UNIVERSITÀ DEGLI STUDI DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE

ICT International Doctoral School

CYCLE XXXII

MODELING HUMAN AND CITIES' BEHAVIORS: FROM COMMUNICATION SYNCHRONIZATION TO SPATIO-TEMPORAL NETWORKS

LORENZO CANDEAGO

Advisor:

Bruno Lepri

Fondazione Bruno Kessler, Trento

Industrial advisor:

Mattia Pasolli

TIM - Telecom Italia, Trento

ACADEMIC YEAR 2018/2019

ABSTRACT

Recent years have seen a huge increase in the amount of data collected from multiple sources: mobile phones are ubiquitous, social networks are widely used, cities are more and more connected and the mobility of people and goods has risen to a global scale. The *Big Data Era* has opened the doors to new kinds of studies that were unthinkable with previous qualitative methods: human behavior can now be analyzed with a fine-grained resolution, patterns of mobility and behavior can be extracted from the incredible amount of data collected every day. Modern large cities are becoming more and more interconnected and this phenomenon leads to an increasing communication and activities' synchronization. Due to the amount of data available or for anonymization reasons, it is often necessary to aggregate data spatially and temporally. A natural representation of clustered mobility data is the temporal network representation. In this thesis we focus on these two aspects of spatial distance in human mobility: (i) we study the synchronization of 76 Italian cities, using mobile phone data, showing that both distance between cities and city size determine the synchronization in communication rhythms. Moreover, we show that the effect of the distance in synchronization decreases when the size of the city increases; (ii) we investigate how clustering continuous spatio-temporal data affects spatio-temporal network measures for real-life and synthetic datasets and analyze how spatio-temporal networks' measures vary at different aggregation levels.

PUBLICATIONS

This thesis is based on ideas and figures presented in the following publications:

[1] Lorenzo Candeago, Giulia Bertagnolli, Paolo Bosetti, Michele Vescovi, Francesco Sacco, and Bruno Lepri. "Cities of a feather flock together: a study on the synchronization of communication between Italian cities." In: *EPJ Data Science* 8.1 (2019), p. 19.

[2] Lorenzo Candeago, Lorenzo Lucchini, Bruno Lepri, and Mirco Musolesi. "Spatial clustering and network topology: a study on dynamical networks." 2020. To be submitted.

Other publications that are not included in this manuscript:

[1] Nadia C Fabrizio, Bruno Lepri, Elisa Rossi, Andrea Martini, Dimitar Anastasovski, Paolo Cappello, and Lorenzo Candeago. "Invoice Discounting: a blockchain-based approach." In: *Frontiers in Blockchain* 2 (2019), p. 13.

[2] Lorenzo Candeago, Daniel Larraz, Albert Oliveras, Enric Rodrıéguez-Carbonell, and Albert Rubio. "Speeding up the constraint-based method in difference logic." In: *International Conference on Theory and Applications of Satisfiability Testing*. Springer. 2016, pp. 284–301.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

1.1 CONTEXT

The core aim of the social sciences has always been to unveil and explore the complex pattern of people's behavior [1]. Social scientists have used all the tools in their possession to investigate a range of human actions ranging from financial activities to buying preferences [2] and social mobility [3], in order to uncover new knowledge on these issues. Technological advancements have been deeply integrated into the life of the individual as well as the society as a whole: as technological progress goes by, human life is shaped by modern technologies that not only influence people's behaviors [4], but also allows for innovative approaches to study human mobility [5], psychological aspects of human life, such as stress levels [6], personality traits [7] and mental health [8]. Through collecting data from mobile phones [9, 10, 11, 8], social media [12, 13] and credit cards [14, 15], innovative research is made possible, giving birth to what is called *computational social science* [16, 14]. The *Big data era* has allowed for studies that would have been inconceivable with qualitative approaches, expanding greatly in sample size and span of time that can be observed. Unobtrusive data collection leads to reducing the possibility of tampering with the habits of study subjects [17, 18] that, knowing they are being observed, might change their behavior [19].

In our work, we analyze how spatial distance influences the relationship between cities and, in a temporal network framework, how spatial clustering affects temporal network's metrics for continuous spatio-temporal data.

One of the possible subjects of study, thanks to the global availability of data, are cities, and in particular *global cities*. Modern large cities are becoming more interconnected [20] leading to an increase in communication and activities' synchronization [21, 22, 23, 10, 24]. Urban sociology literature [25, 26] has investigated the synchronization due to globalization, where large companies tend to spread their headquarters in different cities and countries. This literature [25,

26] has also introduced the concept of *gateway cities*, namely central nodes for all the communication and economical activities, and for the flow of people to and from the region where the city is located. To explore synchronization and mobility, *Call Detail Records* (CDR) are often used [22, 10]; mobile network operators collect these for billing purposes. CDRs contain information about user's calls, locations, and communications. We will see that both spatial distance and city size play a key role in determining cities' synchronization.

Another natural approach to spatial data, and in particular mobility data, is graph theory. Each stop point or check-in location can be represented as a node in a graph and the movement between two locations can be represented as a directed edge. If we do not consider the temporal component of the mobility and pick only the links between nodes, no matter the time when they have been activated, we have what is called *static graph*. Static graphs have been successfully used to explore human mobility (e.g. in [5, 27, 28]). To take into account the temporal aspect of mobility, we can employ what is called *spatio-temporal graphs* or *spatio-temporal network* [29]. A spatio-temporal network is an ordered collection of graphs. Each graph of the network contains the same set of nodes and captures the dynamics of the underlying system during a time frame t. At each time frame t, the graph contains the edges that were active in the corresponding time interval.

It has been shown that temporal graphs capture different aspects of mobility [30]. Spatio-temporal networks have been successfully employed as an example to optimize the immunization process of a population [31], to enhance routing algorithms [32] and to explore how different brain regions interact [33]. Classical static graph metrics, such as betweenness or characteristic path length can be adapted to the temporal-network case [34], capturing also the temporal relationship between each node and event of a network. Recent regulations, such as the General Data Protection Regulation (GDPR) [35], focused on a more careful diffusion of individuals' data, stressing the importance of data ownership and share. A standard approach to anonymization of individual CDR and mobility data is to aggregate the individual data at a coarser spatial and temporal resolution. Although it has been shown that anonymization techniques based on coarser data aggregation are often insufficient [36, 15], these techniques are still widely used, hence the need of a study on how these aggregations influence the spatio-temporal networks' metrics. For both privacy and computational

reasons, mobility traces such as gps traces are often aggregated spatially and temporally. In our work, we consider different continuous mobility traces and discretize them both temporally and spatially. The effect of temporal aggregation has already been explored in [37, 38, 39]. The novelty of our study is dual: to our knowledge, this is the first work on the effects of spatial aggregations on spatio-temporal networks' metrics; furthermore, we explore synthetic models that reproduce the behavior of temporal network's measures for real-life datasets or characterize different types of mobility (high mixing, low mixing, global-scale distance, and city-scale distance) based on temporal network metrics.

1.2 RESEARCH DIRECTIONS

Given the availability of CDR data and large scale spatio-temporal datasets, we took advantage of the information collected to investigate the spatial aspects on human behavior. In this thesis we focus on two main topics:

i. Understanding and analyzing global cities' synchronization using CDRs.

ii. Studying how spatio-temporal networks' metrics behave when clustering continuous GPS data at different resolutions, exploring two different synthetic mobility models (random waypoint and s-EPR) and characterizing the mobility based on temporal network measures.

In our first work, we explore the communication synchronization between 76 Italian cities of different sizes by using mobile phone data. Our results show that both the spatial distance and the size of the city influence the synchronization: larger cities are more similar to larger cities in communication rhythms than medium cities are to medium cities, and medium cities are more similar to medium cities than smaller cities are to smaller cities. Furthermore, for all the cities' sizes we observe a drift in similarity due to spatial distance. Interestingly, the drift due to distance over similarity is less strong in large cities, that act as gateway nodes for the Italian economical system, hence having an emerging strongly connected and synchronized network, than for medium and small cities, that are more bound to local industries. Finally, our results also show that highly synchronized cities have greater trends of immigration and increase in wealth.

In our second work, we focus on temporal networks generated from geo-spatial data. In particular, we study how spatial clustering affects the topological features of location-based dynamic networks. We show how a coarse-graining process based on geographical distance redesigns the network structure by changing both the nodes and the edges among them. We found that different type of mobility–high mixing mobility, where a trajectory rarely returns to the same point and low mixing mobility, where a set of preferential locations such as home or work place are visited more often–have different behavior with respect to the metrics considered. This study is intended to be a guide to mobility modeling and clustering calibration in a dynamic network environment.

1.3 DISSERTATION OUTLINE

This thesis is organized as follows:
In Chapter 2 we present our first study about cities' synchronization. In section 2.3 we describe the mobile phone data and the socio-economic indicators (section 2.3.1), we introduce the Dynamic Time Warping distance algorithm (section 2.3.2) that we use for computing the synchronization among the communication activity timeseries of our cities, and finally we describe the bootstrap resampling procedure used (section 2.3.3). In section 2.4 we present our results on communication synchronization and on the influence played by the city size and the spatial distances as well as the associations between the synchronization of calling patterns of a city and the city's socio-economic indicators. Moreover, in section 2.5 we discuss the obtained results with regard to urban sociology literature. In Chapter 3 we present our preliminary results on spatio-temporal netwoks. In section 3.3 we introduce the methodology and data used for this study. In particular, we give an introduction on temporal networks (section 3.3.2) and we present the spatio-temporal metrics we considered (section 3.3.3); we then introduce the datasets we used (section 3.3.1), we describe the sampling procedure we implemented for real-life datasets in section 3.3.4 and describe the synthetic models we used in section 3.3.5. We present our results on temporal networks in section 3.4 and discuss them.

Finally, in Chapter 4, we present a summary of the main results of this dissertation, initially discussing the practical and theoretical implications of our work, and then pointing out some limitations.

2

CITIES OF A FEATHER FLOCK TOGETHER: A STUDY ON THE SYNCHRONIZATION OF COMMUNICATION BETWEEN ITALIAN CITIES

2.1 BACKGROUND

Synchronization is a spontaneous process that emerges in many domains in nature [40], from neurons [41], trees [42], animals [43], and up to human beings [24, 21, 23].

Contemporary large cities are becoming more and more interconnected [20] and this phenomenon leads to an increase in communication and activities' synchronization, as observed in Morales *et al.* [22]. Recent urban sociology literature [25, 26] has investigated the synchronization due to globalization, where large companies tend to spread their headquarters in different cities and countries. This literature [25, 26] has also introduced the concept of *gateway cities*, namely central nodes for all the communication and economical activities, and for the people flows from and to the region where the city is located. Recently, studies on cities' and human activity synchronization based on CDR have shown that, within communities, social capital measures (referendum turnout, blood donations and association density) correlate with high community synchronization[23]. In [22], it has been shown using geolocated Twitter data that there exist a global synchronization of large cities across the world, leading to an interdependence of behavior that can be observed in synchronization of communication; furthermore, synchronization between global cities has been explored by the means of mobile phone usage patterns in [44]: in the study, Grauwin *et al.* show that three global cities (New York, London and Hong Kong) share common mobile phone usage patterns in business areas. In our work, we explore the communication synchronization between 76 Italian cities of different sizes by using mobile phone data (i.e., Call Detail Records) and investigate if and which Italian cities act as *gateway cites*. We also explored how the synchronization between couples of cities changes depending on the

size of the city and the spatial distance between them. We found that both spatial distance and city size influence the synchronization: larger cities are more similar to larger cities in communication rhythms than medium cities to medium cities, and medium cities are more similar to medium cities than smaller cities to smaller cities. Moreover, for all the cities' sizes we observed a drift in similarity due to spatial distance. In addition, we have also investigated if cities with a higher average synchronization tend to be richer and to attract more people from other places. Our results show that highly synchronized cities have a higher percentage of foreign-born population and higher levels of average yearly income per tax payer.

2.2 CONTRIBUTION

In our work, we preset a novel study exploring the communication synchronization between 76 Italian cities of different sizes by using mobile phone data: building upon the results of the previous studies, we further explore cities' synchronization, focusing on Italy, by considering not only global cities as previous studies, but also medium and small cities, and by exploring the effects of distance and city size on cities' synchronization. Our results show that both the spatial distance and the size of the city influence the synchronization: larger cities are more similar to larger cities in communication rhythms than medium cities are to medium cities, and medium cities are more similar to medium cities than smaller cities are to smaller cities. Furthermore, for all the cities' sizes we observe a drift in similarity due to spatial distance. Interestingly, the drift due to distance over similarity is less strong in large cities, that act as gateway nodes for the Italian economical system, hence having an emerging strongly connected and synchronized network, than for medium and small cities, that are more bounded to local industries. Finally, our results also show that highly synchronized cities are more attractive for foreign-born population and are richer.

ABBREVIATIONS

CDR: Call Detail Records; CRM: Customer Relations Management; DTW: Dynamic Time Warping; LAU: Local Administrative Unit; OECD Organization for Economic Co-operation and Development; WLS: Weighted Least-Square Regression.

2.3 METHODOLOGY

2.3.1 *Data Description*

Our dataset consists of 24 consecutive days (18 weekdays and 6 weekend days) of Call Detail Records (CDRs) data, inclusive of 11.4B outgoing mobile calls of TIM, one of the major Italian telecommunication companies (30.8% of market share in Italy[1]).

CDRs are collected for billing purposes by mobile network operators: more specifically, a CDR record of the user is created every time a phone interacts with the network, recording (i) the type of the event (incoming/outgoing call, transmission of a text message, consumption of a certain amount of data traffic), (ii) the pseudonym of the users involved (the one producing traffic and, eventually, e.g., in case of voice traffic, the other party involved), (iii) the timestamp of the event, and (iv) the cell network's antenna accessed for the event (i.e., to which the caller's phone was connected), that, to a wider extent, represents the location of the user [45, 46].

The CDRs of our dataset are limited to voice traffic and have been provided by TIM after some pre-processing steps. First of all, CDRs have been enriched with demographic data from the Customer Relations Management, in order to be able to represent users in terms of gender and age range. CDRs have then been filtered at the 99th percentile based on the number of daily calls per user, in order to remove edge cases that are not representative of the general population (e.g., call centers). In particular, if the number of calls for a user during a day exceeds the threshold, all the CDRs associated with that user for that day are removed from the dataset. Finally, data have been aggregated by

1 https://www.statista.com/statistics/710559/mobile-network-provider-market-share-italy/

city, hour, gender and age-range, getting rid of the identities (even if already pseudoanonimized) of users. Thus, for each city and hour, the dataset contains: (i) the number of outgoing calls divided by gender, (ii) the number of outgoing calls divided by age range, and (iii) the total number of outgoing calls.

Regarding the identification of our cities, we have adopted the 2012 definition developed jointly by the European Commission and the Organization for Economic Co-operation and Development (OECD) [47]: a city is a local administrative unit (LAU) where the majority of the population lives in an urban centre of at least 50 000 inhabitants. The definition provides also a division of European cities into 6 size classes: S, M, L, XL, XXL and Global City. We have considered 76 Italian cities that fall into the OECD definition and grouped them in *Small* (S), *Medium* (M), *Large* (L, XL, XXL). According to OECD definition no city in Italy can be categorized as Global City, since no Italian city has more than 5 million inhabitants.

Hence, if we define $calls_h(c,d)$ as the number of calls for a city c, during a day d and an hour h, the timeseries of the calls, or *city's activity pattern* (A), is a timeseries of the values $A_h(c,d)$ where

$$A_h(c,d) = \frac{calls_h(c,d)}{\sum_{h \in [0,23]} calls_h(c,d)}$$

It is worth highlighting that we are considering the percentage of calls over the day for each hour and city. Thus, we can compare different cities independently of the absolute number of outgoing calls.

Finally, we identify the following socio-economic indicators to investigate the economic role (i.e., city's wealth), the attractiveness for foreigners (e.g., immigrants), and the incoming and outgoing commuting patterns of the highly synchronized Italian cities:

- **Resident population**: The absolute number of the resident population in a city[2].
- **Foreign population**: The absolute number of the foreign-born population in a city[2].

2 15° Censimento generale della popolazione e delle abitazioni, ISTAT, 2011, https://www.istat.it/it/censimenti-permanenti/censimenti-precedenti/popolazione-e-abitazioni/popolazione-2011

- **Population density**: The ratio between the resident population and the city surface[2].
- **Foreign percentage**: The percentage of the foreign population over the resident population for a city[2].
- **Average income**: The average yearly income per tax payer[2].
- **In-out commuters ratio**: The ratio between commuters moving to a city X for work or study reasons and commuters moving from a city X for work or study reasons[3].
- **Incoming commuter ratio**: The ratio between commuters moving to a city X for work or study reasons and the resident population of that city[3].
- **Outgoing commuter ratio** The ratio between commuters moving out from a city X for work or study reasons and the resident population of that city[3].

2.3.2 *Dynamic Time Warping*

In order to compute the synchronization between the activity patterns of each pair of our cities, we have used the Dynamic Time Warping (DTW) distance algorithm [48]. DTW has been extensively adopted in speech recognition [49], computer vision [50, 51], natural language processing [52, 53], and image matching and handwritten recognition [54] as a measure of similarity between timeseries. The algorithm provides an estimate of the optimal match between two timeseries, including possible compression, expansion or lags in sections of the sequences. For example, DTW can capture similarities in walking activities, even if an individual is walking faster than the other. Thus, DTW can remove the lag due to the circadian rhythms characterizing our timeseries [55, 56]. DTW provides a more correct notion of similarity between cities' activity patterns than an approach based on sliding-window correlation due to the less strict assumptions:it has been shown [57, 58] that the DTW has similar results to other time series alignment methods such as windowed cross-correlation, but

3 15° Censimento generale della popolazione e delle abitazioni, Matrici del pendolarismo, ISTAT, 2011 https://www.istat.it/it/archivio/139381

doesn't require the velocity of the perturbation to be homogeneous nor the time shifts to be constant or linearly increasing over a given analysis window.

More specifically, assuming two timeseries $X = (x_1, \cdots, x_M)$ and $Y = (y_1, \cdots, y_N)$ a DTW path $P = (p_1, \cdots, p_K)$ is a sequence of tuples of indices where $p_k = (m_k, n_k) \in [1, \cdots, M] \times [1, \cdots, N]$ is subject to the following constraints:

1. $p_1 = (1, 1)$ and $p_K = (M, N)$
2. $m_1 \leq m_2 \leq \cdots m_K$ and $n_1 \leq n_2 \leq \cdots n_K$
3. $p_{k+1} - p_k \in \{(1, 0), (0, 1), (1, 1)\}$ for $k \in [1 : K - 1]$

Given a distance function d (e.g., Euclidean distance), the cost of a path c_p is defined as $c_p(X, Y) = \sum_{k=1}^{K} d(x_{m_k}, y_{n_k})$. The DTW distance between X and Y is hence defined as the cost of the wrapping path $p^\star$ having minimal total cost among all the possible wrapping paths.

By considering the activity pattern timeseries associated with the activity level of a city, we have obtained the DTW distance between the timeseries of all the couples of cities for a given day. Therefore, the higher the DTW distance between a couple of cities, the lower the synchronization of their activity pattern timeseries. Moreover, we have computed the mean and variance of the DTW distances, during weekdays and weekends, for each couple of cities. Mean and variance are estimated by using the jackknife resampling procedure [59].

In order to investigate the association between the DTW distances and the socio-economic indicators listed in section 2.3.1, for each city we have considered the average of the means previously computed using the jacknife resampling method. Then, we have computed the variance-weighted average of the DTW distances for each city by using the inverse-variance weighting procedure [60]. This method permits aggregation of two or more random variables (i.e., DTW distances) to minimize the variance of the weighted average.

Finally, the variance-weighted average of the DTW distances for each city is associated to each of the socio-economic indicators by means of Spearman's bivariate correlations. The Spearman's bivariate correlation measures the strength and direction of the association between two variables. Specifically, the Spearman's coefficient is a number between -1 and +1, where -1 means perfect negative correlation, +1 indicates perfect positive correlation and 0 indicates no correlation.

2.3.3 *Bootstrap Procedure*

Since we didn't have a way to asses the error on the dataset, we estimated the parameters' variance we used a bootstrap procedure. Bootstrap resampling method is a widely used technique to infer properties of an estimator by sampling the original data repeatedly. We have performed a group bootstrap by extracting a city and adding to our bootstrap sample all the couples containing the extracted city. This procedure guarantees to preserve, at each bootstrap iteration, all the correlations that a city has with other cities within the bootstrap sample is preserved, since no couples that include the selected city are left out. Our bootstrap procedure follows three steps:

(i) For each group of n cities of the same size (*Large*, *Medium* and *Small*) extract n cities with replacement;

(ii) Create the dataset with couples of cities for the bootstrap iteration using all the possible combinations of extracted cities (excluding the couples with the same city);

(iii) Perform a Weighted Least-Square Regression (WLS) on DTW and cities' distance using as weights the variances previously computed using the jackknife sampling method.

As an example, in one bootstrap iteration we extract *Milan, Naples, Rome*, all the possible tuples of city will be added to the dataset (i.e. {Milan-Rome, Milan-Naples, Milan-Naples, Naples-Rome}) For each bootstrap iteration we implement a Weighted Least-Square Regression on DTW and cities' distance and collect the values of the slope m and the intercept q of the fit. Finally, obtained results were evaluated by performing a T-test to asses whether the slope differs from zero.

2.4 RESULTS

The activity level of a city is the result of the combined behavioural patterns of different agents (i.e., individuals) and external constraints such as working schedules, school timetables and vacations. Such activity is mirrored by the number of calls placed in a city during a day: therefore, we have considered the

percentage of outgoing calls per hour in each city (*city's activity timeseries*) as a proxy of the activity level of the city over time.

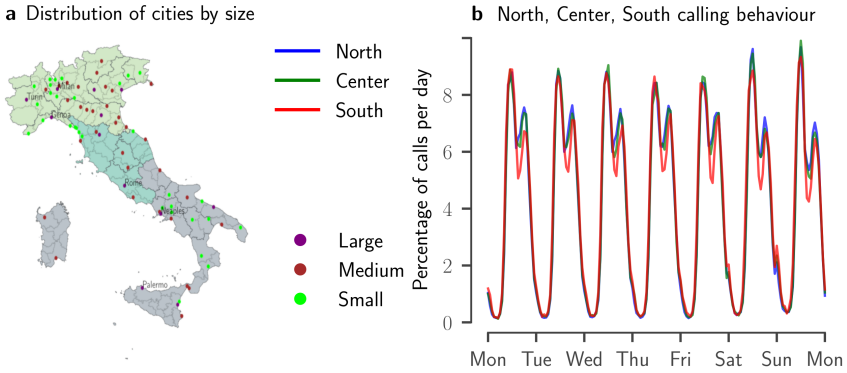

Figure 2.1: **a)** The 76 selected cities, their geographical position in Italy, and their size according to the OECD categorization. **b)** Average volume of outgoing calls as a percentage by hour for the cities in the North, Center and South of Italy. We can see that during weekdays the drops in percentage of calls for South of Italy cities are delayed compared to the behavior of the ones from the North and the Center.

Our sample equally represents cities of classes *Medium* and *Small*, while we have less *Large* cities. Moreover, the cities we have considered are evenly spread across all regions of Italy, as can be seen in Figure 2.1a. Interestingly, by examining the activity timeseries associated with North, Center and South of Italy (see Figure 2.1b), a similar pattern emerges for North and Center while South is characterized by a shift in the drop during lunchtime. For this reason, we have used DTW distance to compute the synchronization between *cities' activity timeseries*. Indeed, DTW distance removes the influence of the observed lags due to circadian rhythms (see section 2.3.2).

In Figure 2.2, each cell of the heatmap represents the mean value over weekdays of the DTW distance between two cities. The cities on both axes are ordered by total volume of calls from lowest (top left corner) to highest (bottom right corner). Two emerging clusters can be observed: one at the top left corner, where smaller cities with lower call volume are located, tends to have larger average distance between the communication activity timeseries; the other one can be observed in the bottom right corner, where large cities (with higher call volume) tend to have smaller mean DTW distance. The mean value of the DTW distance roughly increases when the volume of calls for a city decreases (i.e., it roughly scales with the size of the city): thus, the mean DTW distance for

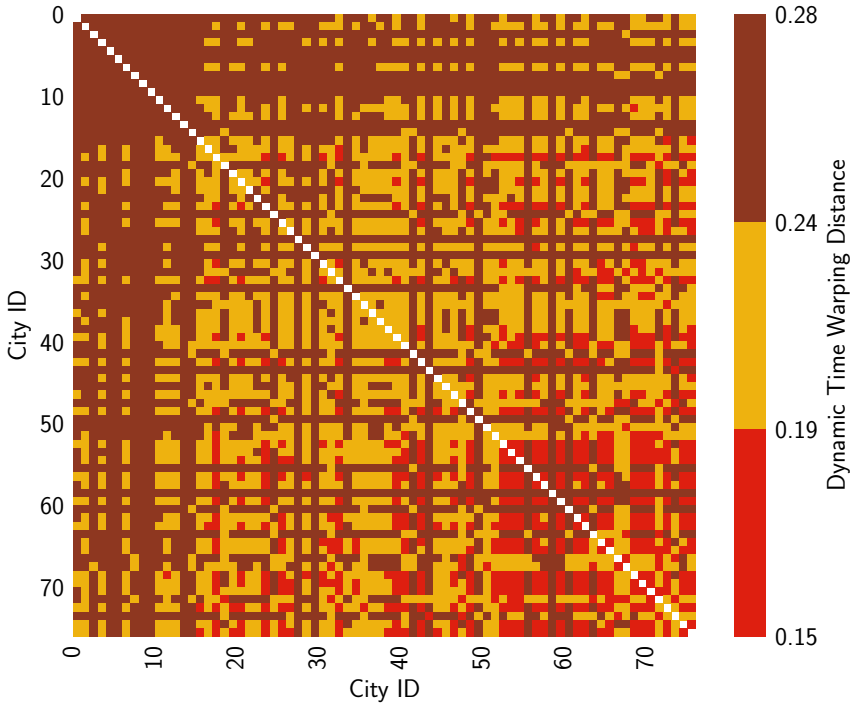

Figure 2.2: **Heatmap for Dynamic Time Warping Distance** Each cell of this heatmap represents the mean value over weekdays of the DTW distance between two cities. In the x and y axes the cities are sorted by volume of calls from lower (top left corner) to higher (bottom right corner). Blanks in the diagonal represent the DTW distance between each city and itself.

medium-call-volume cities is lower than the one for smaller cities, but higher than the DTW distance for larger cities.

As previously said, we have performed a bootstrap for cities of the same size (*Large*, *Medium* and *Small*) and a WLS is fitted using the mean and the variance of the DTW distance between cities' activity pattern timeseries. As seen in Figure 2.2, DTW distance roughly decreases when the number of outgoing calls increases and we can roughly divide the cities into three clusters based on DTW distance. Remarkably, this relationship still holds when considering the division of Italian cities into three size classes (*Large*, *Medium* and *Small*) according to OECD definition (see section 2.3.1 for details). Cities of the same size appear to be more similar: two large cities (such as Turin and Milan) are more similar than two medium cities (such as Padua and Modena), and medium cities are more similar than small cities.

As we can see in Figure 2.3, the similarity decreases as the distance between cities increases, in a consistent way for all classes (see Table 2.A.1 for a complete

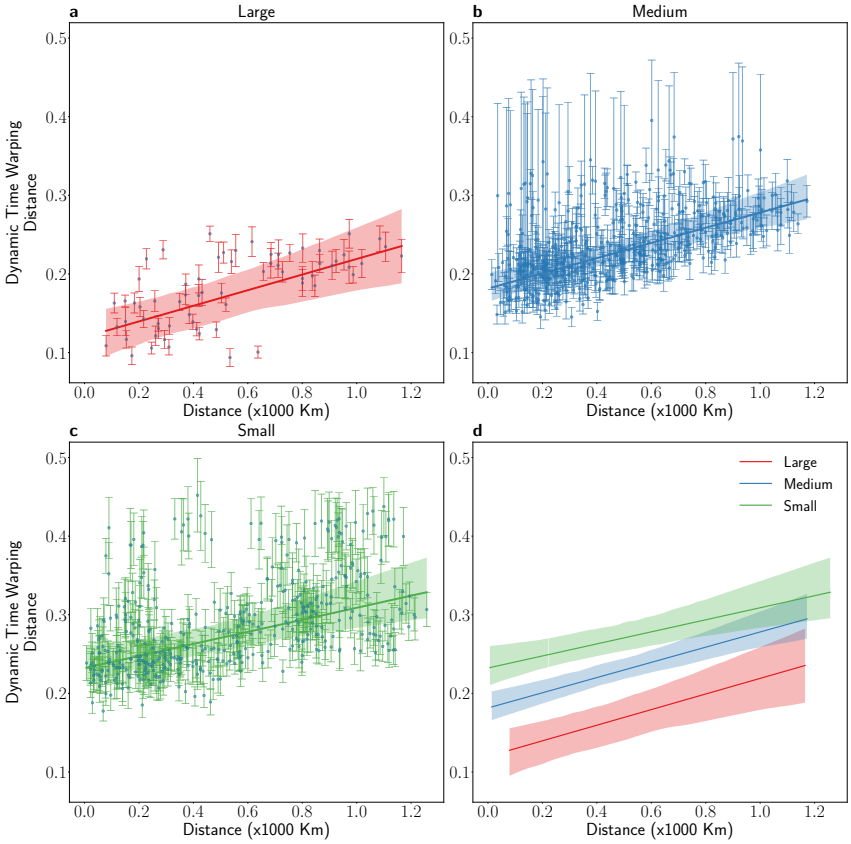

Figure 2.3: **Bootstrap for weekdays a)** shows the results of the bootstrap for couples of Large cities during the weekdays. The points and the bars represent the mean and the variance computed for each couple through the jackknife sampling method. The shaded area represents the 95% confidence interval obtained using the bootstrap method described in *Materials and Methods* and the line represents the bootstrapped Weighted Least-Square Regression (WLS) fit. **b)** as **a)** but for the Medium cities. **c)** as **a)** but for Small cities. **d)** summarizes the estimates of the bootstrap regression previously obtained in Figure **a-c)**. Note that 95% confidence intervals for all the classes mostly do not overlap. Detailed statistics are reported in Table 2.A.1.

report of the computed statistical measures) and the 95% confidence intervals for all classes are mostly not overlapping. Two large cities that are close to each other such as Milan and Turin are more similar than more distant large cities such as Milan and Rome. However, the similarity between large cities is higher than the similarity between medium cities, independently of the distance, i.e. Milan and Rome (two distant large cities) are more similar than Padua and Modena (two medium cities that are closer). The same relationship holds for medium and small cities: two medium cities are more similar to each other than

two small cities, and the similarity decreases when spatial distance between cities increases.

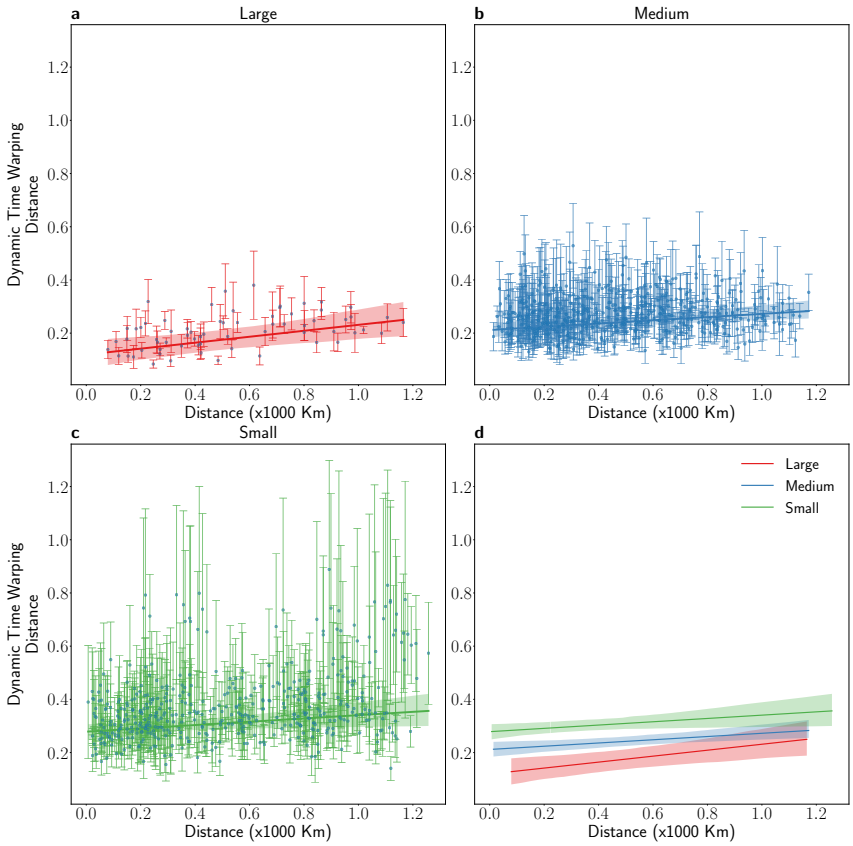

Figure 2.4: **Bootstrap for weekends a)** shows the results of the bootstrap for couples of Large cities during the weekends. The points and the bars represent the mean and the standard deviation obtained for each couple using the jackknife sampling method. The shaded area represents the 95% confidence interval obtained using the bootstrap method described in *Materials and Methods* and the line represents the bootstrapped Weighted Least-Square Regression (WLS) fit. **b)** as **a)** but for the Medium cities. **c)** as **a)** but for Small cities. **d)** summarizes the estimates of the bootstrap regression previously obtained in Figure **a-c)**. Note that error bars and confidence intervals are larger, due to the lower data availability for weekends (our dataset consists of 6 weekend days and 18 weekdays).

In Figure 2.4, during weekends, we can observe a similar pattern as for weekdays (see Table 2.A.2 for a complete report of the computed statistical measures). In particular, the ordering of the cities' classes is consistent with weekdays and the drift due to distance in the similarity is still visible, although

confidence intervals are not as clearly separated as the weekdays' ones reported in Figure 2.3.

Furthermore, Table 2.A.3 reports the Spearman's correlation scores between the variance-weighted average of the DTW distances and several socio-economic indicators characterizing (i) the size of a city (i.e., resident population and population density), (ii) the city's attractiveness for foreigners and immigrants (i.e., foreign-born population and percentage of foreigners per resident population), (iii) the city's wealth (i.e., yearly average income per tax payer), and (iv) the outgoing and incoming commuting patterns.

Our results show that the variance-weighted average of the DTW distances is negatively associated with both the absolute number of resident population (Spearman's ρ = -0.614***) and the population density (Spearman's ρ = -0.477***) as well as with the absolute number of foreign-born population (Spearman's ρ = -0.653***) and the foreign percentage (Spearman's ρ = -0.409***). Thus, cities which are more synchronized seem to be both more populated and more attractive for foreigners (e.g., tourists, immigrants). Interestingly, the variance-weighted average of the DTW distances also correlates negatively with yearly average income per tax payer (Spearman's ρ = -0.349**), showing a relationship between highly synchronized cities and the rich ones. We have also investigated the effect of *activity timeseries* synchronization on the commuting (incoming and outgoing) patterns of each city. As shown again in Table 2.A.3, we have not found significant correlations, with the exception of a slightly significant negative association (Spearman's ρ = -0.236*) between the variance-weighted average of the DTW distances and the outgoing commuters' ratio (computed as the ratio between the number of outgoing commuters and the number of incoming+outgoing commuters).

We have also tested the correlation between the mean DTW distance and the spatial distance for each couple of cities. Our results (Spearman's ρ = 0.205***) show that the increasing spatial distance is associated with a lower communication synchronization. Hence, it seems that regional economies are playing a role in the communication synchronization of our cities.

Finally, in Table 2.A.4 we report the 20 cities with the lowest variance-weighted average of the DTW distances, namely the cities with higher level of activity timeseries synchronization. Interestingly, Rome appears as the most synchronized city in Italy (mean DTW distance = 0.168) and this may be ex-

plained by a mixture of several factors such as its political and economic role (i.e., Rome is the national capital and the second wealthiest city in Italy), its size (i.e., Rome is the most populous city in Italy), and its attractiveness for foreigners (e.g., tourists). Other cities showing high degree of activity time-series' synchronization are important cities for the maritime trade routes (i.e., Trieste and Genoa are the major Italian seaports for trade of goods and flows of people). Again, relevant touristic cities, such as Florence, Rimini, Ravenna, Venice, Verona, Milan are among the ones with higher levels of activity time-series' synchronization. Finally, it is worth noting that only one city (Palermo) located in the South of Italy appears in the list of the 20 most synchronized ones.

2.5 DISCUSSION

Cities' synchronization and similarity using mobile phone and social media (i.e., Twitter) data has been recently investigated in [44] and in [22]. In Grauwin *et al.* [44], three global cities (New York, London and Hong Kong) are studied by means of mobile phone usage patterns. The paper shows that these three large cities, despite the distance, have comparable and common usage patterns, especially in the core business districts of the cities. In Morales *et al.* [22], an analysis of the synchronization of large world cities is presented using Twitter data. In this work, a cluster of similar large cities (Middle Eastern, European and African cities) is detected.

Our results, based on CDR data for 76 Italian cities, provide some evidence in support of these findings, at least for a subset of European cities (Italian large cities), by considering only the DTW distance. Indeed, we can observe, based on the outgoing calls' similarity patterns, an emerging cluster of similar Italian large cities (see Figure 2.2). We further investigated the city's similarity and synchronization concept, by considering different scales of cities (Large, Medium and Small) and exploring how distance influences similarity between cities. After removing the effects of circadian rhythms by using DTW as time-series distance measurement, we have analyzed the effect of spatial distance over cities' similarity. Tobler's First Law of Geography[61] states that *everything is related to everything else, but near things are more related than distant things.*". This phenomena can be observed in particular in the distance decay (i.e. the

effect of distance on spatial or cultural interactions), that plays a central role in distribution of population, time and spatial interactions as shown in [62, 63, 64]. It is therefore reasonable to expect that the similarity between cities decrease when cities' distance increase (such as in the gravity model [65]). In our work we show that when we consider cities' synchronization, distance is not the only factor that drives cities' distance decay, but also city's size and economic centrality play an important role in the communication pattern similarity. We can suppose that Italian large cities act as *gateway cities* [26]: a gateway city is a city that plays the role of hub and central node for resources and capital circulation for the whole region where the city is located. *Gateway cities* host tertiary services such as banks, trading centers, headquarters of large companies that require a high degree of synchronization [66]. In our paper, we observed this behavior for Italian large cities, that are more similar in rhythms to each other, despite the distance, than medium cities to medium cities or small cities to small cities. This is confirmed also by Grauwin et al. [67], that describes North Italy as a complex interconnected area (*city-region*), where larger cities provide advanced services–such as financial trading centers–for the whole area and act as a gateway for information and commercial flows for the whole region. Interestingly, in our study the 20 cities showing the highest degree of activity timeseries' synchronization are all located in the North (15 out of 20) or in the Center (4 out of 20) of Italy, with the exception of Palermo. Again, our results indicate that highly synchronized cities play a relevant economic role (these cities have higher levels of average yearly income per tax payer) and are more attractive for foreign-born people (i.e. immigrants, tourists, etc.).

As shown in [66], cities follow a common development trajectory and after the population reaches a certain threshold (1.2 million people for the US case analyzed in [66]), the economical development path moves from primary industries (e.g., agriculture, mining, etc.) to tertiary industries such as banks and services. These kinds of industries require a higher level of synchronization, as in the case of brokers trading stocks, or large industries that have headquarters spread all over Italy and the world, thus reducing the influence of distance over cities' similarity.

Conversely, when considering small cities, distance has a larger role in determining the communication synchronization and the similarity, since economy is more bound to local productions and smaller industries. In the case of

medium cities, as shown in [66], the cities are ongoing an industry transformation from primary to tertiary, hence the contribution of the distance over similarity is less than for the case of small cities but still bigger than the one for large cities.

These findings support also the theory exposed in [68], where, by using a scaling model and analyzing social and economical factors, such as GDP, wages, number of crimes, it is shown that large cities have a temporal self-similarity, in terms of higher and faster patterns of social interaction, walking speed of pedestrians, number of employees in research and development. It is also shown that smaller cities, when growing, follow a common social dynamic as the larger ones: when a city increases in population, it tends to accelerate its rhythms and have faster behaviors and technical innovation rates. This is confirmed by our observations, showing a scaling and division of cities based on the city's size: large cities are more similar in rhythms to large cities, despite the distance, than medium cities are to medium cities, and medium cities are more similar to medium cities than small cities are to small cities.

APPENDIX

2.A TABLES

Table 2.A.1: **Bootstrap results for weekdays.** The table reports the mean values of intercept *q* and slope *m* for 500 bootstrap iterations. The bootstrapped 95% Confidence Intervals (CI) and the p-value *p* for the t-test on the slope *m* ($p \leq 0.05$: * , $p \leq 0.01$: ** , $p \leq 0.001$: ***) are also reported.

size	q	m	p
Large	0.120 CI [0.084 - 0.151]	0.099 CI [0.043 - 0.153]	**
Medium	0.181 CI [0.164 - 0.201]	0.097 CI [0.068 - 0.131]	***
Small	0.232 CI [0.210 - 0.260]	0.077 CI [0.041 - 0.121]	***

Table 2.A.2: **Bootstrap results for weekends.** The table reports the mean values of intercept *q* and slope *m* for 500 bootstrap iterations. The bootstrapped 95% Confidence Intervals (CI) and the p-value *p* for the t-test on the slope *m* ($p \leq 0.05$: * , $p \leq 0.01$: ** , $p \leq 0.001$: ***) are also reported.

size	q	m	p
Large	0.119 CI [0.067 - 0.174]	0.112 CI [0.046 - 0.192]	*
Medium	0.212 CI [0.183 - 0.240]	0.061 CI [0.023 - 0.114]	**
Small	0.279 CI [0.249 - 0.307]	0.062 CI [0.011 - 0.131]	*

Table 2.A.3: Spearman's correlation between the variance-weighted average of the DTW distances and (i) resident population (absolute number), (ii) foreign population (absolute number), (iii) population density, (iv) percentage of foreigners per resident population, (v) yearly average income per tax payer, (vi) ratio between commuters coming to a city and commuters leaving a city, (vii) percentage of incoming commuters per resident population, and (viii) percentage of outgoing commuters per resident population ($p \leq 0.05$: * , $p \leq 0.01$: ** , $p \leq 0.001$: ***).

	DTW variance-weighted avg
Resident population	-0.614 ***
Foreign population	-0.653 ***
Population density	-0.477 ***
Foreign percentage	-0.409 ***
Average income	-0.349 **
In-out commuters ratio	-0.038
Incoming commuter ratio	-0.087
Outgoing commuter ratio	-0.236 *

Table 2.A.4: List of the 20 cities with the lowest variance-weighted average of the DTW distances. Cities with lower variance-weighted average DTW distances are more synchronized.

Ranking	City	DTW variance-weighted avg.
1	Rome	0.168
2	Genoa	0.183
3	Florence	0.186
4	Brescia	0.190
5	Turin	0.193
6	Trieste	0.197
7	Rimini	0.200
8	Ravenna	0.208
9	Bologna	0.210
10	Udine	0.211
11	Milan	0.217
12	Como	0.218
13	Venice	0.218
14	Verona	0.224
15	Terni	0.225
16	Bolzano	0.227
17	Modena	0.230
18	Bergamo	0.231
19	Ancona	0.231
20	Palermo	0.231

3

SPATIAL CLUSTERING AND NETWORK TOPOLOGY: A STUDY ON DYNAMICAL NETWORKS

3.1 BACKGROUND

Network theory has proven to be a valuable framework for capturing several structural features of geo-spatial systems[69, 70]. Notably, dynamical networks are becoming increasingly important in understanding and describing real world behavioral and social phenomena [71, 72]. As a matter of fact, the explosive adoption of dynamical network as a modeling tool for human behavioral studies has grown along with the increasing data availability from personal technological devices[73]. More and more data are used for scientific purposes by researchers collaborating with companies. Some of these are made available for research in aggregated forms and other are released to the public to benefit communities as a whole[74, 75, 76, 77, 78, 79]. For privacy reasons, data aggregation and clustering has often been suggested as a possible solution to some aspects of these problems and this approach has already been widely adopted[80].

With this work we aim at bridging the theory of spatio-temporal networks and clustering by extensively analyzing the impact of an aggregation process on the topology of the resulting network. In general, given a two-dimensional space, S, we can represent motion of objects in terms of continuous trajectories. In the real world, continuous trajectories are collected as discrete sequences of locations, e.g. GPS sampling. Potentially, each of these locations might be represented as a node while a connection can be activated from one node to another following the time ordered sequence of visited locations. A network approach to dynamical processes on graphs also requires a discretization of the temporal component: a slice of the network is obtained by sampling times within a certain time interval and considering them as contemporaneous. Each slice contains a copy of all the locations considered and all the edges that were active in the corresponding time interval. As a consequence, how we design

temporal sampling affects the network structure and the diffusion processes on it. This aspect has been extensively studied in previous works[37, 38, 39]. Here we capitalize over those results and explore more deeply the spatial aspect of spatio-temporal networks.

In this setting a clustering algorithm has the role of aggregating nodes based on some heuristics, usually the metric defined over S. In this work, we adopt a similar approach. Based on the Haversine approximation of distance on a polar coordinate reference system, we cluster together nodes that lie within a maximum radius of length ϵ. By varying the ϵ parameter we measure how the topological features of the network and its dynamical properties change. We also investigate how the small world behavior of spatio-temporal networks [81], and how it changes depending on the clustering radius.

3.2 CONTRIBUTION

In this work we propose an initial study on how clustering and discretization of continuous mobility datasets affects spatio-temporal network measures. In particular, we show how a coarse-graining process based on geographical distance redesigns the network structure by changing both the nodes and the edges among them. Additionally, we compare these measures with two synthetic models–random waypoint ad s-EPR–as well as presenting an algorithm for computing temporal betweenness centrality. This study is intended as a guide to mobility modeling and clustering calibration in a dynamical network environment.

3.3 METHODOLOGY

3.3.1 *Individual Trajectory Data and Sampling*

For our experiments we picked three real-life open data sources: T-Drive and Brightkite and Gowalla.

T-Drive [75] is a dataset featuring taxi drivers in Beijing. It contains a high number of users (>10k) with data recorded over a very short period of time (from 2008-02-02 to 2008-02-08). 75.36% of the traces have a time resolution

of 1 minute [75]. To clean the dataset and reduce noise we picked only the GPS points within Beijing where $39° \leq$ latitude $\leq 41°$ and $115° \leq$ longitude $\leq 118°$. We discarded Saturday, 2008-02-02 and Sunday, 2008-02-03 since the mobility and activity behavior might differ at the weekends [82]. We select a time window $W_{taxi} = 1$ day, hence a time frame will represent $\frac{24 \text{ hours}}{100} = 14.4$ minutes. We then sampled 100 datasets, each one containing the mobility of 100 taxis.

Brightkite [74] and Gowalla [83] were World-wide location-based social networks where users could register their check-ins. The datasets consists of 4.8 millions geolocated check-ins, starting from March 2008 to October 2010 for Brightkite and 6.4 millions check-ins starting from February 2009 to October 2010 for Gowalla. Due to the nature of these two datasets, each user has few points, sparse both in time and space. Therefore, we considered a lager time window of 6 consecutive months fo both datasets ($W_{Brightkite} = 6$months), where we have peaks in the number of check-ins registered. We focused on America and Europe, that contain most of the points of the dataset ($23° \leq$ latitude $\leq 70°$ and $-130° \leq$ longitude $\leq 45°$), and sampled 100 datasets, each containing the traces of 1500 users.

For all the datasets, we consider the first time frame to begin at midnight of the first day selected for the sample and the last time frame to end at midnight of the last day of the sample. Furthermore, we discard the traces with less than two distinct GPS points.

Each dataset is then divided in 100 equally-sized time frames, hence, for each data source, a time frame of the generated network represents a different span of time. For the both Brightkite and Gowalla, we selected 180 consecutive days, hence each time frame has a duration of $\frac{180 \text{ days}}{100} = 43.2$ hours. For the taxi dataset each time frame spans $\frac{24 \text{ hours}}{100} = 14.4$ minutes

3.3.2 *Spatio-Temporal Networks*

A temporal (or dynamical) network is a time-ordered sequence of graphs. Each graph corresponds to a static snapshot of the network in a given time frame. Typically, each snapshot consists of N nodes where N is equal to the number of different objects under study. These might be e-mail senders and recipients, calls among mobile phone devices, users in social media platforms, as well as genes, functional brain regions and species of an ecological system [84].

Similarly, *spatio-temporal networks* are time-ordered sequences in which each element of the sequence is represented by a location. In this context, edges represent movements or connections among locations at specific times. They can encode mobility information or other structural relationships among the two. In our study, edges represent individual mobility between two locations. They are drawn from one node to another on a given slice, which corresponds to a specific time frame, if an individual traveled among them at that time. These individual-based connections are naturally directional, from the origin (the earliest visited location among the two) to the destination (the later visited among the two). As a consequence, in our approach we consider only directed graphs in which edges can only be crossed along their direction.

More formally, following the notation introduced in [85], we define V to be the set of all nodes belonging to a spatio-temporal network. Let $N = |V|$ be the number of nodes and T be the number of time frames we divided our network into. Since spatio-temporal networks are ordered sequences of directed graphs slices, we can define it as $G = \{G(t_1), \cdots, G(t_T)\}$, where $G(t) = (V, E(t))$ is a graph slice corresponding to a particular time window t and $E(t)$ the edges active at that time. We assume that, given a time frame t, all edges in $G(t)$ can be explored when searching for the shortest paths by following the edge direction. In our work we consider the edges to have all the same weight–i.e. we are considering unweighted graphs– and the formalism can easily be extended to weighted graphs [85]. We additionally make the simplifying assumption that the velocity of traveling across one edge is negligible with respect to the temporal size of t, i.e. $t_e \ll t$ where t_e is the time required to go from one node to the following one individual's sequence of locations.

For each time frame t we define the corresponding adjacency matrix $A(t)_{ij}$, where $A(t)_{ij} = 1$ if there exists an active edge in $G(t)$ from node i to node j, $A(t)_{ij} = 0$ otherwise.

3.3.3 *Spatio-Temporal Metrics*

Here we introduce the temporal measures we studied with a brief discussion on the algorithms implemented, while details of the clustering process are presented in section 3.3.4. As it is true for static networks, the properties and topological structure of a temporal network can hardly be captured by a single

metric. Thus, in order to have a deeper understanding of how a clustering process affects the structure and relationship among nodes, in this section we introduce the extension to temporal networks of several established metrics in network theory.

Except for average edge density estimation, all the other measures for temporal networks involve the concept of temporal shortest path.

Definition 3.3.1. *Temporal path*
Given a spatio-temporal network $\mathcal{G}$, a temporal path from node i to node j in the time interval $[1, T]$ is a sequence of edges

$$p_{ij} = (a_0 = i, a_1, t_1 = 1), (a_1, a_2, t_2), \cdots, (a_{k-1}, a_k = j, t_k) \tag{1}$$

such that $\forall k \in [1, T]$ it holds that $t_{k-1} \leq t_k$ [86].

We stress that, here, we consider all temporal paths as starting at the same initial time t_1.

Definition 3.3.2. *Temporal length*
We define the length of a temporal path p_{ij} as $l(p) = k$, where k is the number of time frames needed to reach j starting from i at time frame $t_1 = 1$.

Definition 3.3.3. *Temporal shortest path*
Temporal shortest paths, $\sigma_{temp}(i, j)$, are defined as temporal paths whose temporal length, $l(p_{ij})$ is minimal:

$$\sigma_{temp}(i, j) = p'_{ij} \in \{p_{ij}\} : l(p'_{ij}) = \min_{\{p_{ij}\}}(l(p_{ij})) \tag{2}$$

If two temporal shortest paths have the same temporal length, the path that requires less hops across different nodes is defined to be the shortest one (as defined in [86, 87]).

Temporal shortest paths defined above are often also called *earliest arrival path* [88], or *temporal foremost path* in [87].

Definition 3.3.4. *Temporal distance*
We define the temporal distance d_{ij} as the temporal length of the temporal shortest path $\sigma_{temp}(i, j)$.

$$d_{ij} = l(\sigma_{temp}(i, j)) \tag{3}$$

If a temporal path between the two nodes does not exists, the temporal distance is defined as $d_{ij} = \infty$.

Conveniently, we also introduce the concept of *temporal connected component* with the following definition:

Definition 3.3.5. *Temporal connected component*
We define $\mathcal{C}$ as the *temporal connected component*, i.e. the set of origin-destination pairs (i, j) of a temporal network for which $d_{ij} < \infty$.

Inspired by the work of [69, 89], we propose an algorithm that allows to compute together both shortest temporal path and temporal betweenness centrality, defined later in this section. We provide the pseudo-code of the algorithms in Appendix 2.5.

Using these definition we can easily introduce several measures that capture different aspects of a spatio-temporal network.

3.3.3.1 *Temporal Characteristic Path Length*

The *temporal characteristic path length* is proposed as a measure that collectively describes the average temporal path length of all the paths among nodes of a temporal connected component $\mathcal{C}$. It is defined as follows:

$$L = \frac{1}{N(N-1)} \sum_{i,j \in \mathcal{C}} d_{ij} \tag{4}$$

where the sum runs over all the pairs of nodes (the set of origin-destination nodes) $(i, j) \in \mathcal{C}$. Here, d_{ij} is the temporal distance between i and j, as defined in Definition 3.3.4. We can imagine temporal characteristic path length as a measure of how quick the information can flow over a temporal network: the larger the temporal characteristic path length, the larger the number of timestamps it will take, on average, for the information to reach a node of the temporal connected component $\mathcal{C}$.

3.3.3.2 *Temporal Diameter*

A different measure that captures the temporal size of the network is the *temporal diameter*. It is defined as the length of the longest temporal shortest path and can be easily computed from the distance matrix d_{ij}.

$$D = \max_{i,j \in \mathcal{C}} [d_{ij}]$$

We can imagine the temporal diameter as the shortest time it takes (as number of time-frames) starting from the initial time t_1 to reach every node of the connected component $\mathcal{C}$.

All these measures are readily obtained once the temporal shortest paths of a temporal network are computed. Indeed, each of these definitions capture different aspects of the same objects, i.e. the temporal shortest paths. However, to better understand the relationships among nodes and their relative importance in bridging the dynamics that might occur on the network, a more complex measure must be studied.

3.3.3.3 *Temporal Betweenness Centrality*

To this extent, we choose to focus our attention also on centrality measures, in particular, betweenness centrality, that represents the importance of a node in supporting connectedness within the network. Keeping in mind the definitions introduced above and following the ideas of previous works [90, 34, 91, 92], let us introduce $\sigma_{temp}(i, j)$ as the number of shortest temporal paths between node i and j in a temporal network G. Let $\sigma_{temp}(i, j, k)$ be the number of temporal shortest paths from i to j going through k. The temporal betweenness centrality for a node k can hence be defined as:

$$C_B(k) = \frac{1}{(N-1)\cdot(N-2)} \sum_{\substack{m\neq n\neq k\in V:\\ \sigma_{temp}(m,n)>0}} \frac{\sigma_{temp}(m,n,k)}{\sigma_{temp}(m,n)}$$

As displayed in the formula, the temporal component adds a layer of complexity to the already high computational costs of static betweenness centrality. Inspired by the Brandes solution for static betweenness centrality[89], we implement a temporal version of the algorithm. The algorithm's pseudo-code is reported in Appendix 2.5.

3.3.3.4 *Topological Overlap and Temporal Network Overlap*

Real world networks, such as networks representing contacts between people, often have links that repeat over time [93]. The topological overlap is a measure intended to quantify to what extent links in one time frame t of the network are persistent in the following time frame $t+1$. More formally, if we consider a node i and two adjacent time frames t and $t+1$, the *topological overlap* for node i is defined as [34, 81]:

$$C_i(t_i, t_{i+1}) = \frac{\sum_j \left[a_{ij}(t)a_{ij}(t+1)\right]}{\sqrt{\left[\sum_j a_{ij}(t)\right]\left[\sum_j a_{ij}(t+1)\right]}}$$

where $a_{ij}(t)$ is the ij coefficient of the adjacency matrix of frame t, being 1 if an edge from i to j is active at t, and 0 otherwise. Being dependent on time t and characterizing node i (and the edges originating from it), this measure does not bring a global description of the whole network itself. However, by simply averaging over all the time frames of the temporal network, we can introduce the *average topological overlap* $\bar{C}_i$ of node i. Let then

$$\bar{C}_i = \frac{1}{T-1} \sum_{m=1}^{T-1} C_i(t_m, t_m + 1)$$

be the *average topological overlap* for node i. Again, by averaging over all the nodes of the network, we can introduce the *temporal-correlation coefficient* for a temporal network as

$$C = \frac{1}{N} \sum_{i=0}^{N} \bar{C}_i$$

We stress that through the averaging processes we obtain a quantity that is 1 only if all the snapshots of the temporal network have the same edges. Conversely, it is equal to zero if no active edge at any time frame appears in the subsequent snapshot.

3.3.3.5 *Average Edge Density*

Let $|E(t)|$ be the number of edges at time frame t in $G(t) = (V, E(t))$, excluding self-loops. The *average edge density* D is defined as follow:

$$D = \frac{1}{T} \sum_{t=1}^{T} \frac{|E(t)|}{N \cdot (N-1)}$$

The average edge density is 0 for a spatio-temporal network without edges and 1 for a complete spatio-temporal network.

The measures presented here were all computed as descriptors of the network temporal structure. Aiming at presenting a first deep assessment on how clustering affects the spatio-temporal networks' structure, we will dedicate the following sections to introducing both our clustering approach and the synthetic and real data spatio-temporal networks before presenting the results of our analysis.

3.3.4 *Clustering Trajectories and Aggregated Networks*

To asses how spatial clustering influences the spatio-temporal metrics described in section 3.3.3, for each trajectory we cluster the latitude and longitude coordinates using a vanilla DBSCAN [94], which is a well known standard clustering method for clustering geographical data [95]. Differently from [96], where the behaviour of a temporal network is explored at different temporal resolutions, we consider the duration t of the time frames to be fixed for each data source, and we focus only on the spatial clustering aspect. Each cluster corresponds to a node in our temporal network G, and for each movement happening during a time frame t in a trajectory, we add a corresponding edge to the network $G(t)$.

In order to investigate the behavior of the network when the ε parameter of DBSCAN varies, we fix to 1 DBSCAN's min_pts parameter. This parameter represents the minimum number of points for each cluster. Then we average $max\_\varepsilon_{S_i}$ and $min\_\varepsilon_{S_i}$ to obtain max_ε_S and min_ε_S that will be used to perform the experiments for the data source S. Since we observed that the distribution of the number of clusters decreases in a superlinear way when ε increases, we pick 20 logarithmically spaced values $\hat{\varepsilon_S}$ from the interval $[max\_\varepsilon_S, min\_\varepsilon_S]$.

For each $\hat{\varepsilon_S}$ so obtained, we cluster every sample of the data source S, generate the network as described in section 3.3.4 and compute the metrics listed in section 3.3.2.

3.3.5 *Synthetic Models of Mobility*

3.3.5.1 *Random waypoint*

In order to generate a synthetic model that mimics our real-world datasets, we use a random waypoint with Gaussian attractors, similarly to [97, 98] and adopt importance sampling to decide the next point of a trace [99]. For each simulation, we position one Gaussian bi-modal attractor for the T-DRIVE dataset in a location corresponding to the center of Beijing and two Gaussian bimodal attractors for the Brightkite dataset, corresponding to America and Europe. The mean and standard deviation of the attractors are chosen by computing the mean and standard deviation of latitude and longitude. At each iteration, we

sample a tuple $(\delta\_lat, \delta\_lon, v)$, representing respectively the shift in latitude, in longitude and the speed from the corresponding dataset.

Each trajectory is built as follows:

1. Initially, we pick an origin point by sampling the Gaussians.
2. For each iteration we pick one point $h = (lat_i, lon_i)$ by sampling uniformly on a circumference of radius $r = \sqrt{\delta\_lat^2 + \delta\_lon^2}$ and center in the previous point of the trajectory and evaluate the probability density function of the attractors in h.
3. We multiply the pdfs, obtaining $p_{ki} = pdf\_lat_{ki} \cdot pdf\_lon_{ki}$. In the case of random waypoint for Brightkite, the values of p_{ki} for the attractors in America and Europe are summed by weighting on the percentage of users in the corresponding countries in the Brightkite dataset.
4. We compute the maximum value of the pdf $m = max(pdf_{lat_k}, pdf_{lon_k})$ and pick a value u form a uniform distribution in the interval $[0, m]$
5. Until $p_{ki} < u$, repeat the process from point 2
6. The shift $(\delta\_lat_i, \delta\_lon_i)$ is added to the latitude and longitude of the previous point of the trace
7. Finally the timestamp for a mobility trace is obtained by dividing the spatial distance between the origin point and the chosen destination point by the sampled velocity v and summing it to the timestamp of the previous point of the trace.

The process terminates when the timestamp reaches a fixed value: 24 hours for the T-DRIVE simulations and 6 months for the Brightkite simulations. For both T-Drive and Brightkite, we generate 100 synthetic datasets, each consisting of respectively 100 traces for T-Drive and 200 traces for the Brightkite. We finally perform a spacial clustering as described in section 3.3.4, create the corresponding spatio-temporal network and compute the metrics described in section 3.3.3.

3.3.5.2 *Spatial Exploration and Preferential Return*

Another synthetic model we consider is the Exploration and Preferential Return (EPR), and in particular the spatial version of EPR (s-EPR) [100]. EPR is a

synthetic model where an user at each timestamp can either return to previously visited locations (preferential return) or explore a new location at a given distance from the current location (exploration), according to a truncated power law distribution of standard mobility measures such as the waiting time and the jump length [100, 101].

The s-EPR model is an evolution of the EPR model, where the movement happens within a fixed bounding box and for each individual a gravity model is used to assess the probability of returning to a given location. The exploring phase is achieved by sampling points within the bounding box where the jump length and waiting time between movements follow a truncated power law.

To simulate the inaccuracy of GPS data, we add Gaussian noise to each point with σ equal to the 5% of each dimension's range. We consider the movements to happen on a sphere and create a dataset, simulating movements of the scale of Brightkite and Gowalla.

We decided to use s-EPR only for datasets where the individual mobility patterns are well describe by a preferential return behavior, rather that using s-EPR for collective behavior such as the one described by T-Drive taxi dataset, that might need more specific models to properly describe the taxis' trajectory[102, 103].

3.4 RESULTS

In this section we show the results obtained for the T-Drive, Brightkite and Gowalla datasets presented in section 3.3.1, for the random waypoint model (section 3.3.5.1) and for the s-EPR model (section 3.3.5.2). After generating the spatio-temporal networks by clustering with DBSCAN, as described in section 3.3.4, for each data source we compute the metrics described in section 3.3.3 over 20 log-spaced ε values. To be able to plot and compare different scales of ε, for each dataset we rank the values of ε. The rank value takes the name of *cluster size* in the following figures. As an example, all points for a data source S in the plot having *cluster size* = *0* correspond to networks generated with $\varepsilon = min\_\varepsilon_S$, as described in section 3.3.5.1, and the points with *cluster size* = *1* correspond to the measures resulting from clustering with $\varepsilon = max\_\varepsilon_S$. We stress again that the larger the value of the local cluster radius ε (and consequently the *cluster size*), the fewer clusters we will have. We used the same values of min_ε and max_ε

for both the synthetic and real-life datasets. Finally, for each value we report the mean and Standard Error of the Mean (SEM) over 100 runs for all datasets and synthetic models. The ε values considered range from 4km to 400km for Brightkite and Gowalla and from 60 meters to 5km for the taxi dataset. We will divide the datasets presented into two categories: *low mixing* (Gowalla and Brightkite dataset), where each trace consists of fewer and recurrent locations (mirroring personal mobility patterns) and one with *high mixing*, such as the taxi dataset, where each trace is not the results of the mobility of a single individual and therefore each trace contains more non-repeating locations.

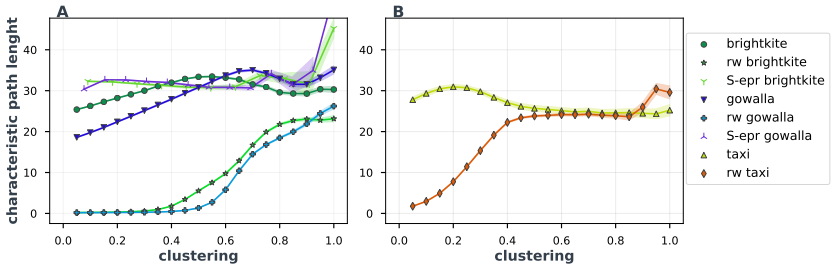

Figure 3.4.1: **Temporal characteristic path length** mean and standard error of the mean for 100 runs for all datasets and synthetic models.

In Figure 3.4.1A,that high mixing mobility (taxi dataset in Figure 3.4.1B), is well described by the random waypoint model. We can notice also that the random waypoint model does not capture correctly the changes in distances for the global datasets. Furthermore, the temporal characteristic path length is mostly stable when the cluster size is between 0.5 and 0.9 for the taxi dataset, corresponding to an ε between 0.6km and 3km.

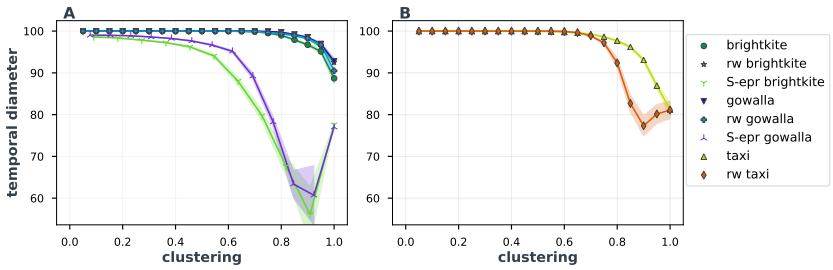

Figure 3.4.2: **Temporal diameter** mean and standard error of the mean for 100 runs for all datasets and synthetic models.

In Figure 3.4.2 we can see that the temporal diameter stays constant during an initial phase and super-exponentially decreases when ε increases. Temporal

diameter is well captured for Brightkite and Goawalla by the random waypoint model while the same is not true for the s-EPR model, that decreases smoothly and gradually when clustering radius increases. The real-life datasets and random-waypoint models maintain some pairs of node that are at maximum distance (100 time frames) throughout most of the range of the cluster size. The same is true for the taxi dataset, where we can find nodes that are at maximum distance until ε reaches 1.3km.

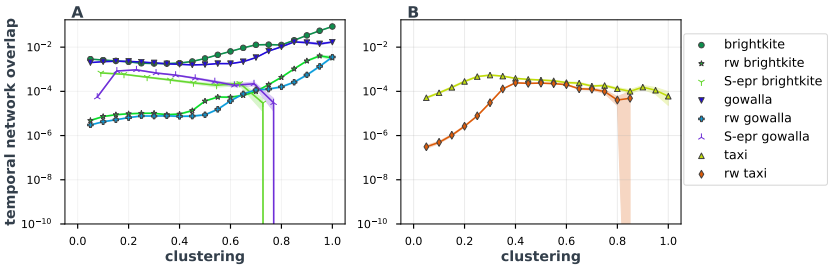

Figure 3.4.3: **Temporal Network Overlap** mean and standard error of the mean for 100 runs for all datasets and synthetic models.

In Figure 3.4.3 we can see that the temporal network overlap is mostly stable for both the high and low mixing datasets, with a slight increasing trend in the low mixing dataset and a slight decreasing trend for the high mixing datasets. We can see that the random waypoint model in Figure 3.4.3A better captures the increasing trend in the correlation coefficient. The s-epr model, although closer as value to the real-life dataset when the ε value is between 7km and 110km, misses the increasing trend and generates networks with no repeated edges in consecutive time frames for $\varepsilon > 110km$.

By looking at averaged measures of the network structure we can conclude that, while for low mixing personal mobility patterns the temporal structure is better preserved at lower values of ε, corresponding to smaller cluster-size aggregation, for high mixing non-personal patterns the structure rapidly changes both at small cluster sizes as well as at larger cluster size, thus suggesting to carefully tune the cluster size depending on the specific characteristics of the system under study and on the specific goals of the analysis. For these high mixing datasets reduced effects on network topology are shown for intermediate values of epsilon (between 350mt and 2km).

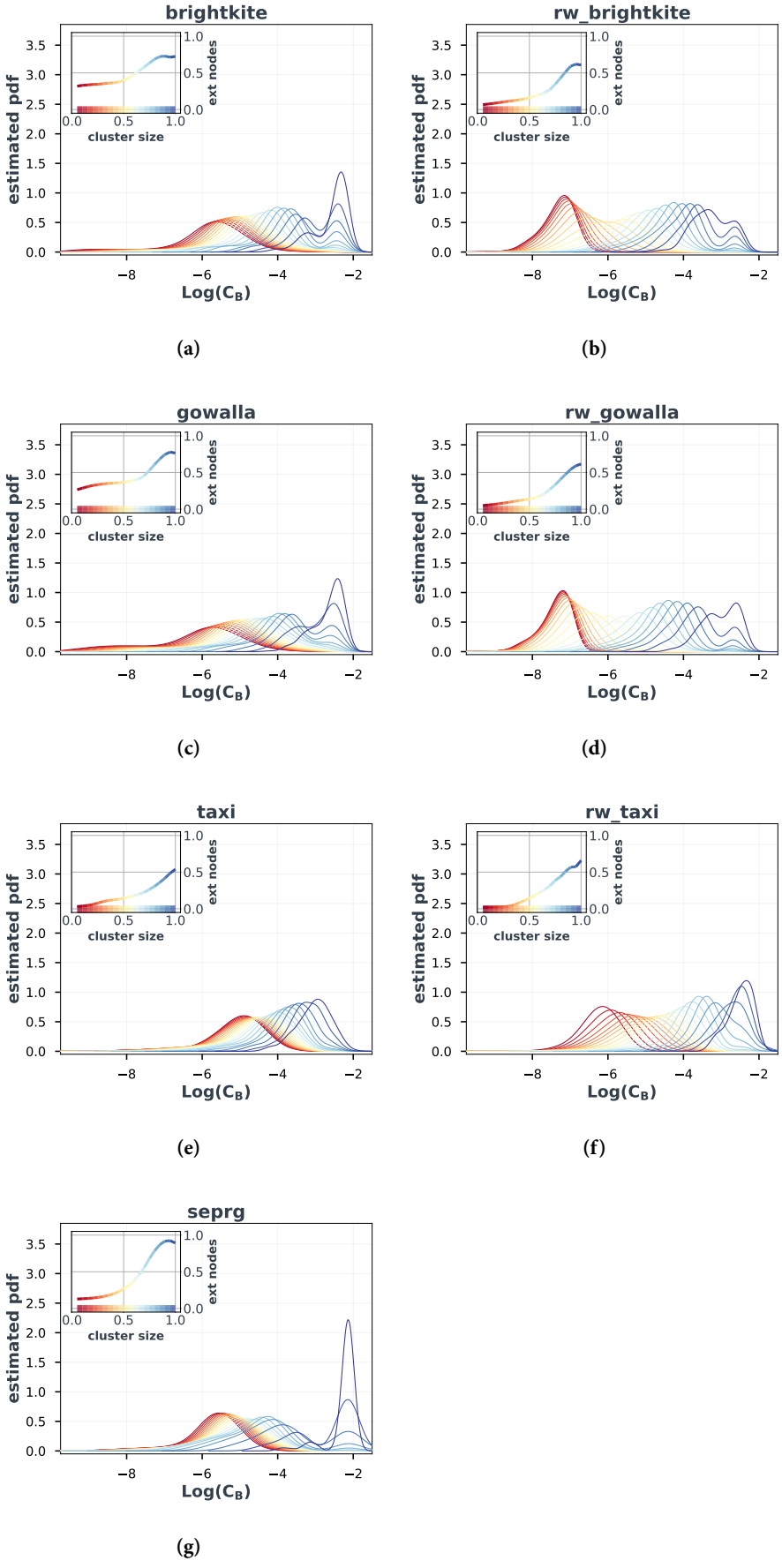

Figure 3.4.4: **Temporal betweenness centrality**. Each line corresponds to the kernel density estimation for non zero values of temporal betweenness centrality for one clustering radius ε. In the inset we can observe the behavior of the fraction of nodes with zero betweenness (*ext nodes*).

In Figure 3.4.4 we can see that the low-mixing datasets' behavior of the temporal betweenness centrality is well captured both as fraction of nodes with zero betweenness and distribution by the s-EPR model (Figures 3.4.4a, 3.4.4c,3.4.4g), while the random waypoint model misses the peaks in the distribution that we can see in 3.4.4a, 3.4.4c. The random waypoint model better mimics the metric for the taxi dataset (Figure 3.4.4e)

3.5 DISCUSSION

In our work, based on the spatio-temporal metrics described in 3.3.3, we identified two types of mobility and behavior in the temporal network representation of real world and synthetic mobility traces: one with *low mixing* (Gowalla and Brightkite dataset)–where each trace consists of fewer preferential locations, as in the case of human mobility[100]–and one with *high mixing* where each trace contains mostly non-repeating locations (random waypoint models and T-Drive dataset). We assume the taxi dataset to be a high mixing dataset: although a taxi might work in a specific area, the locations visited by it are the resulting mobility of different people. Similarly, the random waypoint model can also be considered an high mixing model, since the Gaussian attractors for each user doesn't have memory of previously visited location. For each of the metrics considered, we noticed different behaviors for the synthetic models and the non synthetic traces, that depend both on the metric and on the clustering radius: the behavior of some metrics, such as temporal characteristic path length, is better modeled by the s-EPR model, while others like temporal network overlap is better modeled by the random waypoint model.

This distinction between preferential return traces (lower mixing) and higher mixing traces can be seen also in the temporal diameter: the steep drop in the value of the temporal diameter, corresponds to a super-exponential decrease in the number of clusters, for large enough values of ε. We can observe that the larger the scale of movements within the dataset (Brightkite and Gowalla), the later the temporal diameter decreases. We may suppose that this is due to the nature of low mixing datasets: they are generally more disconnected and characterized by users having fewer points per trace. When considering temporal network overlap, we can observe that for Brightkite and Gowalla datasets, even though the shape of the corresponding random waypoint model

is similar, the scale differs significantly. As previously remarked, this may suggest that random waypoint for Brightkite and Gowalla captures the mobility at level of countries and continents due to the presence of only two Gaussian attractors for the same geographical area; on the contrary, in real-life mobility the movements are less random and more structured at different resolutions: we have big cities that act as attractors (e.g. Los Angeles, New York, London, etc.) and within a city we have micro-mobility towards each area of the city (e.g. the industrial area, the residential zones, etc.): all this shorter-scale mobility might be missed by a model with only two attractors. We can observe that in low mixing datasets the characteristic path length is mostly constant. In particular, for short range mobility, that models a large city such as Beijing, the characteristic path length is constant for a clustering radius between 0.6km and 3km and within that range, the measure is well described by the random waypoint model.

The analyses performed over a wide range of clustering-size values point out that, when using spatial clustering and temporal network, a coordinated harmonization of the clustering scale with the scale of the dynamics under study plays a major role in shaping the structure of the system. While small-scale clustering mildly affect networks describing personal mobility patterns, this does not hold for non-personal mobility patterns (high mixing networks). In this case we revealed that a more structurally stable region is present at intermediate clustering scales (i.e., with a clusters with radii between 0.35km and 2km for the taxi dataset).

Our work highlighted the importance of these considerations, providing a systematic analysis of the structural properties of the temporal networks underlying different dynamical systems: while working with temporal network and spatial clustering is a powerful tool for describing and understanding the behavior of the various components living the system, it should always be kept in mind that the manipulations we are making over the raw data have a strong impact on the description of reality and, in particular, we shown that when working in a temporal network framework clustering scale might affect our conclusions.

APPENDIX

3.A FIGURES

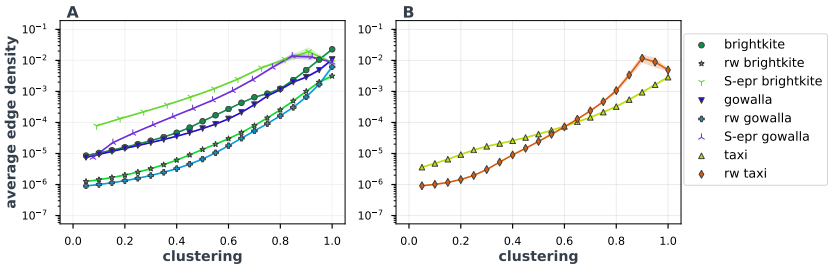

Figure 3.A.1: **Average edge density** mean and standard error of the mean for 100 runs for all datasets and synthetic models. We can observe an exponential increase for all the datasets.

3.B ALGORITHMS

Algorithm 1 All pairs shortest paths

```
1:  function SINGLE_NODE_TEMPORAL_SHORTEST_PATH(s, G)
2:                                          // All initialized lists have size N = nodes ∈ G
3:      min_t[] ← ∞                         // Contains d_sj for all j ∈ G
4:      min_t[s] ← 0
5:      reached ← {s}
6:      Q ← queue
7:      for t ∈ [1 . . . T] do
8:          for u in reached do
9:              Q.push(u)
10:         while Q is not empty do
11:             u ← Q.pop(0)                // Pop u from the bottom of the queue
12:             for w in neigh(u, G(t)) do
13:                 if w not in reached then
14:                     reached.add(w)
15:                     min_t[w] = t
16:                     Q.push(w)
17:                     if length(reached) = N then
18:                                         // Stops if all nodes are reached
19:                         return min_t
20:     return min_t
```

Algorithm 2 Single node temporal betweenness centrality

```
1: function SINGLE_NODE_BETWEENEESS(s, G)
   // All the lists initialized have size N where N is the number of nodes of the network G
2:     N ← number of nodes of the network G
3:     T ← number of time frames of the network G
4:     δ[] ← 0
5:     σ[] ← 0
6:     σ[s] ← 1
7:     min_t[] ← ∞
8:     min_t[s] ← 0                    // Contains the temporal distance of any
                                          node from the starting node s
9:     reached ← {s}
10:    Q ← queue
11:    S ← stack
12:    P[] ← []                        // List of lists of parents of each node
13:    for t ∈ [1...T] do
14:        for u in reached do
15:            Q.push(u)
16:        while Q is not empty do
17:            u ← Q.pop(0)            // Pop u from the bottom of the queue
18:            if u not in S then
19:                S.push(u)
20:            for w in neigh(u, G(t)) do
21:                if t < min_t[w] then
   // First time we reach a node
22:                    min_t[w] ← t
23:                    Q.push(w)
24:                    dist[w] ← dist[u] + 1
25:                    reached.append(w)
26:                if t = min_t[w] and dist[w] = dist[u] + 1 then
27:                    σ[w] ← σ[w] + σ[u]
28:                    P[w].append(u)
29:    return S, P, σ, min_t
```

Algorithm 3 Temporal betweenness centrality and temporal distance

function ACCUMULATE_BETWEENEESS($betweenness$, S, P, σ, s)
  while S is not empty **do**
    $w \leftarrow S.pop()$
    for v in $P[w]$ **do**
      $\delta[v] \leftarrow \delta[v] + \frac{\sigma[v]}{\sigma[w]} \cdot (1 + \delta[w])$
    if $w \neq s$ **then**:
      $betweenness[w] \leftarrow betweenness[w] + \delta[w]$
function BETWEENEESS_CENTRALITY_AND_TEMPORAL_DISTANCE(G)
  $T \leftarrow$ number of time frames of the network G
  $N \leftarrow$ number of nodes of the network G
  $temporal\_distance[] \leftarrow []$
  $betweeneess[] \leftarrow 0$
  for all $s \in V$ **do**
    $S, P, \sigma, min\_t \leftarrow$ SINGLE_NODE_BETWEENEESS(s, G)
    $betweeneess \leftarrow$ ACCUMULATE_BETWEENEESS_AND_TEMPORAL_DISTANCE($betweenness$, S, P, σ, s)
    $temporal\_distance[s] \leftarrow min\_t$
  return $betweeneess \cdot \frac{1}{\cdot N \cdot (N-1)}$, $temporal\_distance$

4

CONCLUSIONS

In this thesis we analyzed different aspects of spatio-temporal data, focusing on the role of spatial distance. We investigated how distance between cities influences cities' synchronization and how the clustering radius, when creating a temporal network description of continuous GPS data, influences temporal networks' metrics.

Firstly, we have investigated Italian cities' communication synchronization and the influence of spatial distance over this communication synchronization by analyzing the CDR data of 76 Italian cities of different sizes. We found that larger cities tend to be more similar to each other than medium cities when considering only similarity between call patterns. We also found that similarity decreases when spatial distance between cities increases. The drift due to distance over similarity is less strong in large cities, that are gateway nodes for the Italian economic system, hence they have an emerging strongly connected and synchronized network, compared to medium and small cities, which are more bound to local industries. We observed that the similarity decreases in a consistent way according to size of the city: large cities are more similar to large cities than medium cities to medium cities, independently of the spatial distance, and the same holds for medium cities and small cities. Finally, our results have shown that cities with higher average synchronization tend to be richer and to attract more people from other places (e.g. tourists, business people, and migrants).

However, our work has several limitations: for example, while we could hypothesize that the high synchronization of some cities (e.g. Trieste, Genoa, Ancona) is due to their importance for maritime trade routes, this result cannot be validated by checking per city correlations between the variance-weighted average of the DTW distances and the amount of traded goods (or similar indicators about commercial trades). Indeed, to the best of our knowledge, there is no available dataset that provides information about commercial trades at city-level granularity. Furthermore, even if our method could be applied for any

city for which the CDR data is available, for our study we did not have access to any CDR data, for the same time period, for other world cities. Thus, we cannot investigate the *gateway* role played by Italian cities worldwide. Another possible limitation is that the telecommunication company TIM, which provided the CDR data, covered only 30.8% of the Italian market, so we cannot have a full picture of all users for all operators. Though, it is a fair assumption that this sample is representative enough of the population of the cities considered.

In our second study, we adopted a temporal network description of spatial data and studied how spatial clustering affects the topological features of location-based dynamical networks. Namely, we showed how a coarse-grain process based on geographical distance redesigns the network structure changing both the nodes and the edges among them. We analyzed the behavior of two synthetic models: random waypoint and s-EPR. Based on the analysis of the temporal network measures, we identified two categories of behavior: high mixing mobility (T-Drive dataset and random waypoint), where a single trace is the result of the movements of multiple individuals and low mixing mobility (Brightkite and s-Epr), where each trace corresponds to a single person, and hence has fewer preferential return locations. We have shown that the metrics' behavior is dependent both on the metric and on the clustering radius and therefore future studies based on spatio-temporal network should take into account this result when dealing with spatially clustered data. Our results shows how a fine tuning of clustering techniques is a key step for preserving systemic representatives, maintaining individual privacy as well as for reducing and speeding up the computational effort. Finally, we proposed a modified version of the Brandes algorithm to efficiently compute temporal betweenness centrality and shortest temporal paths.

Nonetheless, this latter study is focused only on three specific real-life datasets; in the future it will be expanded to include a wider range of data sources. Furthermore, we limited our analysis to subsamples of the whole networks due to computational and time constraints. For future analysis, we plan to expand the study to include larger samples and other metrics (such as motifs). Finally, we considered the duration of each time frame to be constant and simplified the model by studying only unweighted directed temporal networks. Future work could further explore the subject by investigating also weighted networks and

extending the study to both spatial and temporal clustering with algorithms other than DBSCAN.

BIBLIOGRAPHY

[1] Max Weber. *From Max Weber: essays in sociology*. Routledge, 2013.

[2] John W Slocum Jr and H Lee Mathews. "Social class and income as indicators of consumer credit behavior." In: *Journal of Marketing* 34.2 (1970), pp. 69–74.

[3] Raymond Boudon. "Education, opportunity, and social inequality: Changing prospects in western society." In: (1974).

[4] Janet Fulk, Joseph Schmitz, and Charles W Steinfield. "A social influence model of technology use." In: *Organizations and communication technology* 117 (1990), p. 140.

[5] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. "Human mobility, social ties, and link prediction." In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. Acm. 2011, pp. 1100–1108.

[6] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Sandy Pentland. "Daily stress recognition from mobile phone data, weather conditions and individual traits." In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM. 2014, pp. 477–486.

[7] Jacopo Staiano, Bruno Lepri, Nadav Aharony, Fabio Pianesi, Nicu Sebe, and Alex Pentland. "Friends don't lie: inferring personality traits from social network structure." In: *Proceedings of the 2012 ACM conference on ubiquitous computing*. ACM. 2012, pp. 321–330.

[8] Luca Canzian and Mirco Musolesi. "Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis." In: *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM. 2015, pp. 1293–1304.

[9] Olle Järv, Rein Ahas, and Frank Witlox. "Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records." In: *Transportation Research Part C: Emerging Technologies* 38 (2014), pp. 122–135.

[10] Lorenzo Candeago, Giulia Bertagnolli, Paolo Bosetti, Michele Vescovi, Francesco Sacco, and Bruno Lepri. "Cities of a feather flock together: a study on the synchronization of communication between Italian cities." In: *EPJ Data Science* 8.1 (2019), p. 19.

[11] Simone Centellegher, Marco De Nadai, Michele Caraviello, Chiara Leonardi, Michele Vescovi, Yusi Ramadian, Nuria Oliver, Fabio Pianesi, Alex Pentland, Fabrizio Antonelli, et al. "The Mobile Territorial Lab: a multilayered and dynamic view on parents' daily lives." In: *EPJ Data Science* 5.1 (2016), p. 3.

[12] Derek Ruths and Jürgen Pfeffer. "Social media for large studies of behavior." In: *Science* 346.6213 (2014), pp. 1063–1064.

[13] Norjihan Abdul Ghani, Suraya Hamid, Ibrahim Abaker Targio Hashem, and Ejaz Ahmed. "Social media big data analytics: A survey." In: *Computers in Human Behavior* 101 (2019), pp. 417–428.

[14] Ray M Chang, Robert J Kauffman, and YoungOk Kwon. "Understanding the paradigm shift to computational social science in the presence of big data." In: *Decision Support Systems* 63 (2014), pp. 67–80.

[15] Coco Krumme, Alejandro Llorente, Manuel Cebrian, Esteban Moro, et al. "The predictability of consumer visitation patterns." In: *Scientific reports* 3 (2013), p. 1645.

[16] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. "Computational social science." In: *Science* 323.5915 (2009), pp. 721–723.

[17] Eugene J Webb, Donald T Campbell, Richard D Schwartz, and Lee Sechrest. *Unobtrusive measures*. Vol. 2. Sage Publications, 1999.

[18] Steven D Levitt and John A List. "What do laboratory experiments measuring social preferences reveal about the real world?" In: *The journal of economic perspectives* (2007), pp. 153–174.

[19] Robert Rosenthal. "Experimenter effects in behavioral research." In: (1976).

[20] David Levinson. "Network structure and city size." In: *PloS one* 7.1 (2012), e29721.

[21] Serguei Saavedra, Kathleen Hagerty, and Brian Uzzi. "Synchronicity, instant messaging, and performance among financial traders." In: *Proceedings of the National Academy of Sciences* (2011), p. 201018462.

[22] Alfredo J Morales, Vaibhav Vavilala, Rosa M Benito, and Yaneer Bar-Yam. "Global patterns of synchronization in human communications." In: *Journal of The Royal Society Interface* 14.128 (2017), p. 20161048.

[23] Marco Mamei, Francesca Pancotto, Marco De Nadai, Bruno Lepri, Michele Vescovi, Franco Zambonelli, and Alex Pentland. "Is social capital associated with synchronization in human communication? An analysis of Italian call records and measures of civic engagement." In: *EPJ Data Science* 7.1 (2018), p. 25.

[24] Z. Néda, E. Ravasz, Y. Brechet, T. Vicsek, and A.L. Barabási. "Self-organizing processes: The Sound of many Hands Clapping." In: *Nature* 403.6772 (2000), pp. 849–850.

[25] Saskia Sassen. "The global city: Introducing a concept." In: *Brown J. World Aff.* 11 (2004), p. 27.

[26] John Rennie Short, Carrie Breitbach, Steven Buckman, and Jamey Essex. "From world cities to gateway cities: Extending the boundaries of globalization theory." In: *City* 4.3 (2000), pp. 317–340.

[27] Jing Tian, Jorg Hahner, Christian Becker, Illya Stepanov, and Kurt Rothermel. "Graph-based mobility model for mobile ad hoc network simulation." In: *Proceedings 35th Annual Simulation Symposium. SS 2002*. IEEE. 2002, pp. 337–344.

[28] Bei Pan, Yu Zheng, David Wilkie, and Cyrus Shahabi. "Crowd sensing of traffic anomalies based on human mobility and social media." In: *Proceedings of the 21st ACM SIGSPATIAL international conference on advances in geographic information systems*. ACM. 2013, pp. 344–353.

[29] Vassilis Kostakos. "Temporal graphs." In: *Physica A: Statistical Mechanics and its Applications* 388.6 (2009), pp. 1007–1023.

[30] Petter Holme. "Modern temporal network theory: a colloquium." In: *The European Physical Journal B* 88.9 (2015), p. 234.

[31] Sungmin Lee, Luis EC Rocha, Fredrik Liljeros, and Petter Holme. "Exploiting temporal network structures of human interaction to effectively immunize populations." In: *PloS one* 7.5 (2012), e36439.

[32] Betsy George, Sangho Kim, and Shashi Shekhar. "Spatio-temporal network databases and routing algorithms: A summary of results." In: *International Symposium on Spatial and Temporal Databases*. Springer. 2007, pp. 460–477.

[33] Danielle S Bassett, Nicholas F Wymbs, M Puck Rombach, Mason A Porter, Peter J Mucha, and Scott T Grafton. "Task-based core-periphery organization of human brain dynamics." In: *PLoS computational biology* 9.9 (2013), e1003171.

[34] Vincenzo Nicosia, John Tang, Cecilia Mascolo, Mirco Musolesi, Giovanni Russo, and Vito Latora. "Graph metrics for temporal networks." In: *Temporal networks*. Springer, 2013, pp. 15–40.

[35] *2018 reform of EU data protection rules*. European Commission. May 25, 2018. URL: https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf (visited on 06/17/2019).

[36] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. "Unique in the crowd: The privacy bounds of human mobility." In: *Scientific reports* 3 (2013), p. 1376.

[37] Rajmonda Sulo, Tanya Berger-Wolf, and Robert Grossman. "Meaningful selection of temporal resolution for dynamic networks." In: *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*. ACM. 2010, pp. 127–136.

[38] Bruno Ribeiro, Nicola Perra, and Andrea Baronchelli. "Quantifying the effect of temporal resolution on time-varying networks." In: *Scientific Reports* 3.1 (2013), p. 3006.

[39] Luis EC Rocha, Naoki Masuda, and Petter Holme. "Sampling of temporal networks: Methods and biases." In: *Physical Review E* 96.5 (2017), p. 052302.

[40] SH Strogatz. *Sync: the emerging science of spontaneous order*. New York: Theia, 2003.

[41] Elad Schneidman, Michael J Berry II, Ronen Segev, and William Bialek. "Weak pairwise correlations imply strongly correlated network states in a neural population." In: *Nature* 440.7087 (2006), p. 1007.

[42] Stefan Van Dongen, Thierry Backeljau, Erik Matthysen, and Andre A Dhondt. "Synchronization of hatching date with budburst of individual host trees (Quercus robur) in the winter moth (Operophtera brumata) and its fitness consequences." In: *Journal of Animal Ecology* (1997), pp. 113–121.

[43] David JT Sumpter. *Collective animal behavior*. Princeton: Princeton University Press, 2010.

[44] Sebastian Grauwin, Stanislav Sobolevsky, Simon Moritz, István Gódor, and Carlo Ratti. "Towards a comparative science of cities: Using mobile traffic records in New York, London and Hong Kong." In: *Computational approaches for urban environments*. New York: Springer, 2015, pp. 363–387.

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

[46] G. Barlacchi, M. De Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrisi, F. Antonelli, A. Vespignani, A. Pentland, and B. Lepri. "A multi-source dataset of urban life in the city of Milan and the Province of Trentino." In: *Scientific Data* 2 (2015), p. 150055.

[47] Lewis Dijkstra and Hugo Poelman. "Cities in Europe: the new OECD-EC definition." In: *Regional Focus* 1.2012 (2012), pp. 1–13.

[48] Meinard Müller. "Dynamic time warping." In: *Information retrieval for music and motion* (2007), pp. 69–84.

[49] Santosh K Gaikwad, Bharti W Gawali, and Pravin Yannawar. *A review on speech recognition technique*. Vol. 10. 3. International Journal of Computer Applications, 2010, pp. 16–24.

[50] Gianluca Antonini and Jean-Philippe Thiran. "Counting pedestrians in video sequences using trajectory clustering." In: *IEEE Transactions on Circuits and Systems for Video Technology* 16.8 (2006), pp. 1008–1020.

[51] Trista P Chen, Horst Haussecker, Alexander Bovyrin, Roman Belenov, Konstantin Rodyushkin, Alexander Kuranoc, and Victor Eruhimov. "Computer Vision Workload Analysis: Case Study of Video Surveillance Systems." In: *Intel Technology Journal* 9.2 (2005).

[52] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. "A word at a time: computing word relatedness using temporal semantic analysis." In: *Proceedings of the 20th international conference on World wide web*. ACM. 2011, pp. 337–346.

[53] Cory S Myers and Lawrence R Rabiner. "A comparative study of several dynamic time-warping algorithms for connected-word recognition." In: *Bell System Technical Journal* 60.7 (1981), pp. 1389–1409.

[54] Toni M Rath and Raghavan Manmatha. "Word image matching using dynamic time warping." In: *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*. Vol. 2. IEEE. 2003, pp. II–II.

[55] T Aledavood, E Lopez, S.G.B Roberts, F Reed-Tsochas, E Moro, R.I. Dunbar, and J. Saramaki. "Daily rhythms in mobile telephone communication." In: *PloS One* 10 (9 2015), e0138098.

[56] Daniel Monsivais, Asim Ghosh, Kunal Bhattacharya, Robin IM Dunbar, and Kimmo Kaski. "Tracking urban human activity from mobile phone calling patterns." In: *PLoS computational biology* 13.11 (2017), e1005824.

[57] Ilaria Bartolini, Paolo Ciaccia, and Marco Patella. "WARP: Accurate retrieval of shapes using phase and time warping distance." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.1 (2005), pp. 142–147.

[58] T Dylan Mikesell, Alison E Malcolm, Di Yang, and Matthew M Haney. "A comparison of methods to estimate seismic phase delays: Numerical examples for coda wave interferometry." In: *Geophysical Journal International* 202.1 (2015), pp. 347–360.

[59] AF Bissell. "The jacknife." In: *Journal of Applied Statistics* 4.1 (1977), pp. 55–64.

[60] Joachim Hartung, Guido Knapp, and Bimal K Sinha. *Statistical meta-analysis with applications*. Vol. 738. John Wiley & Sons, 2011.

[61] Waldo Tobler. "On the first law of geography: A reply." In: *Annals of the Association of American Geographers* 94.2 (2004), pp. 304–310.

[62] Bob Mckercher and Alan A Lew. "Distance decay and the impact of effective tourism exclusion zones on international travel flows." In: *Journal of Travel Research* 42.2 (2003), pp. 159–165.

[63] J Douglas Eldridge and John Paul Jones III. "Warped space: A geography of distance decay." In: *The Professional Geographer* 43.4 (1991), pp. 500–511.

[64] Michael Iacono, Kevin Krizek, and Ahmed M El-Geneidy. "Access to destinations: How close is close enough? Estimating accurate distance decay functions for multiple modes and different purposes." In: (2008).

[65] Gautier Krings, Francesco Calabrese, Carlo Ratti, and Vincent D Blondel. "Urban gravity: a model for inter-city telecommunication flows." In: *Journal of Statistical Mechanics: Theory and Experiment* 2009.07 (2009), p. L07003.

[66] Inho Hong, Morgan R Frank, Iyad Rahwan, Woo-Sung Jung, and Hyejin Youn. "A common trajectory recapitulated by urban economies." In: *arXiv preprint arXiv:1810.08330* (2018).

[67] Luca Garavaglia. "The distribution of advanced business services in Northern Italy: towards a polycentric metropolis model?" In: *Métropoles* 14 (2014).

[68] Luıés MA Bettencourt, José Lobo, and Geoffrey B West. "Why are large cities faster? Universal scaling and self-similarity in urban organization and dynamics." In: *The European Physical Journal B* 63.3 (2008), pp. 285–293.

[69] Mark EJ Newman. "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality." In: *Physical review E* 64.1 (2001), p. 016132.

[70] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks." In: *Rev. Mod. Phys.* 74 (1 2002), pp. 47–97.

[71] Sune Lehmann. "Fundamental Structures in Temporal Communication Networks." In: *Temporal Network Theory*. Ed. by Petter Holme and Jari Saramäki. Cham: Springer International Publishing, 2019, pp. 25–48. ISBN: 978-3-030-23495-9.

[72] John Tang, Ilias Leontiadis, Salvatore Scellato, Vincenzo Nicosia, Cecilia Mascolo, Mirco Musolesi, and Vito Latora. "Applications of temporal graph metrics to real-world networks." In: *Temporal Networks*. Springer, 2013, pp. 135–159.

[73] Huiji Gao and Huan Liu. "Mining human mobility in location-based social networks." In: *Synthesis Lectures on Data Mining and Knowledge Discovery* 7.2 (2015), pp. 1–115.

[74] Hao Wang, Manolis Terrovitis, and Nikos Mamoulis. "Location recommendation in location-based social networks using user check-in data." In: *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM. 2013, pp. 374–383.

[75] Jing Yuan, Yu Zheng, Chengyang Zhang, Wenlei Xie, Xing Xie, Guangzhong Sun, and Yan Huang. "T-drive: driving directions based on taxi trajectories." In: *Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems*. ACM. 2010, pp. 99–108.

[76] Yu Zheng, Xing Xie, Wei-Ying Ma, et al. "Geolife: A collaborative social networking service among user, location and trajectory." In: *IEEE Data Eng. Bull.* 33.2 (2010), pp. 32–39.

[77] Yong Wang, Pei Zhang, Ting Liu, Chris Sadler, and Margaret Martonosi. *CRAWDAD dataset princeton/zebranet (v. 2007-02-14)*. Downloaded from https://crawdad.org/princeton/zebranet/20070214. Feb. 2007.

[78] Piotr Sapiezynski, Arkadiusz Stopczynski, David Dreyer Lassen, and Sune Lehmann. "Interaction data from the Copenhagen Networks Study." In: *Scientific Data* 6.1 (2019), p. 315. ISSN: 2052-4463.

[79] Lorenzo Lucchini, Sara Tonelli, and Bruno Lepri. "Following the footsteps of giants: modeling the mobility of historically notable individuals using Wikipedia." In: *EPJ Data Science* 8.1 (2019), p. 36. ISSN: 2193-1127.

[80] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. "Data clustering: a review." In: *ACM computing surveys (CSUR)* 31.3 (1999), pp. 264–323.

[81] John Tang, Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, and Vito Latora. "Small-world behavior in time-varying graphs." In: *Physical Review E* 81.5 (2010), p. 055101.

[82] Jari Saramäki and Esteban Moro. "From seconds to months: an overview of multi-scale dynamics of mobile telephone calls." In: *The European Physical Journal B* 88.6 (2015), p. 164.

[83] Eunjoon Cho, Seth A Myers, and Jure Leskovec. "Friendship and mobility: user movement in location-based social networks." In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011, pp. 1082–1090.

[84] Petter Holme and Jari Saramäki. "Temporal networks." In: *Physics reports* 519.3 (2012), pp. 97–125.

[85] Matthew J Williams and Mirco Musolesi. "Spatio-temporal networks: reachability, centrality and robustness." In: *Royal Society open science* 3.6 (2016), p. 160196.

[86] Huanhuan Wu, James Cheng, Silu Huang, Yiping Ke, Yi Lu, and Yanyan Xu. "Path problems in temporal graphs." In: *Proceedings of the VLDB Endowment* 7.9 (2014), pp. 721–732.

[87] B Bui Xuan, Afonso Ferreira, and Aubin Jarry. "Computing shortest, fastest, and foremost journeys in dynamic networks." In: *International Journal of Foundations of Computer Science* 14.02 (2003), pp. 267–285.

[88] Huanhuan Wu, James Cheng, Yiping Ke, Silu Huang, Yuzhen Huang, and Hejun Wu. "Efficient algorithms for temporal path computation." In: *IEEE Transactions on Knowledge and Data Engineering* 28.11 (2016), pp. 2927–2942.

[89] Ulrik Brandes. "A faster algorithm for betweenness centrality." In: *Journal of mathematical sociology* 25.2 (2001), pp. 163–177.

[90] Ingo Scholtes, Nicolas Wider, and Antonios Garas. "Higher-order aggregate networks in the analysis of temporal networks: path structures and centralities." In: *The European Physical Journal B* 89.3 (2016), p. 61.

[91] Hyoungshick Kim and Ross Anderson. "Temporal node centrality in complex networks." In: *Physical Review E* 85.2 (2012), p. 026107.

[92] Ioanna Tsalouchidou, Ricardo Baeza-Yates, Francesco Bonchi, Kewen Liao, and Timos Sellis. "Temporal betweenness centrality in dynamic graphs." In: *International Journal of Data Science and Analytics* (2019), pp. 1–16.

[93] Petter Holme. "Network reachability of real-world contact sequences." In: *Physical Review E* 71.4 (2005), p. 046119.

[94] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *Kdd* 96.34 (1996), pp. 226–231.

[95] Rui Xu and Donald Wunsch. "Survey of clustering algorithms." In: *IEEE Transactions on neural networks* 16.3 (2005), pp. 645–678.

[96] Derya Birant and Alp Kut. "ST-DBSCAN: An algorithm for clustering spatial–temporal data." In: *Data & Knowledge Engineering* 60.1 (2007), pp. 208–221.

[97] Ljubica Blazevic, J-Y Le Boudec, and Silvia Giordano. "A location-based routing method for mobile ad hoc networks." In: *IEEE Transactions on mobile computing* 4.2 (2005), pp. 97–110.

[98] Dieter Mitsche, Giovanni Resta, and Paolo Santi. "The random waypoint mobility model with uniform node spatial distribution." In: *Wireless networks* 20.5 (2014), pp. 1053–1066.

[99] Peter W Glynn and Donald L Iglehart. "Importance sampling for stochastic simulations." In: *Management science* 35.11 (1989), pp. 1367–1392.

[100] Luca Pappalardo, Salvatore Rinzivillo, and Filippo Simini. "Human mobility modelling: exploration and preferential return meet the gravity model." In: *Procedia Computer Science* 83 (2016), pp. 934–939.

[101] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. "Modelling the scaling properties of human mobility." In: *Nature Physics* 6.10 (2010), pp. 818–823.

[102] Hua Cai, Xiaowei Zhan, Ji Zhu, Xiaoping Jia, Anthony SF Chiu, and Ming Xu. "Understanding taxi travel patterns." In: *Physica A: Statistical Mechanics and its applications* 457 (2016), pp. 590–597.

[103] Yasuko Matsubara, Lei Li, Evangelos Papalexakis, David Lo, Yasushi Sakurai, and Christos Faloutsos. "F-trail: Finding patterns in taxi trajectories." In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2013, pp. 86–98.