updated infrequently. Nonetheless, alternative data sources, such as call data records (CDR) and mobile app usage also called digital data, can serve as cost-effective and up-to-date sources for identifying socio-economic indicators.

*Call* and *Digital* data has recently been recommended as an alternate source for determining socio-economic status [Uca+21; Mar+17; Mar+20; Sin+19]. Researchers have been able to create socio-economic prediction models with unprecedented temporal and geographical resolutions owing to the increasing usage of mobile devices, social media, and the expanding availability of ubiquitous satellite data [Sot+11; GZZ19; DRZ19; BDK15]. Human mobility and social contacts were shown to be linked to higher income in data from mobile phones and social media [Blu16; Pap+16; Ste+17; Llo+15].

**Research Question.** Considering the literature on CDR [GSA21b; GSA21a] and past research to predict socio-economic features using mobile service consumption data [Uca+21], we show that mobile application usage patterns would be able to predict socio-economic features. The use of mobile app data rather than merely CDR, which has been done by the majority of previous research, is the highlight of this study. The primary goal of using mobile digital data to anticipate socio-economic features is to eliminate or replace costly and time-consuming censuses. This leads to our research question, that is *how accurately we can predict socio-economic conditions using mobile applications usage patterns*.

The rest of the chapter is organized as follows. Next, we explain three datasets utilized in this study: mobile app traffic, geographical, and socio-economic features in Section 4.1. Then, the areal interpolation of IRIS is discussed in Section 4.2. In Section 4.3, mobile app usage patterns are extracted. We cover the result and explainability in Section 4.4. Finally, conclusions are outlined in Section 4.5.

## 4.1. Dataset Description

We conduct a novel and large-scale investigation using nationwide and multi-source datasets. We utilized three different datasets:

### 4.1.1. Mobile Applications Usage

Mobile applications usage data of size 2TB (terabyte) spanning more than 2.5 months by approximately 30 million subscribers distributed over more than 550,000 $km^2$ and served by over 25,000 base stations in a major European country with one of the world's largest economy, France [Ban].

Table 11 shows the screenshot of mobile application usage data. Here, *LocInfo* is a alphanumeric column is used to map the record to the latitude and longitude information of the record which is available in another table. The *PortApp* is used to map the application name from another table. Next, we have timestamp as numeric value. Lastly, *Volume up* and *Volume down* which represents the bytes of data uploaded and downloaded while using the application.

| LocInfo | PortApp | Ts | VolumeUp | VolumeDn |
|---|---|---|---|---|
| 0102f8100068145e | 65759 | 1552753320 | 6901.1599 | 21441.916 |

Table 11: Mobile application usage dataset screenshot.

Therefore, the data consist of time-stamped and georeferenced records of the mobile traffic generated by different applications, such as YouTube, Facebook, or Netflix—including device-specific ones such as Apple Store (run by iOS devices) or Google Play (run by Android devices). The data were collected throughout France and aggregated at the antenna level. Because of their volume and scattered nature, some traffic from different applications was aggregated into common categories such as mail, gaming, news (mainly newspaper outlets), or downloading.

*Mobile applications categorization.* We manually categorized the top 99 percentile mobile applications in terms of traffic load into 19 classes. We chose the top 99 percentile mobile apps since only a small number of apps generate a significant volume of data according to the power law [Mar+17; Sha+11]. The idea of categorizing mobile apps is also explored in past research [Liu+16]. The mobile apps that are selected include heterogeneous apps and encompass *Advertising* (e.g., Web Advertising), *Android download* (e.g., Google Play Store), *Apple cloud* (e.g., iCloud storage), etc. Table in the Appendix A contains the detailed information about application categorization.

**Ethical concern.** Our research follows strong ethical standards. Orange, France provided the data and was responsible for data collection, processing, and storage. They follow the rules of the European Commission's General Data Protection Regulation (GDPR). All these privacy-related activities were also monitored by the Orange Data Privacy Officer and the *CNIL*, the French governmental body responsible for safeguarding privacy in the use of personal data. In this study, we used aggregated data among thousands of people at the antenna level, which does not pose a privacy threat and does not qualify as personal data.

### 4.1.2. Socio-Economic Features

Socio-economic features covering economic status, population, and education information for around 49,000 IRIS[1].

*Economic status:* For economic features, we used the **Revenus, pauvreté et niveau de vie en 2018 (Iris)** dataset [INS18d; Ins], which contains a complete description of the income distribution deciles for residential IRIS zones. Features for areas with less than 1000 inhabitants are not shared by the IGN for privacy reasons. After filtering out IRIS zones with missing values, we have economic status for 9,145 IRIS zones (out of 48,931), which encompass all the main urban areas of France.

The income that we work on is the disposable income that means people declare their total incomes once a year (salary, allowances, rent, financial products, etc.) and they pay taxes on their total incomes. The disposable income is what they keep after redistribution. Next, we select three features Poverty, Median income, and Gini Index from the economic status which are vital for understanding economic status of the IRIS. The selected features along with their definition are shown in Table 12.

*Population information:* For the population structure, we utilised the **Population en 2018** [INS18c] dataset, which contains a description of the population structure by age group, gender and other factors, such as socio-professional category and immigration. The selected features from the population information based on correlation are shown in Table 12. For population information, we have data for 45,508 IRIS zones (out of 48,931).

*Education information:* For education statistics, we used the **Diplômes - Formation en 2018** [INS18b] dataset, which contains the academic level information (BEPC, BAC, SUP, CAPBEP, etc). The selected features from the education information based on correlation are shown in Table 12. Similar to population, for education information, we have data for 45,508 IRIS zones (out of 48,931).

Next, we performed the Pearson's correlation test to study correlation between all 25 socio-economic features as shown in Figure 20. For the ease of interpretation, all weak correlation values (between -0.5 and 0.5) are set to null. From Figure 20, we observe that poverty is negatively correlated with median income, which is obvious. Median income is positively correlated with high intellectual profession and high education, that implies, high professionals and highly educated individuals earn more. Similarly, we can observe that individuals with age above 60 years are retired (correlation value as 0.91).

### 4.1.3. Geographical Information

Geographical information provided publicly by the French Institut national de l'information géographique et forestière (IGN). For the geographical description

[1] IRIS is a region used to divide the country into units of similar population size.

| Socio-Economic Indicator | Definition |
|---|---|
| **Economic status** | |
| Poverty | Poverty rate at the threshold of 60% of disposable income per metropolitian median per person (%) |
| Median Income | Median income per person (in euros) |
| Gini Index | The Gini Index summarizes the dispersion of income across the entire income distribution i.e, it measure income inequality |
| **Population information** | |
| Total population | Total number of people |
| Pop 0-14 | Number of (#) people aged between 0 to 14 years |
| Pop 15-29 | #people aged between 15 to 29 years |
| Pop 30-44 | #people aged between 30 to 44 years |
| Pop 45-59 | #people aged between 45 to 59 years |
| Pop 60-74 | #people aged between 60 to 74 years |
| Pop 75+ | #people aged 75 years or more |
| Immigrants | #people who are immigrants |
| CS1 | #people aged 15 or more who are Farmer operators |
| CS2 | #people aged 15 or more who are Craftsmen, Traders, Company managers |
| CS3 | #people aged 15 or more who are Managers and higher intellectual professions |
| CS4 | #people aged 15 or more who are Intermediate professions |
| CS5 | #people aged 15 or more who are Employees |
| CS6 | #people aged 15 or more who are Worker |
| CS7 | #people aged 15 or more who are Retired |
| CS8 | #people aged 15 or more who fall in Others w/o professional activity |
| Male | #people who are male |
| Female | #people who are female |
| **Education information** | |
| No diploma | #out-of-school people aged 15 or over with no diploma or atleast a CEP |
| BEPC or CAPBEP | #out-of-school people aged 15 or over holding a BEPC, college certificate, DNB, CAP, BEP or equivalent; |
| BAC | #out-of-school people aged 15 or over with a Baccalaureate, professional certificate or equivalent |
| SUP | #out-of-school people aged 15 or over with a higher education diploma at Bac +2 level or more |

Table 12: Selected socio-economic features (per IRIS area) with their definitions from INSEE.

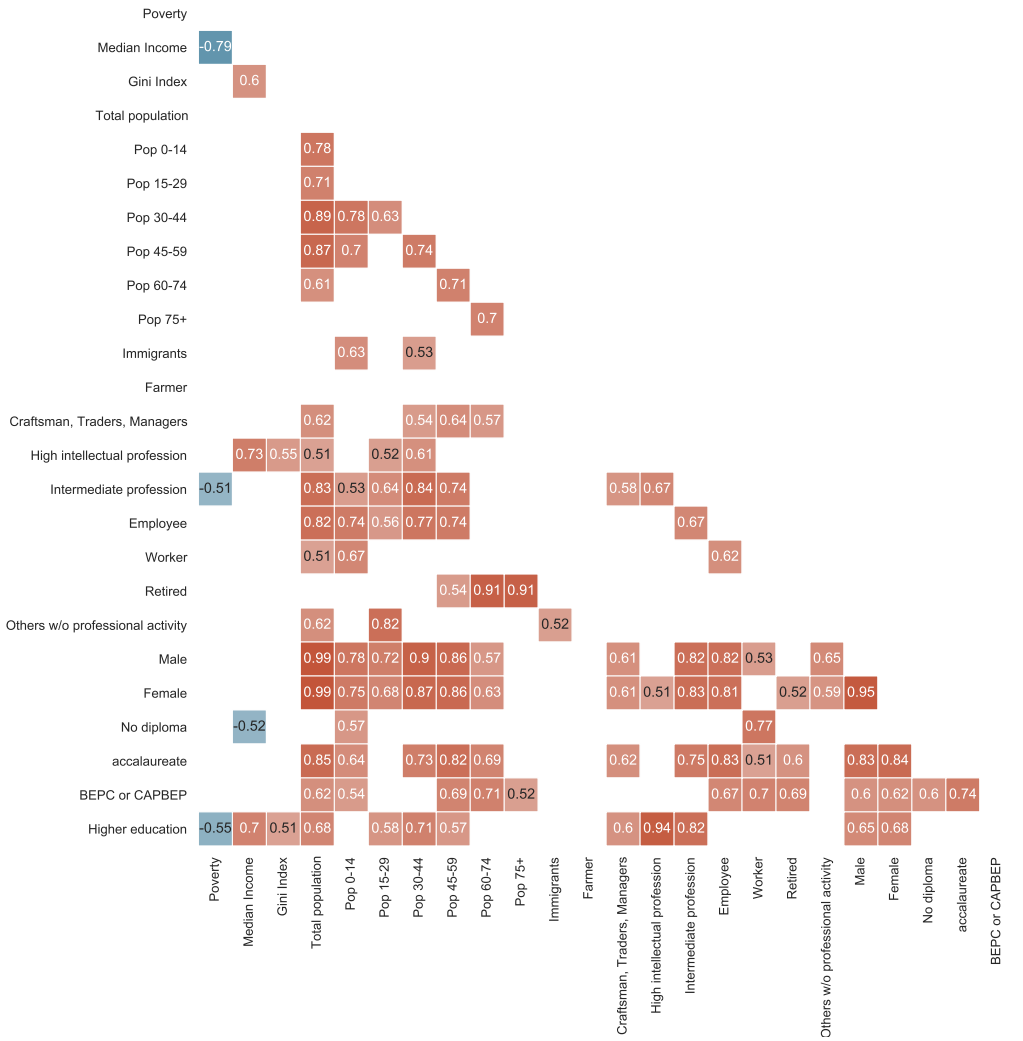



Figure 20: Correlation plot for socio-economic features.

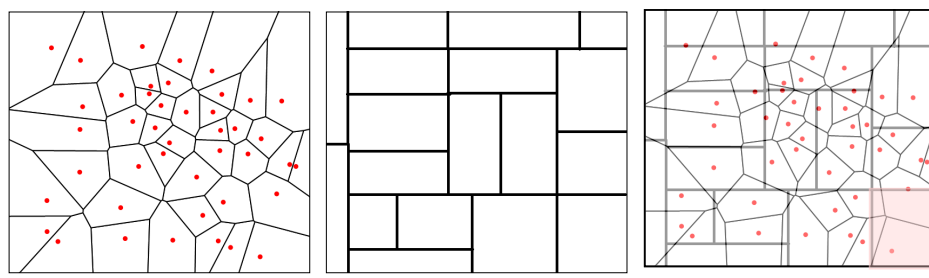
we used IRIS level information, and we downloaded the **Contours IRIS édition 2018** dataset [INS18a], which defines a polygon in a Lambert-93 projection for each IRIS zone (i.e. aggregated unit for statistical information) in France, as well as an associated record containing the IRIS code, name and type among other information. According to the Institut national de la statistique et des études économiques or National Institute of Statistics and Economic Studies (INSEE) in France, an IRIS is defined as a system for dividing the country into units of equal size, also known as IRIS2000. In French, IRIS is an acronym of *aggregated units for statistical information*, and the 2000 refers to the target size of 2000 residents per basic unit.

As we can observe, the Geographical data and socio-economic features mentioned in this section are at the IRIS level. However, the mobile app data is at the

antenna level. Therefore, we convert the mobile app data to IRIS level using the areal interpolation method.

## 4.2. Areal Interpolation of IRIS

In this section, we identified the areal weighted interpolation of IRIS using base station location and IRIS geographic information. As mentioned earlier, our mobile app data is aggregated by antennas. These antennas are located at base stations, which means a base station (BS) can correspond to single or multiple antennas. Therefore, as a first step, we mapped all the antennas to their respective BS based on their latitude (LAT) and longitude (LON) information. Next, we followed the technique mentioned in [Aas+21; Uca+21] for the areal weighted interpolation of IRIS. We begin by modeling the coverage area of each BS using the Voronoi polygon.



(a) Voronoi Polygon (b) IRIS Zone (c) Voronoi and IRIS zone mapping

Figure 21: Areal interpolation. The Voronoi polygons, IRIS region, and weighted interpolation are shown in Figure 21a, 21b, and 21c respectively. (best seen in color)

Let us suppose we have finite number of BS, $N$, in the two-dimensional Euclidean plane, and assume that $2 \leq N < \infty$. The $N$ number of BS are labeled by $bs_1$, $bs_2$,..., $bs_N$ with the Cartesian coordinates $(x_1, y_1)$, $(x_2, y_2)$, ..., $(x_N, y_N)$ or location vectors $X_1$, $X_2$, ..., $X_N$. The $N$ points are distinct in the sense that $X_i \neq X_j$ for $i \neq j$, where $i, j \in I_N = \{1, \cdots, N\}$, where $I_N$ represents a set of $N$ natural numbers. Let $bs$ be an arbitrary BS in the Euclidean plane with coordinates $(x, y)$ or location vector $X$. Then the Euclidean distance from $bs$ to $bs_i$ is given by

$$d(bs, bs_i) = ||X - X_i|| = \sqrt{(x - x_i)^2 + (y - y_i)^2} \tag{4.1}$$

If $bs_i$ is the nearest point from $bs$, we have the relation $||X - X_i|| \leq ||X - X_j||$ for $j \neq i$, where $j \in I_N$. Let B = $\{bs_1, bs_2,..., bs_N\}$, then the Voronoi polygon region associated with $bs_i$ is given by

$$V(bs_i) = \{X \; s.t. \; ||X - X_i|| \leq ||X - X_j|| \; for \; j \neq i, j \in I_N\} \tag{4.2}$$

and the Voronoi diagram generated by $B$ is given by the set

$$V = \{V(bs_1), V(bs_2), ..., V(bs_N)\} \tag{4.3}$$

Figure 21a provides an example of the Voronoi polygon generated using BS in a bounded plane. In equation 4.3, we have defined a Voronoi diagram in an unbounded plane. However, in most cases we deal with a bounded region. Now, the bounded Voronoi diagram generated by $B$ for a bounded region $R$ is given by the set

$$V_{bnd} = V \cap R = \{V(bs_1) \cap R, ..., V(bs_N) \cap R\} \tag{4.4}$$

Let suppose we have finite number of IRIS zones, $Z$, which are labelled as $I_1, I_2, ..., I_Z$. Then, the areal weight for IRIS zone $i$ with respect to all the BS bounded Voronoi polygon is given by

$$W(I_i) = \left\{ \frac{A(V_{bnd}(bs_1) \cap I_i)}{A(V_{bnd}(bs_1))}, ..., \frac{A(V_{bnd}(bs_N) \cap I_i)}{A(V_{bnd}(bs_N))} \right\} \tag{4.5}$$

where, $V_{bnd}(bs_i)$ represents the bounded Voronoi polygon for $bs_1$, $(V_{bnd}(bs_1) \cap I_i)$ represents the overlapping area of Voronoi polygon for $bs_1$ and IRIS zone $i$. A$(\cdot)$ represents the area of the region or polygon within the parenthesis. Hence, $\frac{A(V_{bnd}(bs_1) \cap I_i)}{A(V_{bnd}(bs_1))}$ is the areal weight for IRIS zone $i$ with respect to the $bs_1$ bounded Voronoi polygon. This areal weight is further multiplied by the total traffic recorded for the respective BS to calculate the traffic consumption.

Next, the identified aerial weights are multiplied by the total traffic recorded for an app category on the respective BS of an IRIS to calculate the traffic consumption by an app category in the IRIS zone. For example, let us consider the generic IRIS zone $Z$, containing three BS with aerial weights $W_1, W_2$ and $W_3$ and traffic $T_1, T_2$ and $T_3$ for app category $A$ at any given time. The total traffic consumption for app category $A$ in the considered IRIS $Z$ will amount to $T_1 * W_1 + T_2 * W_2 + T_3 * W_3$. After this, we have per-application traffic data entries recorded as the uplink and downlink at a 1-minute granularity and aggregated by IRIS. As a next step, we also sum up the uplink and downlink byte counts, which results in the dataset with per-application traffic data entries recorded as the total byte counts at a temporal granularity of 1 minute and aggregated by IRIS. In the next section, we refer to this dataset as $\mathscr{D}$.

## 4.3. Mobile Apps Usage Features for IRIS Sectors

Predicting socio-economic features using mobile app data significantly depends upon distinctive usage patterns in mobile applications. In this section, we derived a set of metrics from apps data able to capture useful information from the large-scale data to be later used in the classification stage. In particular, we identified

week signature (Section 4.3.1), Revealed Comparative Advantage (RCA) utilization (Section 4.3.2), and standardized cumulative utilization (Section 4.3.3), formally defined in the following subsections.

### 4.3.1. Typical Week Signature (TWS)

Let us consider our mobile app consumption dataset $\mathscr{D}$, describing the usage of a set of mobile apps $A$ for the population during a set of $d$ days. For the definition of the signature, let us define the signature support $\Delta$ of one week, i.e., from Monday to Sunday, $\Delta = \{\text{mon, tue, wed, thu, fri, sat, sun}\}$. Then, $d^{\delta} \subset d$ denotes the set of days in the dataset $D$ that correspond to the day of the week $\delta \in \Delta$, with $\bigcup_{\delta \in \Delta} d^{\delta} = d$. The generic element in the signature of an IRIS $i$ for mobile app $a$ at time $t$ for day of the week $\delta$ is:

$$s_i(\delta,t,a) = M(\{v_i(d,t,a) | d \in d^{\delta}\}), \quad \forall i \in I, a \in A, \tag{4.6}$$

where, $v_i(d,t,a)$ describes the total mobile app $a$ usage within the IRIS $i$ at time slot (hour in our case) $t$ of day $d$ from dataset $\mathscr{D}$. M$(\cdot)$ represents the median of the set within parenthesis. Also in this case, $\delta$ is small with respect to the overall set of $d$ days, which implies data compression. $I$ and $A$ represents set of IRIS and applications respectively.

Signatures then undergo a standard normalization. To that end, each element obtained in Equation 4.6 is normalized with respect to the mean and standard deviation of all elements referring to the same IRIS and mobile application. Formally, for a generic element of IRIS $i$ and mobile application $a$ is

$$\tilde{s}_i(\delta,t,a) = \frac{s_i(\delta,t,a) - \mu(s_i(a))}{\sigma(s_i(a))}, \ \forall \delta \in \delta, i \in I, a \in A, \tag{4.7}$$

where $\mu(s_i(a))$ and $\sigma(s_i(a))$ denote the mean and standard deviation of the set of elements concatenated in the signature $s_i$.

The techniques for the construction of a TWS for per-application traffic data, at a temporal granularity of 1 hour and for each IRIS [Fur+16; FFS17] requires to process the dataset $\mathscr{D}$ through three phases. These phases aim at summarizing the mobile traffic activity in each unit areas into a meaningful profile, i.e., the IRIS signature for the mobile application category. The steps involved in generating the TWS are as follows:

1. We converted the timestamp value to the corresponding "*Hour within week*" value, which means that all the timestamp values represent the hour and day of the week. For example, all the timestamp values on Monday between 00:00 to 00:59 hours, will become 1 irrespective of the date. Similarly, all the timestamp values on Tuesday between 16:00 to 16:59 hours, will become 41 (24+17) irrespective of the date.
2. Next, we aggregate the data on "*Hour within week*" value, and calculate the median total byte count per-application for each IRIS. We considered the

median value as denoising component and extracts information deemed to be representative of the typical mobile traffic activity in an IRIS, isolating it from the inherent noise in the data.

3. Lastly, we standardised the signature making it independent from the absolute volume of mobile traffic recorded at an IRIS. This allows comparing the per-application activity at IRIS level on the sole basis of the traffic data variations. This we named as *Typical Week Signature*.

The *TWS* is a relevant feature from mobile phone app usage data due to its capability to depict the relative dynamics of mobile app consumption over the course of an entire week under the premise of cyclic (weekly) regularity at the scale of an IRIS region.

### 4.3.2. Revealed Comparative Advantage (RCA)

Depending on the nature of the data, different mobile apps produce a heterogeneous traffic volume over the network. For example, videos generate a significantly larger load per session than messaging. As a result, the traffic volume across app categories might vary by several orders of magnitude. Therefore, we also computed the revealed comparative advantage (RCA) [Bal65], a relative measure that can compare traffic across IRIS zones and app categories. The RCA index for IRIS zone $z$ and app category $a$ is calculated as follows:

$$RCA_{za} = \frac{T_{za} / \sum_{a' \in A} T_{za'}}{\sum_{z' \in Z} T_{z'a} / \sum_{z' \in Z, a' \in A} T_{z'a'}} \tag{4.8}$$

where $T_{za}$ is the median hourly traffic per person in IRIS zone $z$ for apps category $a$; $\sum_{a' \in A} T_{za'}$ is the median hourly traffic per person in zone $z$ combinedly generated by all considered apps; $\sum_{z' \in Z} T_{z'a}$ is the median hourly traffic per person generated by app $a$ in all zones; $\sum_{z' \in Z, a' \in A} T_{z'a'}$ is the median hourly traffic per person, aggregated over all zones and apps. A value of $RCA_{za} > 1$ indicates more than average usage of app $a$ in zone $z$; on the other hand, $RCA_{za} < 1$ indicates lower than average usage of app $a$ in zone $z$.

### 4.3.3. Standardized Cumulative Utilization (SCU)

Next, we extracted the cumulative utilization of apps in different IRIS. This is defined as the total data transmitted in byte counts for each app aggregated for the IRIS. This is instead a simple metric compared to the *week signature* and *RCA*. The motivation behind calculating this information is to check the effectiveness of straightforward metrics to predict socio-economic features. We take the total data transmitted bytes during the 2.5-month observation period for each IRIS zone and mobile service. Finally, we calculate the standardized cumulative utilization (SCU) as follows:

$$SCU_{za} = \frac{C_{za} - mean(\forall_Z C_{Za})}{sd(\forall_Z C_{Za})} \tag{4.9}$$

where $C_{za}$ is the cumulative traffic in IRIS zone $z$ for app category $a$; $mean(\forall_Z C_{za})$ is the mean total traffic in all zones $Z$ generated by app $a$; $sd(\forall_Z C_{za})$ is the standard deviation of total traffic in all zones $Z$ generated by app $a$. Therefore, the SCU value measures how far from the mean the traffic generated by a particular mobile application in a specific IRIS zone is in terms of standard deviations.

## 4.4. Socio-Economic Features Prediction

Here, we build machine learning models using the patterns (TWS, RCA, and SCU) we explored in the previous sections to predict the socio-economic features of IRIS in France.

### 4.4.1. Experimental Setup

To illustrate the predictive power of the different feature sets (TWS, RCA, and SCU), we define a series of models, each with a distinct feature set corresponding to mobile app usage patterns.

We used Python programming language for the implementation. For all the experiments, we use PowerEdge R740 server, 2 * Intel Xeon Gold 5220 (2,2GHz, 18C/36T, cache 24,75MB, 10,4GT/s, 125W, Turbo, HT), DDR4 2666MHz, 768GB Ram RDIMM 3200MT/s, Dual Rank.

Furthermore, we experiment with several regression models, including Linear Regression, Ridge Regression, and CatBoost. We find that CatBoost models are the most effective for our task, and we thus report results from only those models. Because of the regression task, we take care of the trade-off between bias and variance of the model. During experimentation, we split our dataset into 80:20 train and test sets, ensuring that our CatBoost models are not overfitting our training data. To achieve this, we used a 5-fold cross-validation on training data. Finally, we report the R-squared value on test data.

### 4.4.2. Results with Explainability

We subdivide our results into three parts based on economic status, population information, and education information.

**1) Economic status:** Table 13 shows the R-squared value for the prediction of our models. We used four variations of predictive features to predict Poverty, Median income, and the Gini index. The variations of predictive features include (1) *Cumulative:* that represents SCU values, (2) *RCA:* that includes RCA index values, (3) *TWS:* that has typical week signature on hourly basis, and (4) *All:* which contains all the three mentioned features combined, i.e., SCU values, RCA index, and TWS.

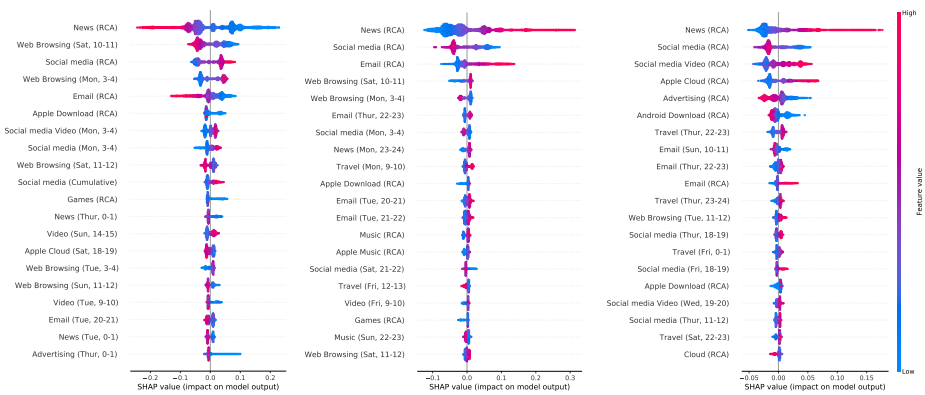
We find the most predictive features to be the week signature or TWS. Strikingly, mobile app usage at different times of day is significantly more predictive than the RCA index and cumulative usage patterns. With a model trained on all

| Socio-Economic Features | Cumulative | RCA | TWS | All |
|---|---|---|---|---|
| *Economic status* | | | | |
| Poverty | 0.306 | 0.366 | 0.458 | **0.482** |
| Median Income | 0.551 | 0.592 | 0.631 | **0.659** |
| Gini Index | 0.569 | 0.592 | 0.623 | **0.642** |
| *Education information* | | | | |
| No diploma | 0.380 | 0.392 | 0.450 | **0.456** |
| Baccalauréat | 0.442 | 0.454 | 0.508 | **0.516** |
| BEPC + CAPBEP | 0.381 | 0.398 | 0.457 | **0.464** |
| College | 0.538 | 0.548 | 0.596 | **0.604** |
| *Population information* | | | | |
| Total population | 0.457 | 0.470 | 0.520 | **0.527** |
| Population 0 to 14 yrs | 0.464 | 0.480 | 0.538 | **0.543** |
| Population 15 to 29 yrs | 0.529 | 0.541 | 0.591 | **0.596** |
| Population 30 to 44 yrs | 0.483 | 0.496 | 0.553 | **0.560** |
| Population 45 to 59 yrs | 0.449 | 0.461 | 0.516 | **0.522** |
| Population 60 to 74 yrs | 0.393 | 0.406 | 0.462 | **0.469** |
| Population 75+ yrs | 0.368 | 0.378 | 0.428 | **0.435** |
| Immigrants | 0.596 | 0.605 | 0.654 | **0.664** |
| Farmer | 0.246 | 0.256 | 0.322 | **0.328** |
| Craftsman, Trader | 0.363 | 0.372 | 0.407 | **0.417** |
| Manager, high profession | 0.544 | 0.555 | 0.593 | **0.602** |
| Intermediate profession | 0.473 | 0.484 | 0.537 | **0.543** |
| Employees | 0.437 | 0.449 | 0.498 | **0.504** |
| Worker | 0.354 | 0.369 | 0.432 | **0.437** |
| Retired | 0.359 | 0.371 | 0.425 | **0.433** |
| Other w/o profession | 0.485 | 0.494 | 0.536 | **0.540** |
| Male population | 0.459 | 0.472 | 0.525 | **0.532** |
| Female population | 0.467 | 0.478 | 0.529 | **0.535** |

Table 13: R-squared scores using *Cumulative*, *RCA*, *TWS*, and *All* predictive features for the socio-economic features. Best score is highlighted using bold text.

available features, we achieve a score of 0.48, 0.66, and 0.64 for Poverty, Median income, and Gini index, respectively. The results outperformed the state-of-the-art, which explored only the RCA index and not TWS. Table in Appendix B contains the RMSE scores using TWS, RCA, SCU, and All predictive features for the socio-economic features. Additionally, the best scores are highlighted using bold text.

Furthermore, to understand the features' importance for the best model with *All* predictive features, we use SHAP (Shapley additive explanations) values [LL17]. A summary plot of SHAP values for top-20 features is shown in Figure 22. The SHAP values and feature names are represented by the x- and y-axes, respectively. Each data point represents a single instance. The red color represents a higher value for the feature than its average value, while the blue color denotes a lower value. Red values on the right side of the x-axis indicate a positive impact on the prediction and vice versa. The features are listed in order of decreasing importance.




(a) Poverty (b) Median Income (c) Gini Index

Figure 22: SHAP plot to determine feature importance for Economic status. (best seen in color).

Each point represents an IRIS, thus for each feature, the figure shows the distribution of SHAP values across the dataset. The wider distribution indicates a larger absolute impact of the feature in the prediction, while the color of each point encodes the feature value (low or high). A feature with high values corresponding to negative SHAP values (to the left) is negatively correlated with economic status. For example, RCA index for *News* app category or *News (RCA)* is negatively correlated with *Poverty* (see Figure 22b and 22c). Conversely, features with high values correspond to positive SHAP values (to the right), which positively correlate with economic status. For example, *News (RCA)* is positively correlated with *Median income* and *Gini index* (see Figure 22b and 22c).

The most important feature for predicting economic status are News, Web Browsing, Social media, Email, Social media video, etc. We observe that *News*

*(RCA)* consumption is more in IRIS with high median income and a high Gini index. On the other hand, *News (RCA)* consumption is less in IRIS with high Poverty. That means News is read more by high-income individuals. Additionally, RCA index for *Social media* app category or *Social media (RCA)* usage pattern is completely opposite compared to *News (RCA)*. That indicates that high-income people do not frequently use social media. Similarly, we find that *Web Browsing* on Saturday from 10 to 12 is positively correlated with *Median income* and negatively correlated with *Poverty*. Conversely, we find that *Web Browsing* on Monday from 3 to 4 is negatively correlated with *Median income* and positively correlated with *Poverty*. This implies that high-income individuals avoid browsing the internet on Monday late afternoon.

**Takeaways.** Social media applications are utilized less by the high-income group than by the low-income group. On the other hand, the high-income group frequently uses News and Productivity apps.

**2) Education:** Table 13 reports the R-squared values for the CatBoost models for four education classes: No diploma, Compulsory Schooling (Baccalauréat), High School (BEPC and CAPBEP), and College. Again, we find the most predictive features to be the weekly signature. That means mobile app usage at different times of the day is significant. With a model trained on all available features, we achieve a score of 0.46, 0.52, 0.46, 0.60 for No diploma, Compulsory Schooling (Baccalauréat), High School (BEPC and CAPBEP), and College, respectively.

**Takeaways.** We find that the consumption of social media videos, social media, music, and travel applications play a vital role in predicting the education status in the IRIS. We also observe that people without a diploma, compulsory education, and high school have more distinguished consumption of music and social media than those with a college education.

**3) Population:** Table 13 report the R-squared values for the CatBoost models for all population features. We achieve scores between 0.44 (age 75+) to 0.66 (Immigrants). Our model predicted the immigrant population and people ages 15 to 29 years more accurately. Similarly, we achieve scores for profession information between 0.33 (Farmers) to 0.60 (Managers, high profession). The model performed better in predicting high profession, Intermediate profession, and Others with high scores.

**Takeaways.** We observe that most of the features among the top 20 are from TWS. This further supports the relevance of TWS. We find that social media video is highly consumed by people aged between 15 to 45 years. The immigrant population, primarily young and working in positions with unprofessional activities, are not frequent targets for the advertisement. Additionally, high profession individuals, such as, Managers like to hear music after working hours.

| Socio-Economic Features | Census | All |
|---|---|---|
| Median income | 0.50 | **0.66** |
| Gini Index | 0.38 | **0.64** |
| College | **0.61** | **0.61** |

Table 14: R-squared scores for CatBoost model using *census*, and *All (Cumulative, RCA, and TWS patterns)* data for predicting Median income, Gini index, and College education. Best score is highlighted using bold text.

### 4.4.3. Relevance of the Findings

To show that the extracted patterns (TWS, RCA, and SCU) from mobile applications are better/comparable than census information for predicting socio-economic features, we further experimented using census data (Total pop, age 0-14, 15-29, 30-44, 45-59, 60-74, and 75+), and All (Signature, RCA, and Cumulative) to predict three socio-economic features: Median income, Gini index, and College education. Table 14 report the R-squared values for the CatBoost models. We find that the model with *All* predictive features achieve high scores.

Please note that, for the sake of completeness, we performed a comparability analysis between the mobile applications usage patterns and census information for predicting three socio-economic features: Median income, Gini index, and College education. However, the use of population information for predicting socio-economic features is not aligned with the purpose of this study. That is, this study tries to replace expensive, and time-consuming census tasks to collect socio-economic information by mobile usage to predict socio-economic features.

## 4.5. Summary

This work proposes a large-scale, quantitative, and predictive study of the relationship between mobile app usage and socio-economic features. We analyzed nationwide data from the leading mobile operators in France and extracted three patterns: TWS, RCA, and SCU. We find that TWS has richer information diversification than RCA and SCU. The best model using all three patterns achieved an R-squared score up to 0.66, thus concluding our study goal that mobile application usage can predict a nation's socio-economic conditions.

Although mobile application usage patterns are successful in predicting socioeconomic features, model drift can occur over time, making these predictions erroneous. Therefore, it is necessary to automatically rectify model drift. As a result, mobile digital data can augment censuses (for example, if a census occurs every ten years, mobile data can help with better interpolation with precise estimates for every year or even every month). The model will need to be calibrated with a physical census on a regular basis (e.g., every ten years, as is done in many countries).

**Impact:** Most countries conduct a census once every ten years and utilize the

same data to drive policies prior to the next census [UN17; SS20]. To overcome the limitation of census, this study makes a case for an inexpensive, privacy-preserving, real-time and scalable method to understand the latest socio-economic conditions and, by extension, poverty, unemployment, literacy, or economic progress in our societies through mobile phone data (application usage).

# 5. CONCLUSION

As mobile phones have become an inseparable part of our lives, the significance of call data and digital data to study people's behaviour worldwide cannot be underestimated. Most people who spend a large chunk of time on mobile leave their usage signature, consciously and unconsciously, which can be analyzed to mine the distinctive patterns. These patterns are later utilized to understand individual or group mobility, demographic segregation, socio-economic conditions, and so on. Hence, mobile phone data is an excellent source to study numerous social well-being topics.

In this thesis, we investigated three societal well-being facets using mobile phone data. First, we conducted modeling research to better understand the propagation of coronavirus. Here, we developed two versions of the mobility-based SIR model, (i) fully-mixed and (ii) for complex networks, which take into account the real-life interactions from CDR. Using the proposed method, we predicted coronavirus cases in Estonia and France.

The second work is a descriptive research in which we use the CDR data to analyze societal segregation. This is another major social problem since it plays a vital part in establishing a country's development trajectories by assisting governments and other relevant entities in developing better-targeted policies for the appropriate populations. Our findings suggest that CDR can effectively portray the real population and also identify social segregation.

The third work is a predictive research where we look at digital traces from mobile applications (such as Twitter, YouTube, and so on) to predict socioeconomic indicators. These indicators include data on education, gender, poverty, employment, and other variables. As a result, precise and trustworthy information is vital for social research and government policing. Today, most public data sources, such as censuses, have limited population coverage or are only updated periodically. Nonetheless, our findings indicate that data sources like digital traces of mobile app activity can be used to detect socioeconomic indicators in a cost-effective and timely manner.

# 6. FUTURE SCOPE

In this thesis, we investigated three social well-being aspects using mobile phone data. This section discusses different potential future directions for modelling and experiments.

There are two dimensions in which we plan to extend our epidemic modeling research in Chapter 2. First, this thesis has been mainly focused on the use of compartmental models to understand epidemics. Generally, there are two types of epidemic modeling: compartmental and agent-based modeling [FWF11; Ker+21]. In this thesis, we have only focused on compartmental models. Hence, we plan to develop *agent-based models* to understand the spread of coronavirus in the future. Second, we would like to test our proposed models on larger datasets with longer time-period coverage to map people's mobility more accurately.

While studying the segregation in Chapter 3, it would be interesting to analyze the data in a more detailed location, such as the municipalities to grasp the economic inequality at a more finer level. We would also like to combine the mobile CDR data with other datasets such as *financial* data to understand the socioeconomic segregation in Estonia.

The outcomes of the socio-economic study in Chapter 4 are confined to France and its specific geographical granularity, IRIS. However, there are a few potential ideas that we would like to investigate. First, using the proposed methodology, study other countries' call data records or digital data. Second, we can extend the proposed methodology by including user temporal network analysis and the investigation of alternate sources of information.

# BIBLIOGRAPHY


[Aas+21] Anto Aasa et al. “Spatial interpolation of mobile positioning data for population statistics”. In: *Journal of Location Based Services* 15.4 (2021), pp. 239–260.

[AB02] Réka Albert and Albert-László Barabási. “Statistical mechanics of complex networks”. In: *Reviews of modern physics* 74.1 (2002), p. 47.

[Aha+08] Rein Ahas et al. “Evaluating passive mobile positioning data for tourism surveys: An Estonian case study”. In: *Tourism Management* 29.3 (2008), pp. 469–486.

[Aha+10] Rein Ahas et al. “Using mobile positioning data to model locations meaningful to users of mobile phones”. In: *Journal of urban technology* 17.1 (2010), pp. 3–27.

[AK20] Jacob Levy Abitbol and Marton Karsai. “Interpretable socioeconomic status inference from aerial imagery through urban patterns”. In: *Nature Machine Intelligence* 2.11 (2020), pp. 684–692.

[Ale+03] Alberto Alesina et al. “Fractionalization”. In: *Journal of Economic growth* 8.2 (2003), pp. 155–194.

[AM05] Rein Ahas and Ülar Mark. “Location based services—new challenges for planning and public administration?” In: *Futures* 37.6 (2005), pp. 547–561. ISSN: 0016-3287.

[Asi+20] Aili Asikainen et al. “Cumulative effects of triadic closure and homophily in social networks”. In: *Science Advances* 6.19 (2020), eaax7310.

[ASS18] Md Arquam, Anurag Singh, and Rajesh Sharma. “Modelling and Analysis of Delayed SIR Model on Complex Network”. In: *International Conference on Complex Networks and their Applications*. Springer. 2018, pp. 418–430.

[AT12] James P Allen and Eugene Turner. “Black–white and Hispanic–white segregation in US counties”. In: *The Professional Geographer* 64.4 (2012), pp. 503–520.

[B P+15] Nicolas B. Ponieman et al. “Mobility and sociocultural events in mobile phone data records”. In: *Ai Communications* 29 (Sept. 2015), pp. 77–86.

[Bal65] Bela Balassa. “Trade liberalisation and “revealed” comparative advantage 1”. In: *The manchester school* 33.2 (1965), pp. 99–123.

[Ban] The World Bank. *GDP (current US$)*. URL: `https://data.worldbank.org/indicator/Ny.Gdp.Mktp.Cd?most_recent_value_desc=true`.

[Bar+05] Marc Barthélemy et al. “Dynamical patterns of epidemic outbreaks in complex heterogeneous networks”. In: *Journal of theoretical biology* 235.2 (2005), pp. 275–288.

[BBC20] BBC. *Life on Estonia's 'corona island'*. [Accessed 09-Mar-2023]. 2020. URL: https://www.bbc.com/news/av/world-europe-52282118.

[BC06] Lawrence A Brown and Su-Yeul Chung. "Spatial segregation, segregation indices and the geographical perspective". In: *Population, space and place* (2006).

[BDK15] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. "A survey of results on mobile phone datasets analysis". In: *EPJ data science* 4.1 (2015), p. 10.

[BF13] Joshua Blumenstock and Lauren Fratamico. "Social and spatial ethnic segregation: a framework for analyzing segregation with large-scale spatial network data". In: *Proceedings of the 4th Annual Symposium on Computing for Development*. 2013, pp. 1–10.

[BG95] Trevor C Bailey and Anthony C Gatrell. *Interactive spatial data analysis*. Vol. 413. 8. Longman Scientific & Technical Essex, 1995.

[Blu16] Joshua Evan Blumenstock. "Fighting poverty with data". In: *Science* 353.6301 (2016), pp. 753–754.

[Bon02] Eric Bonabeau. "Agent-based modeling: Methods and techniques for simulating human systems". In: *Proceedings of the national academy of sciences* 99.suppl 3 (2002), pp. 7280–7287.

[Bou11] Jonathan Bousfield. *The Rough Guide to Estonia, Latvia & Lithuania*. Dorling Kindersley Ltd, 2011.

[BY08] Pleskovic Boris and Justin Yifu Lin. *Annual World Bank Conference on Development Economics 2008, Regional: Higher Education and Development*. The World Bank, 2008.

[Che] Sergey Chernov. *Tartu: Estonia's Intellectual and Theater Capital*. https://www.themoscowtimes.com/2012/12/23/tartu-estonias-intellectual-and-theater-capital-a20396. Published: Dec. 24 2012.

[Col58] James Coleman. "Relational analysis: The study of social organizations with survey methods". In: *Human organization* 17.4 (1958), pp. 28–36.

[COV19] Coronavirus COVID. *Global Cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)*. 19.

[CR07] Magnus Carlsson and Dan-Olof Rooth. "Evidence of ethnic discrimination in the Swedish labor market using experimental data". In: *Labour Economics* 14.4 (2007), pp. 716–729.

[CSS20] JHU CSSE. *Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)*. 2020.

[Czy] Andrew Czyzewski. *Modelling an unprecedented pandemic (Imperial Stories)*. [Accessed 24-Apr-2023]. URL: %5Curl%7Bhttps://www.imperial.ac.uk/stories/coronavirus-modelling/#Explained-Anatomy-of-a-model-VNpjbONLmY%7D.

[dat22] datareportal. *Digital around the world - datareportal – global digital insights*. 2022. URL: https://datareportal.com/global-digital-overview#:~:text=There%5C%20are%5C%205.31%5C%20billion%5C%20unique,in%5C%20the%5C%20past%5C%2012%5C%20months..

[Daw04] Casey J Dawkins. "Measuring the spatial pattern of residential segregation". In: *Urban Studies* 41.4 (2004), pp. 833–851.

[DD55a] Otis Dudley Duncan and Beverly Duncan. "A methodological analysis of segregation indexes". In: *American sociological review* 20.2 (1955).

[DD55b] Otis Dudley Duncan and Beverly Duncan. "Residential distribution and occupational stratification". In: *American journal of sociology* 60.5 (1955).

[Dev+14] Pierre Deville et al. "Dynamic population mapping using mobile phone data". In: 111.45 (2014), pp. 15888–15893. DOI: 10.1073/pnas.1408439111.

[DP] Centers for Disease Control and Prevention. *First global estimates of 2009 H1N1 pandemic mortality released by CDC-led collaboration. 2012*.

[DRZ19] Lei Dong, Carlo Ratti, and Siqi Zheng. "Predicting neighborhoods' socioeconomic attributes using restaurant data". In: *Proceedings of the national academy of sciences* 116.31 (2019), pp. 15447–15452.

[EPL08] Nathan Eagle, Alex (Sandy) Pentland, and David Lazer. "Mobile Phone Data for Inferring Social Network Structure". In: *Social Computing, Behavioral Modeling, and Prediction*. Ed. by Huan Liu, John J. Salerno, and Michael J. Young. Boston, MA: Springer US, 2008, pp. 79–88. ISBN: 978-0-387-77672-9.

[Eps09] Joshua M Epstein. "Modelling to contain pandemics". In: *Nature* 460.7256 (2009), pp. 687–687.

[ERR20a] ERR. *Terviseamet: Eestis on kinnitatud 27 koroonajuhtu ja kohalik levik*. https://www.err.ee/1063204/terviseamet-eestis-on-kinnitatud-27-koroonajuhtu-ja-kohalik-levik. 2020.

[ERR20b] ERR. *Three new cases of coronavirus disease confirmed in Estonia*. [Accessed 09-Mar-2023]. 2020. URL: https://news.err.ee/1062392/three-new-cases-of-coronavirus-disease-confirmed-in-estonia.

[Esta] Statistics Estonia. *Mean annual population by sex and age group*. URL: https://www.stat.ee/esms-metadata?code=30205.

[Estb] Statistics Estonia. *Statistics Estonia*. https://andmed.stat.ee/et/stat. Accessed: 2021-03-20.

[Est18] Statistics Estonia. *Quarterly bulletin of statistics Estonia*. 2018.

[Exp20] Estonian Express. *Täiuslik koroonakolle: kuidas viirus Saaremaa menuüritustelt valla pääses*. [Accessed 09-Mar-2023]. 2020. URL: https://ekspress.delfi.ee/artikkel/89250245/taiuslik-koroonakolle-kuidas-viirus-saaremaa-menuuritustelt-valla-paases?.

[Fek+20] Mariem Fekih et al. "A data-driven approach for origin–destination matrix construction from cellular network signalling data: a case study of Lyon region (France)". In: *Transportation* (2020), pp. 1–32.

[FFS17] Angelo Furno, Marco Fiore, and Razvan Stanica. "Joint spatial and temporal classification of mobile traffic demands". In: *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE. 2017, pp. 1–9.

[fra20a] République française. *Covid-19 : un 2e confinement national à compter du 29 octobre minuit - vie-publique.fr*. https://www.vie-publique.fr/en-bref/276947-covid-19-un-2e-confinement-national-compter-du-29-octobre-minuit. [Accessed in 2020]. 2020.

[fra20b] République française. *OLD Données de laboratoires infra-départementales durant l'épidémie COVID-19*. https://www.data.gouv.fr/fr/datasets/donnees-de-laboratoires-infra-departementales-durant-lepidemie-covid-19/. [Accessed in 2020]. 2020.

[Fre78] Linton C Freeman. "Segregation in social networks". In: *Sociological Methods & Research* 6.4 (1978), pp. 411–429.

[Fu+12] Feng Fu et al. "The evolution of homophily". In: *Scientific reports* 2 (2012), p. 845.

[Fur+16] Angelo Furno et al. "A tale of ten cities: Characterizing signatures of mobile traffic in urban areas". In: *IEEE Transactions on Mobile Computing* 16.10 (2016), pp. 2682–2696.

[FWF11] Enrique Frias-Martinez, Graham Williamson, and Vanessa Frias-Martinez. "An agent-based model of epidemic spread using human mobility and social network information". In: *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE. 2011, pp. 57–64.

[GFS23] Rahul Goel, Angelo Furno, and Rajesh Sharma. "Predicting Socio-Economic Well-being Using Mobile Apps Data: A Case Study of France". In: *arXiv preprint arXiv:2301.09986* (2023).

[GHB08] Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. "Understanding individual human mobility patterns". In: *Nature* 453.7196 (June 2008), pp. 779–782.

[Gil+15] Brian Joseph Gillespie et al. “Homophily, close friendship, and life satisfaction among gay, lesbian, heterosexual, and bisexual men and women”. In: *PloS one* 10.6 (2015), e0128900.

[Goj+09] Marija Zivkovic Gojovic et al. “Modelling mitigation strategies for pandemic (H1N1) 2009”. In: *Cmaj* 181.10 (2009), pp. 673–680.

[GS20a] R. Goel and R. Sharma. “Mobility Based SIR Model For Pandemics - With Case Study Of COVID-19”. In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2020, pp. 110–117. DOI: 10.1109/ASONAM49781.2020.9381457.

[GS20b] Rahul Goel and Rajesh Sharma. “Mobility based sir model for pandemics-with case study of covid-19”. In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE. 2020, pp. 110–117.

[GSA21a] Rahul Goel, Rajesh Sharma, and Anto Aasa. “Overall Behavioural Index (OBI) For Measuring Segregation”. In: *arXiv preprint arXiv:2106.10676* (2021).

[GSA21b] Rahul Goel, Rajesh Sharma, and Anto Aasa. “Studying segregation in Estonia using call data records”. In: *Social Network Analysis and Mining* 11.1 (2021), pp. 1–13.

[GSA21c] Rahul Goel, Rajesh Sharma, and Anto Aasa. “Understanding gender segregation through Call Data Records: An Estonian case study”. In: *PLoS ONE* 16.3 (2021). DOI: 10.1371/journal.pone.0248212.

[GSG19] Rahul Goel, Anurag Singh, and Fakhteh Ghanbarnejad. “Modeling competitive marketing strategies in social networks”. In: *Physica A: Statistical Mechanics and its Applications* 518 (2019), pp. 50–70.

[GZH18] Mohammadhossein Ghahramani, MengChu Zhou, and Chi Tin Hon. “Extracting significant mobile phone interaction patterns based on community structures”. In: *IEEE Transactions on Intelligent Transportation Systems* 20.3 (2018), pp. 1031–1041.

[GZZ19] Jian Gao, Yi-Cheng Zhang, and Tao Zhou. “Computational socioeconomics”. In: *Physics Reports* 817 (2019), pp. 1–104.

[Har16] Richard Harris. “Measuring segregation as a spatial optimisation problem, revisited: a case study of London, 1991–2011”. In: *International Journal of Geographical Information Science* 30.3 (2016), pp. 474–493.

[Has+17] Behrooz Hashemian et al. “Socioeconomic characterization of regions through the lens of individual financial transactions”. In: *PloS one* 12.11 (2017), e0187031.

[Het00] Herbert W Hethcote. “The mathematics of infectious diseases”. In: *SIAM review* 42.4 (2000), pp. 599–653.

[Hii+19] Hendrik Hiir et al. “Impact of Natural and Social Events on Mobile Call Data Records–An Estonian Case Study”. In: *International Conference on Complex Networks and Their Applications*. Springer. 2019, pp. 415–426.

[HJG06] Maureen Hurley, Glen Jacobs, and Melinda Gilbert. “The basic SI model”. In: *New Directions for Teaching and Learning* 2006.106 (2006), pp. 11–22.

[HSS08] Aric Hagberg, Pieter Swart, and Daniel S Chult. *Exploring network structure, dynamics, and function using NetworkX*. Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[Hua+16] Xiaodong Huang et al. “Bayesian estimation of the dynamics of pandemic (H1N1) 2009 influenza transmission in Queensland: A space–time SIR-based model”. In: *Environmental research* 146 (2016), pp. 308–314.

[inf] Nordisch info. *Nordisch.info*. `https://www.nordisch.info/estland/`. Accessed: 2021-03-20.

[Ins] Insee. *Revenus, pauvreté et niveau de vie en 2018 (iris)dispositif fichier localisé social et fiscal (Filosofi)*. URL: `https://www.insee.fr/fr/statistiques/5055909`.

[ins16] Insee - insee.fr. *Population en 2016*. `https://www.insee.fr/fr/statistiques/4228434`. [Accessed in 2020]. 2016.

[INS18a] INSEE. *Contours IRIS - INSEE - IGN - data.gouv.fr - data.gouv.fr*. `https://www.data.gouv.fr/en/datasets/contours-iris-insee-ign/`. [Accessed in 2022]. 2018.

[INS18b] INSEE. *Diploma - Formation en 2018 | Insee - insee.fr*. `https://www.insee.fr/fr/statistiques/5650712`. [Accessed in 2022]. 2018.

[INS18c] INSEE. *Population en 2018 | Insee - insee.fr*. `https://www.insee.fr/fr/statistiques/5650720`. [Accessed in 2022]. 2018.

[INS18d] INSEE. *Revenus, pauvret et niveau de vie en 2018 (Iris) | Insee - insee.fr*. `https://www.insee.fr/fr/statistiques/5055909`. [Accessed in 2022]. 2018.

[Isa+12] Sibren Isaacman et al. “Human Mobility Modeling at Metropolitan Scales”. In: *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*. MobiSys ’12. Low Wood Bay, Lake District, UK: ACM, 2012, pp. 239–252. ISBN: 978-1-4503-1301-8.

[Jar96] Paul A Jargowsky. “Take the money and run: Economic segregation in US metropolitan areas”. In: *American sociological review* (1996), pp. 984–998.

[Jea+16] Neal Jean et al. “Combining satellite imagery and machine learning to predict poverty”. In: *Science* 353.6301 (2016), pp. 790–794.

[JP10] Borae Jin and Jorge F Pena. “Mobile communication in romantic relationships: Mobile phone use, relational uncertainty, love, commitment, and attachment styles”. In: *Communication Reports* 23.1 (2010), pp. 39–51.

[JPF04] Ron Johnston, Michael Poulsen, and James Forrest. “The comparative study of ethnic residential segregation in the USA, 1980–2000”. In: *Tijdschrift voor economische en sociale geografie* 95.5 (2004), pp. 550–569.

[JWX07] Yu Jin, Wendi Wang, and Shiwu Xiao. “An SIRS model with a nonlinear incidence rate”. In: *Chaos, Solitons & Fractals* 34.5 (2007), pp. 1482–1497.

[Kan78] Denise B Kandel. “Homophily, selection, and socialization in adolescent friendships”. In: *American journal of Sociology* 84.2 (1978), pp. 427–436.

[Ker+21] Cliff C Kerr et al. “Covasim: an agent-based model of COVID-19 dynamics and interventions”. In: *PLOS Computational Biology* 17.7 (2021), e1009149.

[Kis14] Maria A Kiskowski. “A three-scale network model for the early growth dynamics of 2014 West Africa Ebola epidemic”. In: *PLoS currents* 6 (2014).

[KM27] William O Kermack and Anderson G McKendrick. “A contribution to the mathematical theory of epidemics”. In: *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences*. Vol. 115. 772. The Royal Society. 1927, pp. 700–721.

[Leo+16] Yannick Leo et al. “Socioeconomic correlations and stratification in social-communication networks”. In: *Journal of The Royal Society Interface* 13.125 (2016), p. 20160598.

[Lie81] Stanley Lieberson. “An asymmetrical approach to segregation.” In: *Ethnic segregation in cities.* (1981), pp. 61–82.

[Liu+16] Xi Liu et al. “Macro-scale mobile app market analysis using customized hierarchical categorization”. In: *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*. IEEE. 2016, pp. 1–9.

[Liu15] Luju Liu. “A delayed SIR model with general nonlinear incidence rate”. In: *Advances in Difference Equations* 2015.1 (2015), p. 329.

[LL17] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems* 30 (2017).

[Llo+15] Alejandro Llorente et al. “Social media fingerprints of unemployment”. In: *PloS one* 10.5 (2015), e0128692.

[LW06] Xiang Li and Xiaofan Wang. “Controlling the spreading in small-world evolving networks: stability, oscillation, and topology”. In: *IEEE Transactions on Automatic Control* 51.3 (2006), pp. 534–540.

[Mar+17] Cristina Marquez et al. “Not all apps are created equal: Analysis of spatiotemporal heterogeneity in nationwide mobile service usage”. In: *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies*. 2017, pp. 180–186.

[Mar+20] Cristina Marquez et al. “Identifying Common Periodicities in Mobile Service Demands with Spectral Analysis”. In: *2020 Mediterranean Communication and Computer Networking Conference (MedComNet)*. IEEE. 2020, pp. 1–8.

[Mel20] Angelo Mele. “Does school desegregation promote diverse interactions? An equilibrium model of segregation within schools”. In: *American Economic Journal: Economic Policy* 12.2 (2020), pp. 228–57.

[Mor+13] Jennifer N Morey et al. “Young adults’ use of communication technology within their romantic relationships and associations with attachment style”. In: *Computers in Human Behavior* 29.4 (2013), pp. 1771–1778.

[Mor75] Barrie S Morgan. “The segregation of socio-economic groups in urban areas: a comparative analysis”. In: *Urban Studies* 12.1 (1975).

[Mor91] Richard Morrill. “On the measure of geographic segregation”. In: *Geography research forum*. Vol. 11. 1991, pp. 25–36.

[MPV02] Yamir Moreno, Romualdo Pastor-Satorras, and Alessandro Vespignani. “Epidemic outbreaks in complex heterogeneous networks”. In: *The European Physical Journal B-Condensed Matter and Complex Systems* 26.4 (2002), pp. 521–529.

[MR13] M Mets and Renita Raudsepp. *Baltic Piling*. CRC Press, 2013.

[MSC01] Miller McPherson, Lynn Smith-Lovin, and James M Cook. “Birds of a feather: Homophily in social networks”. In: *Annual review of sociology* 27.1 (2001), pp. 415–444.

[Nan+08] Amit Anil Nanavati et al. “Analyzing the structure and evolution of massive telecom graphs”. In: *IEEE Transactions on Knowledge and Data Engineering* 20.5 (2008), pp. 703–718.

[Nås96] Ingemar Nåsell. “The quasi-stationary distribution of the closed endemic SIS model”. In: *Advances in Applied Probability* 28.3 (1996), pp. 895–932.

[New03] Mark EJ Newman. “The structure and function of complex networks”. In: *SIAM review* 45.2 (2003), pp. 167–256.

[Nov+13] Jakub Novak et al. “Application of mobile phone location data in mapping of commuting patterns and functional regionalization: a pilot study of Estonia”. In: *Journal of Maps* 9.1 (2013), pp. 10–15.

[Onn+07] Jukka-Pekka Onnela et al. “Structure and Tie Strengths in Mobile Communication Networks”. In: *Proceedings of the National Academy of Sciences of the United States of America* 104 (June 2007), pp. 7332–6. DOI: 10.1073/pnas.0610245104.

[Org+09] World Health Organization et al. *Pandemic H1N1 2009*. Tech. rep. WHO Regional Office for South-East Asia, 2009.

[Org20] World Health Organization. “Coronavirus disease 2019 (COVID-19): situation report, 46”. In: (2020). URL: https://apps.who.int/iris/handle/10665/331443.

[Pag+99] Lawrence Page et al. *The pagerank citation ranking: Bringing order to the web.* Tech. rep. Stanford InfoLab, 1999.

[Pap+16] Luca Pappalardo et al. “An analytical framework to nowcast well-being using mobile phone data”. In: *International Journal of Data Science and Analytics* 2.1 (2016), pp. 75–92.

[Pas+15] Romualdo Pastor-Satorras et al. “Epidemic processes in complex networks”. In: *Reviews of modern physics* 87.3 (2015), p. 925.

[PJF01] Michael Poulsen, Ron Johnston, and James Forrest. “Intraurban ethnic enclaves: introducing a knowledge-based classification method”. In: *Environment and planning A* 33.11 (2001), pp. 2071–2082.

[Pon+16] Nicolas B Ponieman et al. “Mobility and sociocultural events in mobile phone data records”. In: *Ai Communications* 29.1 (2016), pp. 77–86.

[Qui02] Lincoln Quillian. “Why is black–white residential segregation so persistent?: Evidence on three theories from migration data”. In: *Social science research* 31.2 (2002), pp. 197–229.

[RRT12] Andrea Romei, Salvatore Ruggieri, and Franco Turini. “Discovering gender discrimination in project funding”. In: *2012 IEEE 12th International Conference on Data Mining Workshops*. IEEE. 2012, pp. 394–401.

[SA14] Siiri Silm and Rein Ahas. “The temporal variation of ethnic segregation in a city: Evidence from a mobile phone use dataset”. In: *Social Science Research* 47 (2014), pp. 30–43.

[Sch71] Thomas C Schelling. “Dynamic models of segregation”. In: *Journal of mathematical sociology* 1.2 (1971), pp. 143–186.

[SDC08] Hongjing Shi, Zhisheng Duan, and Guanrong Chen. “An SIS model with infective medium on complex networks”. In: *Physica A: Statistical Mechanics and its Applications* 387.8-9 (2008), pp. 2133–2144.

[Sha+11] M Zubair Shafiq et al. “Characterizing and modeling internet traffic dynamics of cellular devices”. In: *ACM SIGMETRICS Performance Evaluation Review* 39.1 (2011), pp. 265–276.

[Sho+12] David A Shoham et al. “An actor-based model of social network influence on adolescent body size, screen time, and playing sports”. In: *PloS one* 7.6 (2012), e39795.

[Sil20] Janet Siltanen. *Locating gender: Occupational segregation, wages and domestic responsibilities*. Routledge, 2020.

[Sim05] Patrick Simon. “The measurement of racial discrimination: the policy use of statistics”. In: *International Social Science Journal* 57.183 (2005), pp. 9–25.

[Sin+19] Rajkarn Singh et al. “Urban vibes and rural charms: Analysis of geographic diversity in mobile service usage at national scale”. In: *The World Wide Web Conference*. 2019, pp. 1724–1734.

[SJ18] Anand Sahasranaman and Henrik Jeldtoft Jensen. “Ethnicity and wealth: The dynamics of dual segregation”. In: *PloS one* 13.10 (2018), e0204307.

[Son+10] Chaoming Song et al. “Limits of Predictability in Human Mobility”. In: *Science* 327.5968 (2010), pp. 1018–1021.

[Sot+11] Victor Soto et al. “Prediction of socioeconomic levels using cell phone records”. In: *International conference on user modeling, adaptation, and personalization*. Springer. 2011, pp. 377–388.

[SS12] Anurag Singh and Yatindra Nath Singh. “Nonlinear spread of rumor and inoculation strategies in the nodes with degree dependent tie strength in complex networks”. In: *arXiv preprint arXiv:1208.6063* (2012).

[SS20] MOULTON SEAN and LONG STEVEN. *The Importance of the 2020 Census, Explained in Dollars and Cents*. https://www.pogo.org/analysis/2020/03/the-importance-of-the-2020-census-explained-in-dollars-and-cents/. 2020.

[ST17] Ivan Smirnov and Stefan Thurner. “Formation of homophily in academic performance: Students change their friends rather than performance”. In: *PloS one* 12.8 (2017), e0183473.

[Sta21] Eesti Statistika. *Statistika andmebaas - Vali tabel - pub.stat.ee*. http://pub.stat.ee/. 2021.

[Ste+17] Jessica E Steele et al. “Mapping poverty using mobile phone and satellite data”. In: *Journal of The Royal Society Interface* 14.127 (2017), p. 20160690.

[Sun+06] Ron Sun et al. *Cognition and multi-agent interaction: From cognitive modeling to social simulation*. Cambridge University Press, 2006.

[Ter20] Terviseamet. *COVID-19 infoleht - terviseamet.ee*. https://www.terviseamet.ee/et/uuskoroonaviirus. 2020.

[Tow+11] S Towers et al. “Antiviral treatment for pandemic influenza: Assessing potential repercussions using a seasonally forced SIR model”. In: *Journal of theoretical biology* 289 (2011), pp. 259–268.

[Uca+21] Iñaki Ucar et al. “News or social media? Socio-economic divide of mobile service consumption”. In: *Journal of the Royal Society Interface* 18.185 (2021), p. 20210350.

[UN17] UN. *Principles and Recommendations for Population and Housing Censuses, Revision 3*. UN, 2017.

[Ves12] Alessandro Vespignani. “Modelling dynamical processes in complex socio-technical systems”. In: *Nature physics* 8.1 (2012), p. 32.

[Voi+20] Ivan Voitalov et al. “Weighted hypersoft configuration model”. In: *Physical Review Research* 2.4 (2020), p. 043157.

[Vou+20] Vasiliki Voukelatou et al. “Measuring objective and subjective well-being: Dimensions and data sources”. In: *International Journal of Data Science and Analytics* 11.4 (2020), pp. 279–309. DOI: 10.1007/s41060-020-00224-2.

[VSC16] Cécile Viboud, Lone Simonsen, and Gerardo Chowell. “A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks”. In: *Epidemics* 15 (2016), pp. 27–37.

[WF94] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press, 1994.

[Whi83] Michael J White. “The measurement of spatial segregation”. In: *American journal of sociology* 88.5 (1983), pp. 1008–1018.

[Won02] David WS Wong. “Modeling local segregation: a spatial interaction approach”. In: *Geographical and Environmental Modelling* 6.1 (2002), pp. 81–97.

[Won93] David WS Wong. “Spatial indices of segregation”. In: *Urban studies* 30.3 (1993).

[Won99] David WS Wong. “Geostatistics as measures of spatial segregation”. In: *Urban geography* 20.7 (1999), pp. 635–647.

[Xia+12] Chengyi Xia et al. “An SIR model with infection delay and propagation vector in complex networks”. In: *Nonlinear Dynamics* 69.3 (2012), pp. 927–934.

[Zha+10] Jin-Zhu Zhang et al. “Analysis of a Delayed SIR Epidemic Model”. In: *Computational Aspects of Social Networks (CASoN), 2010 International Conference on*. IEEE. 2010, pp. 192–195.

[Zho+16] Ningnan Zhou et al. “A general multi-context embedding model for mining human trajectory data”. In: *IEEE transactions on knowledge and data engineering* 28.8 (2016), pp. 1945–1958.

# Appendix A. APPLICATION CATEGORIZATION

| S.No. | Category | Application Name |
|---|---|---|
| 1 | Social media Video | Instagram Videos MP4, Youtube TLS, DailyMotion Stream http, Twitter Videos, YouTube WEB |
| 2 | Web Browsing | TLS, Google Web, Images Web, Apple Web, Default http 80, e-Commerce, Clothes, Default http 8080, Web Audience, Web Microsoft, Yahoo Web, Amazon Web, Women Websites, AMP Project, Adult Content, Other443, Home equipment, TLS Orange |
| 3 | Video | Encrypted Videos, NetFlix Video, HTTP_MP4, HTTP_STREAMING, Molotov Streaming, Streaming TS, Orange TV over TLS, GoogleVideo Web, Silverlight Adaptive streaming, WebM Streaming, Periscope Web Port, TV, Clips from miscellaneous domains, Plex |
| 4 | Social media | Facebook, SnapChat, Apple Adaptive streaming, LinkedIn, Streaming AVSP_TS, WhatsApp, Pinterest, Twitter, iMessage, VKontakte, Blogging, Odnoklassniki, imgur, Telegram, Badoo |
| 5 | Apple Download | AppStore, Apple FOTA, SIRI |
| 6 | Android Download | Google Play Store |
| 7 | Music | Deezer Streaming, Icecast Shoutcast, Spotify, Deezer Web, Radios |
| 8 | Download | DownloadWeb, HTTP File Sharing, Default HTTP Download, BITORRENT, OrangeMonReseau, FTP_Data_Passive, P2P_? |
| 9 | Advertising | Web Advertising, Small Ads |
| 10 | Apple Cloud | iCloud Storage, iCloud command |
| 11 | Games | Twitch, PlayStation Games, Steam Games, World of Warcraft, Pokemon GO, HTTP_GAMES, EA Games HTTP |
| 12 | Apple Music | iTunes Store, Apple Music Download, Apple Music Streaming |
| 13 | Cloud | Google+, Drive, We Transfer, Dropbox, SoundCloud sound sharing, Uptobox, OneFichier, Microsoft Skydrive, MegaUpload Cloud Storage |
| 14 | Apple Video | AppleAdaptive AVSP, Apple Video Download |
| 15 | Window Download | Windows Store, WindowsUpdate |
| 16 | News | NewsPaper, NewsMag, Sport News |
| 17 | Email | HTTP Mail Microsoft, HTTPS MAIL, IMAPS_Orange, IMAPS Google, IMAPS Microsoft, HTTP Mail Google, IMAPS Yahoo, IMAPS_Other, POP3_Orange, HTTP_MAIL, HTTP Mail Yahoo |
| 18 | Travel | Google NAV, Waze GPS, Transport, Tripadvisor, Uber |
| 19 | Productivity | Meet, Microsoft Office, Sharepoint, Banks, Weather, Google Docs |

# Appendix B. RMSE SCORES

| | RMSE | | | |
|---|---|---|---|---|
| Socio-Economic Features | TWS | RCA | SCU | All |
| Poverty | 6.494 | 7.082 | 7.475 | **6.399** |
| Median Income | 2733.635 | 2877.143 | 3031.349 | **2598.945** |
| Gini Index | 0.026 | 0.027 | 0.027 | **0.024** |
| No Diploma | 0.047 | 0.049 | 0.050 | **0.046** |
| BEPC or CAPBEP | 0.023 | 0.024 | 0.025 | **0.023** |
| BAC | 0.037 | 0.037 | 0.038 | **0.036** |
| SUP | 0.050 | 0.052 | 0.053 | **0.048** |
| Total population | 626.195 | 657.502 | 666.741 | **613.923** |
| Pop 0-14 | 0.030 | 0.032 | 0.032 | **0.030** |
| Pop 15-29 | 0.027 | 0.028 | 0.029 | **0.027** |
| Pop 30-44 | 0.026 | 0.028 | 0.028 | **0.026** |
| Pop 45-59 | 0.027 | 0.027 | 0.028 | **0.027** |
| Pop 60-74 | 0.034 | 0.036 | 0.037 | **0.034** |
| Pop 75+ | 0.030 | 0.032 | 0.032 | **0.030** |
| Immigrants | 0.030 | 0.033 | 0.035 | **0.029** |
| CS1 | 0.019 | 0.019 | 0.020 | 0.019 |
| CS2 | 0.018 | 0.018 | 0.018 | **0.017** |
| CS3 | 0.027 | 0.028 | 0.028 | **0.026** |
| CS4 | 0.033 | 0.034 | 0.034 | **0.032** |
| CS5 | 0.032 | 0.033 | 0.033 | **0.031** |
| CS6 | 0.036 | 0.036 | 0.037 | **0.035** |
| CS7 | 0.060 | 0.064 | 0.064 | **0.060** |
| CS8 | 0.038 | 0.039 | 0.039 | **0.037** |
| Male | 0.020 | 0.020 | 0.020 | **0.019** |
| Female | 0.020 | 0.020 | 0.020 | **0.019** |

Table 15: RMSE scores using TWS, RCA, SCU, and All predictive features for the socio-economic features. Best score is highlighted using bold text.

# ACKNOWLEDGEMENTS

I want to thank my supervisor Rajesh Sharma for his constant support and feedback throughout my Ph.D. Also, I would like to thank Anto Aasa, Angelo Furno, and Tymofii Brik for their guidance.

I also want to thank my family, friends, and especially all my teachers. As teachers don't just teach us; they help us get ready for future challenges.

# SISUKOKKUVÕTE

## Sotsiaalse heaolu kaevandamine kasutades mobiilseid andmeid

Heaolu määratletakse kui rahulikku, tervet ja õnnelikku olekut. Heaolu dimensioone on erinevaid. Näiteks tähendab füüsiline heaolu meie keha tervena hoidmist, mida saab saavutada trenni ja õige toitumisega. Samamoodi defineeritakse sotsiaalset heaolu, mis on käesoleva töö keskmes, kui võimet edukalt suhelda kohalikes ja globaalsetes kogukondades. Tänapäeva maailmas suhtlevad inimesed sõprade ja perega personaalselt , telefoni või sotsiaalmeedia platvormide, nagu Twitter, Facebook ja Instagram, kaudu.

Mobiiltelefonid on muutunud meie elu lahutamatuks osaks, seetõttu kasutame oma mobiiltelefone enamasti teistega suhtlemiseks mitte ainult helistades, vaid ka sotsiaalmeedia platvormide kaudu. Enamik inimesi, kes veedavad suure osa ajast mobiilis jätavad oma unikaalse kasutusjälje koos oma iseloomulike mustritega . Teadlased kasutavad neid mustreid laialdaselt heaolu uurimiseks kogu maailmas. Käesolevas lõputöös keskendume mobiiliandmete abil sotsiaalse heaolu uurimise kolmele mõõtmele.

Töö esimene mõõde on modelleerimine, mille käigus pakume välja mudelid, mis aitavad mõista mis tahes nakkushaiguse või epideemia levikut. Oma uurimistöös laiendasime olemasolevaid epideemiamudeleid, kaasates inimeste liikuvuse ja sotsiaalse ühenduvuse, mis on võetud CDR-i andmetest. Meie pakutud meetod on tuntud kui mobiilsuspõhine SIR-mudel. Laiendasime mudelit SIR (Susceptible – Infected – Recovered), kuna see on koronaviiruse leviku mõistmiseks kõige asjakohasem mudel. Selles mudelis on kolm klassi, milles võib inimene olla: vastuvõtlik, mis tähistab inimesi, kes ei ole nakatunud, nakatunud, mis esindab nakatunud inimesi, ja taastunud, mis tähistab nakatumisest taastunud inimesi. Esialgu on suurem osa elanikkonnast vastuvõtlikud ja ainult vähesed inimesed on nakatunud. Aja möödudes haigestuvad vastuvõtlikud inimesed ja lõpuks saavad nakatunud inimesed terveks.

Töö teine mõõde on oma olemuselt kirjeldav, kus uurime CDR-i abil ühiskondlikku segregatsiooni. Segregatsioon on määratletud kui inimeste eraldamine soo, keele või mõne muu demograafilise teabe alusel. Paljud varasemad segregatsiooni käsitlevad uurimistööd põhinevad tavapärastel valitsuse rahvaloenduse andmetel. Siiski võivad loendusandmed jäädvustada füüsilise asustuse täpse mustri, kuid harva registreerivad sotsiaalse suhtluse suundumused, mis on vajalikud sotsiaalse suhtluse olemuse põhjalikuks mõistmiseks. Uurime oma uurimistöös Eesti ühiskondlikku segregatsiooni nelja demograafilise näitaja (sugu, vanuserühm, keel ja asukoht) alusel, kasutades CDR-i andmeid. Andmestikku pakuvad Eesti suuremad sideoperaatorid, et uurida sotsiaalse suhtluse ja inimkäitumise keerukust. CDR-andmete kasutamise peamine panus segregatsiooni ennetamiseks on kulukate ja aeganõudvate loenduste kaotamine või asendamine.

Töö kolmas mõõde on olemuselt ennustav, kus prognoosime mobiilse andmeside abil piirkonna sotsiaal-majanduslikke tingimusi (SEC). Alternatiivsed andmeallikad, nagu demograafilised loendused või uuringud, mida kasutatakse SEC jälgimiseks, on hõredalt hõlmatud rahvastikuga või neid uuendatakse harva nende kõrgete kulude ja aeganõudva protsessi tõttu. Seetõttu ei sobi need kiire arenguga, mida ühiskonnad praegu kogevad. Oma uuringus näitame, et mobiilirakenduste kasutusmustrid suudaksid ennustada SEC-i piirkonnas. Mobiilsete digitaalandmete kasutamise esmane eesmärk SEC-i ennetamiseks on kaotada või asendada kulukad ja aeganõudvad loendused, et ennustada sotsiaal-majanduslikke tingimusi digitaalsete jälgede abil.

# CURRICULUM VITAE

## Personal data

| | |
|---|---|
| Name: | Rahul Goel |
| Email: | rahul.goel@ut.ee |
| Date of Birth: | 11-06-1992 |
| Citizenship: | Indian |
| Language: | Hindi, and English |

## Education

| | |
|---|---|
| 2019–Present | **Ph.D.** at Institute of Computer Science, University of Tartu, Estonia. Advised by Dr. Rajesh Sharma. |
| 2015–2017 | **Master of Technology** in Computer Science and Engineering, Analytics at National Institute of Technology Delhi, India. Advised by Dr. Anurag Singh. Thesis: Information Diffusion in Social Network using Game Theory. |
| 2010–2014 | **Bachelor of Technology** in Information Technology , YMCA University of Science and Technology, Haryana, India. |

## Employment and Research Visit

| | |
|---|---|
| Sep, 2019–Present | Junior Research Fellow, University of Tartu, Estonia. |
| Oct, 2021–Dec, 2021 | Research Visit at LICIT Lab, University Gustave Eiffel, Lyon, France. |
| Aug, 2017–Sep, 2019 | ASE at Centre for Railway Information System (CRIS), Delhi, India. |
| Jun, 2017–Aug, 2017 | Data Scientist at Impact Big Data Analytics, Delhi, India. |

## Scientific work

Main fields of interest:

- Data Science
- Machine Learning
- Social Network Analysis
- Online Social Media Platforms Analysis

## Publications

I **Rahul Goel**, and Rajesh Sharma. "Mobility based sir model for pandemics-with case study of covid-19." In 2020 **IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)**, pp. 110-117. IEEE, 2020.

II **Rahul Goel**, Loïc Bonnetain, Rajesh Sharma, and Angelo Furno. "Mobility-based SIR model for complex networks: with case study Of COVID-19." **Social Network Analysis and Mining** 11, no. 1 (2021): 1-18.

III **Rahul Goel**, Rajesh Sharma, and Anto Aasa. "Understanding gender segregation through call data records: an Estonian case study." **Plos one** 16, no. 3 (2021): e0248212.

IV **Rahul Goel**, Rajesh Sharma, and Anto Aasa. "Studying segregation in Estonia using call data records." **Social Network Analysis and Mining** 11, no. 1 (2021): 1-13.

V **Rahul Goel**, Angelo Furno, and Rajesh Sharma. "Predicting Socio-Economic Well-being Using Mobile Apps Data: A Case Study of France." under review in **IEEE transactions on Knowledge and Data Engineering**.

VI **Rahul Goel**, and Rajesh Sharma. "Understanding The MeToo Movement Through The Lens Of The Twitter." In **International Conference on Social Informatics**, pp. 67-80. Springer, Cham, 2020.

VII **Rahul Goel**, and Rajesh Sharma. "Studying leaders & their concerns using online social media during the times of crisis-A COVID case study." Social network analysis and mining 11, no. 1 (2021): 1-12.

VIII Christian Ritter, **Rahul Goel**, Rajesh Sharma. "Decolonizing Newsmaking: The Case of Climate Change Communication on YouTube during the COP26 Summit." Association of Internet Researchers (2022).

IX **Rahul Goel**, Vijayachitra Modhukur, Katrin, Andres Salumets, Rajesh Sharma, and Maire Peters. "Assessment of user's awareness of endometriosis on social media: Reddit case study" under review in **Journal of Medical Internet Research**.

X **Rahul Goel**, Modar Sulaiman, Kimia Noorbakhsh, Mahdi Sharifi, Rajesh Sharma, Pooyan Jamshidi, and Kallol Roy. "Pre-Trained Language Transformers are Universal Image Classifiers." (arxiv preprint).

XI Raj Jagtap, Abhinav Kumar, **Rahul Goel**, Shakshi Sharma, Rajesh Sharma, and Clint P. George. "Misinformation Detection on YouTube Using Video Captions." (arxiv preprint).

# ELULOOKIRJELDUS

## Isiklikud andmed

| | |
|---|---|
| Nimi: | Rahul Goel |
| E-post: | rahul.goel@ut.ee |
| Sünniaeg | 11-06-1992 |
| Kodakondsus: | Indialane |
| Keelteoskus: | Hindi, Inglise |

## Haridus

| | |
|---|---|
| 2019–hetkel | Ph.D. arvutiteaduses, Tartu Ülikool, Eesti. Nõustab dr Rajesh Sharma. |
| 2015–2017 | Tehnoloogia magister arvutiteaduses ja -tehnoloogias, analüütika Delhi Riiklikus Tehnoloogiainstituudis, India. Nõustaja dr Anurag Singh. Lõputöö: Info levitamine sotsiaalvõrgustikus, kasutades mänguteooriat. |
| 2010–2014 | Infotehnoloogia bakalaureusekraad, YMCA Teadus- ja Tehnoloogiaülikool, Haryana, India. |

## Teenistuskäik

| | |
|---|---|
| Sept, 2019–hetkel | Nooremteadur, Tartu Ülikool, Eesti. |
| Okt, 2021–Dets, 2021 | Teaduskülastus LICIT Lab'i, Gustave Eiffel Ülikool, Lyon, Prantsusmaa. |
| Aug, 2017–Sept, 2019 | Tarkvarainseneri assistant (ASE), Raudtee Infosüsteemi Keskus (CRIS), Delhi, India. |
| Juuni, 2017–Aug, 2017 | Andmeteadlane ettevõttes Impact Big Data Analytics, Delhi, India. |

## Teadustegevus

Peamised uurimisvaldkonnad:

- Andmeteadus
- Masinõpe
- Sotsiaalvõrgustike analüüs
- Veebipõhiste sotsiaalmeedia platvormide analüüs

## Väljaanded

I **Rahul Goel**, and Rajesh Sharma. "Mobility based sir model for pandemics-with case study of covid-19." In 2020 **IEEE/ACM International Confe-**

**rence on Advances in Social Networks Analysis and Mining (ASONAM)**, pp. 110-117. IEEE, 2020.

II **Rahul Goel**, Loïc Bonnetain, Rajesh Sharma, and Angelo Furno. "Mobility-based SIR model for complex networks: with case study Of COVID-19." **Social Network Analysis and Mining** 11, no. 1 (2021): 1-18.

III **Rahul Goel**, Rajesh Sharma, and Anto Aasa. "Understanding gender segregation through call data records: an Estonian case study." **Plos one** 16, no. 3 (2021): e0248212.

IV **Rahul Goel**, Rajesh Sharma, and Anto Aasa. "Studying segregation in Estonia using call data records." **Social Network Analysis and Mining** 11, no. 1 (2021): 1-13.

V **Rahul Goel**, Angelo Furno, and Rajesh Sharma. "Predicting Socio-Economic Well-being Using Mobile Apps Data: A Case Study of France." under review in **IEEE transactions on Knowledge and Data Engineering**.

VI **Rahul Goel**, and Rajesh Sharma. "Understanding The MeToo Movement Through The Lens Of The Twitter." In **International Conference on Social Informatics**, pp. 67-80. Springer, Cham, 2020.

VII **Rahul Goel**, and Rajesh Sharma. "Studying leaders & their concerns using online social media during the times of crisis-A COVID case study." Social network analysis and mining 11, no. 1 (2021): 1-12.

VIII Christian Ritter, **Rahul Goel**, Rajesh Sharma. "Decolonizing Newsmaking: The Case of Climate Change Communication on YouTube during the COP26 Summit." Association of Internet Researchers (2022).

IX **Rahul Goel**, Vijayachitra Modhukur, Katrin, Andres Salumets, Rajesh Sharma, and Maire Peters. "Assessment of user's awareness of endometriosis on social media: Reddit case study" under review in **Journal of Medical Internet Research**.

X **Rahul Goel**, Modar Sulaiman, Kimia Noorbakhsh, Mahdi Sharifi, Rajesh Sharma, Pooyan Jamshidi, and Kallol Roy. "Pre-Trained Language Transformers are Universal Image Classifiers." (arxiv preprint).

XI Raj Jagtap, Abhinav Kumar, **Rahul Goel**, Shakshi Sharma, Rajesh Sharma, and Clint P. George. "Misinformation Detection on YouTube Using Video Captions." (arxiv preprint).

# DISSERTATIONES INFORMATICAE PREVIOUSLY PUBLISHED IN DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 lk.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.

74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.

113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.

# DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

1. **Abdullah Makkeh**. Applications of Optimization in Some Complex Systems. Tartu 2018, 179 p.
2. **Riivo Kikas**. Analysis of Issue and Dependency Management in Open-Source Software Projects. Tartu 2018, 115 p.
3. **Ehsan Ebrahimi**. Post-Quantum Security in the Presence of Superposition Queries. Tartu 2018, 200 p.
4. **Ilya Verenich**. Explainable Predictive Monitoring of Temporal Measures of Business Processes. Tartu 2019, 151 p.
5. **Yauhen Yakimenka**. Failure Structures of Message-Passing Algorithms in Erasure Decoding and Compressed Sensing. Tartu 2019, 134 p.
6. **Irene Teinemaa**. Predictive and Prescriptive Monitoring of Business Process Outcomes. Tartu 2019, 196 p.
7. **Mohan Liyanage.** A Framework for Mobile Web of Things. Tartu 2019, 131 p.
8. **Toomas Krips.** Improving performance of secure real-number operations. Tartu 2019, 146 p.
9. **Vijayachitra Modhukur.** Profiling of DNA methylation patterns as biomarkers of human disease. Tartu 2019, 134 p.
10. **Elena Sügis.** Integration Methods for Heterogeneous Biological Data. Tartu 2019, 250 p.
11. **Tõnis Tasa.** Bioinformatics Approaches in Personalised Pharmacotherapy. Tartu 2019, 150 p.
12. **Sulev Reisberg.** Developing Computational Solutions for Personalized Medicine. Tartu 2019, 126 p.
13. **Huishi Yin.** Using a Kano-like Model to Facilitate Open Innovation in Requirements Engineering. Tartu 2019, 129 p.
14. **Faiz Ali Shah.** Extracting Information from App Reviews to Facilitate Software Development Activities. Tartu 2020, 149 p.
15. **Adriano Augusto**. Accurate and Efficient Discovery of Process Models from Event Logs. Tartu 2020, 194 p.
16. **Karim Baghery.** Reducing Trust and Improving Security in zk-SNARKs and Commitments. Tartu 2020, 245 p.
17. **Behzad Abdolmaleki.** On Succinct Non-Interactive Zero-Knowledge Protocols Under Weaker Trust Assumptions. Tartu 2020, 209 p.
18. **Janno Siim.** Non-Interactive Shuffle Arguments. Tartu 2020, 154 p.
19. **Ilya Kuzovkin.** Understanding Information Processing in Human Brain by Interpreting Machine Learning Models. Tartu 2020, 149 p.
20. **Orlenys López Pintado.** Collaborative Business Process Execution on the Blockchain: The Caterpillar System. Tartu 2020, 170 p.
21. **Ardi Tampuu.** Neural Networks for Analyzing Biological Data. Tartu 2020, 152 p.

22. **Madis Vasser.** Testing a Computational Theory of Brain Functioning with Virtual Reality. Tartu 2020, 106 p.
23. **Ljubov Jaanuska.** Haar Wavelet Method for Vibration Analysis of Beams and Parameter Quantification. Tartu 2021, 192 p.
24. **Arnis Parsovs.** Estonian Electronic Identity Card and its Security Challenges. Tartu 2021, 214 p.
25. **Kaido Lepik.** Inferring causality between transcriptome and complex traits. Tartu 2021, 224 p.
26. **Tauno Palts.** A Model for Assessing Computational Thinking Skills. Tartu 2021, 134 p.
27. **Liis Kolberg.** Developing and applying bioinformatics tools for gene expression data interpretation. Tartu 2021, 195 p.
28. **Dmytro Fishman.** Developing a data analysis pipeline for automated protein profiling in immunology. Tartu 2021, 155 p.
29. **Ivo Kubjas.** Algebraic Approaches to Problems Arising in Decentralized Systems. Tartu 2021, 120 p.
30. **Hina Anwar.** Towards Greener Software Engineering Using Software Analytics. Tartu 2021, 186 p.
31. **Veronika Plotnikova.** FIN-DM: A Data Mining Process for the Financial Services. Tartu 2021, 197 p.
32. **Manuel Camargo.** Automated Discovery of Business Process Simulation Models From Event Logs: A Hybrid Process Mining and Deep Learning Approach. Tartu 2021, 130 p.
33. **Volodymyr Leno.** Robotic Process Mining: Accelerating the Adoption of Robotic Process Automation. Tartu 2021, 119 p.
34. **Kristjan Krips.** Privacy and Coercion-Resistance in Voting. Tartu 2022, 173 p.
35. **Elizaveta Yankovskaya.** Quality Estimation through Attention. Tartu 2022, 115 p.
36. **Mubashar Iqbal.** Reference Framework for Managing Security Risks Using Blockchain. Tartu 2022, 203 p.
37. **Jakob Mass.** Process Management for Internet of Mobile Things. Tartu 2022, 151 p.
38. **Gamal Elkoumy.** Privacy-Enhancing Technologies for Business Process Mining. Tartu 2022, 135 p.
39. **Lidia Feklistova.** Learners of an Introductory Programming MOOC: Background Variables, Engagement Patterns and Performance. Tartu 2022, 151 p.
40. **Mohamed Ragab.** Bench-Ranking: A Prescriptive Analysis Approach for Large Knowledge Graphs Query Workloads. Tartu 2022, 158 p.
41. **Mohammad Anagreh.** Privacy-Preserving Parallel Computations for Graph Problems. Tartu 2023, 181 p.