Cyprus University of Technology

Faculty of Engineering and Technology

Doctoral Dissertation

LATENT GEOMETRY OF HUMAN PROXIMITY NETWORKS

Marco Antonio Rodríguez Flores

Limassol, November 2020

CYPRUS UNIVERSITY OF TECHNOLOGY

FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING, COMPUTER ENGINEERING AND INFORMATICS

Doctoral Dissertation

LATENT GEOMETRY OF HUMAN PROXIMITY NETWORKS

Marco Antonio Rodríguez Flores

Limassol, November 2020

Approval Form

Doctoral Dissertation

LATENT GEOMETRY OF HUMAN PROXIMITY NETWORKS

Presented by

Marco Antonio Rodríguez Flores

Supervisor: Dr. Fragkiskos Papadopoulos, Associate Professor at Cyprus University of Technology

Signature:____

Member of the committee: Dr. Andreas Andreou, Professor at Cyprus University of Technology

Signature:____

Member of the committee: Dr. George Pallis, Associate Professor at University of Cyprus

Signature:____

Cyprus University of Technology
Limassol, November 2020

Copyrights

The approval of the dissertation by the Department of Electrical Engineering, Computer Engineering and Informatics does not imply necessarily the approval by the Department of the views of the writer.

To my wife and family

Acknowledgments

First of all I want to thank God for giving me the opportunity to study in Cyprus. This would be impossible by my own means or the means of my family. I also thank my parents for their sacrifices to support my education to the best of their financial means. I am also grateful to my wife Maria, the reason I came to Cyprus, for her support and encouragement in this journey. I am also grateful to my brother Jorge, my sister Marlene and all the people that always pray for me, especially pastors Arturo, Mario and Nelson.

Academically, my deepest gratitude is by far to my advisor Fragkiskos Papadopoulos for giving me the opportunity to work with him. His knowledge and determination are admirable and push me to go beyond the rest and solve complex problems. Also he goes beyond his duty and has helped me to find financial support to complete my studies. The opportunity to work with him and becoming his PhD student would not have been possible without the financial support of the EU H2020 NOTRE project (grant 692058). My thanks go to everyone involved with the project and to the European Union's Horizon 2020 research and innovation programme for funding NOTRE.

Also I want to thank all my professors from Cyprus University of Technology. In particular professor Michael Sirivianos for giving me the opportunity to be a secondee during the summer of 2020 in the EU H2020 SECONDO project (grant 823997), and to everyone involved with the project that made this experience possible.

I have also to express my gratitude to my advisor during my bachelor studies in Benemérita Universidad Autónoma de Puebla, Mexico. She is now a good friend, professor María Beatríz Bernábe Loranca. She was the mentor that set me on the research and academia path, without her advice and support my professional choices would have been very different.

And last but not least, I would like to thank professor Vassos Soteriou, a kind soul that we lost too soon and unexpectedly this year, may he rest in peace. His classes were some of the most interesting during my studies but beyond that, he was the friendliest professor to me and would never miss an opportunity to talk to me whenever we bumped into each other at the university. My condolences for his family.

Abstract

Understanding the dynamics of human contact and movement patterns in a physical space is crucial to better understand the spread of contagious diseases, information transfer from person to person, social behavior and influence. To this end, in the last 15 years temporal networks known as human proximity networks have been captured in different settings and have been extensively studied. These networks are characterized by similar structural and dynamical properties regardless of the setting. Many of these properties are well understood and can be reproduced with simple models. However, when we examine complex social group dynamics, such as the observed recurrent formation of groups (components) that consists of the same people, simple descriptions have been elusive.

In this thesis, we elucidate the emergence of the observed properties of real human proximity networks and their complex group dynamics through geometric approaches. In the first part of this thesis, we explore the human movement patterns responsible for the emergence of the main properties of the networks but in particular the formation of recurrent components. We propose a model of mobile agents, where agents reside in a hidden metric similarity space. In this space the distances between the agents abstract their similarities and these similarities act as forces that direct their motion towards each other in the physical space, and determine the duration of their interactions. We show that this *force-directed motion* model reproduces the main properties of human proximity networks and simultaneously forms the elusive recurrent components observed in reality. Interestingly, results with this model point to a connection with the popular $\mathbb{S}^1$ model of traditional (non-mobile) complex networks, which is isomorphic to random hyperbolic graphs. In the second part of this thesis, we explore this connection and propose a minimal latent space model which reproduces all the main properties of human proximity networks as well as the formation of recurrent components. The simplicity of the model facilitates its mathematical analysis, allowing us to prove three important properties of the generated networks. These findings lead to the third part of this thesis, where we address the problem of mapping real human proximity networks into hyperbolic spaces. We show that this embedding process can be done using methods developed for traditional complex networks based on the $\mathbb{S}^1$ model. We justify the compatibility theoretically and experimentally. We produce hyperbolic maps of six different real systems, which can be used to identify communities, facilitate greedy routing, and predict future links with significant precision. Further, we show that the time when nodes become infected are positively correlated with their hyperbolic distance from the source of the infection in epidemic spreading simulations on the temporal network.

Contents

List of Publications

Title	Authors	Journal	Publication Date	DOI
Similarity Forces and Recurrent Components in Human Face-to-Face Interaction Networks	Marco Antonio Rodríguez Flores, Fragkiskos Papadopoulos	Physical Review Letters	18 December 2018	https://doi.org/10.1103/PhysRevLett.121.258301
Latent Geometry and Dynamics of Human Proximity Networks	Fragkiskos Papadopoulos, Marco Antonio Rodríguez Flores	Physical Review E	26 November 2019	https://doi.org/10.1103/PhysRevE.100.052313
Hyperbolic Mapping of Human Proximity Networks	Marco Antonio Rodríguez Flores, Fragkiskos Papadopoulos	Scientific Reports	20 November 2020	https://doi.org/10.1038/s41598-020-77277-7

Table 1: List of publications produced from this dissertation.

List of Tables

List of Figures

List of Variables

Variable	Description
N	Denotes the number of agents/nodes in the network or model
τ	Denotes the number of time slots/snapshots in the network or model
$\bar{n}$	Denotes the average number of interacting agents per snapshot in the network
$\bar{l}$	Denotes the average number of links (edges) per snapshot in the network
t	Denotes a specific time slot/snapshot number
$\bar{k}$	Denotes the average degree per snapshot
$\bar{k}_t$	Denotes the average degree in a specific network snapshot/time-slot t
$\bar{k}_{\text{aggr}}$	Denotes the average degree in the time-aggregated network
θ_i	Is the hidden variable of a node denoting its position on a circle, the latent metric space considered in the FDM and the dynamic-$\mathbb{S}^1$ models
κ_i	Is the hidden variable of a node denoting its expected per-snapshot degree in the network in the dynamic-$\mathbb{S}^1$ model
$\tilde{\kappa}_i$	Denotes the hidden degree of a node in the time-aggregated network
r_i	Denotes the radial position of a node in the hyperbolic space in the $\mathbb{H}^2$ model
χ_{ij}	Is the effective distance between two nodes. Proportional to the angular distance between the nodes over the product of their hidden degrees
$\tilde{\chi}_{ij}$	Denotes the effective distance between two nodes in the time-aggregated network
a_i	Is the activation probability of a node in the FDM model
μ_1	Is the exponential decay of the bonding forces in the FDM model
μ_2	Is the exponential decay of the attractive forces in the FDM model
F_0	Is the magnitude of the attractive forces in the FDM model
v	Is the magnitude of the random displacement in the FDM model
d	Is the interaction radius in the FDM model
L	Is the size of the Euclidean space in the FDM model
R	Is the radius of the latent space, a circle in the FDM and dynamic-$\mathbb{S}^1$ models
T	Is the temperature parameter in the dynamic-$\mathbb{S}^1$
α	Denotes the infection probability in the compartmental epidemic spreading models: SIS, SIR, SEIR or SI
β	Denotes the recovery probability in the compartmental epidemic spreading models: SIS, SIR, SEIR or SI

Table 2: List of the main variables used in throughout the thesis.

Chapter 1

Introduction

Human proximity networks are temporal networks representing the close-range proximity among humans in a physical space. They have been extensively studied in the past 15 years as they are critical for understanding the transmission of airborne diseases, the efficiency of information dissemination, social behavior, and influence [1, 9, 16, 24, 41, 42, 44, 48]. To this end, human proximity networks have been captured in different environments over days, weeks or months [1, 16, 24, 36, 45, 64, 102, 107]. Such time-varying networks are represented as a series of static graph snapshots. Each snapshot corresponds to an observation interval or time slot, which typically spans a few seconds to several minutes depending on the devices used to collect the data. The nodes in each snapshot are people and an edge between two nodes signifies that they had been within proximity range during the corresponding slot. At the finest resolution, each slot spans 20 seconds and the proximity range is 1.5 m. Such networks have been captured by the SocioPatterns collaboration [97] in closed settings, such as hospitals, schools, scientific conferences and workplaces, and correspond to face-to-face interactions [36, 45, 64, 102, 107]. At a coarser resolution, each snapshot spans several minutes and proximity range can be up to 10 m or more. Such networks have been captured in university dormitories, residential communities and university campuses [1, 24, 39].

Irrespective of the context, measurement period and measurement method, different human proximity networks have been shown to exhibit similar structural and dynamical properties [9, 101]. Examples of such properties include the broad distributions of contact and intercontact durations, and properties of the time-aggregated network such as weight and strength distributions [16, 44, 48, 101]. Interestingly, these and other properties of human proximity systems can be well reproduced by simple generative models [99, 100, 103, 112]. For example, in a model of mobile agents known as the *attractiveness model* [99, 100], modeling the motion patterns of individuals as random walks in a two dimensional space, is sufficient to reproduce these and many other properties of human proximity networks. However, in recent years, more complex characteristics of these networks have been investigated, which originate from motion patterns far from random [94]. Specifically, the recurrent formation of groups that consists of the same people, in other terms, connected components that appear recurrently throughout the network snapshots. These recurrent components are fundamental structures of human proximity networks that are crucial for tasks such as community detection and predicting future behavior [40, 94]. Thus, in this thesis we address the following research questions: i) Can we model human proximity networks with latent geometry approaches?; ii) Can the similarities abstracted in the latent metric space be the driving forces that form recurrent components that previous models in the literature do not capture?; iii) Is the underlying geometry of human proximity networks hyperbolic, like in the case of traditional (non-mobile) complex networks?, iv) How can we embed real human proximity networks into their latent geometry? and v) Can the embeddings be efficiently used for applications such as community detection, information dissemination, link prediction and epidemic spreading?

In the first part of this thesis, we study recurrent components and propose a model of mobile agents capable of forming them as well as reproducing other main properties of human proximity networks. The results obtained with this model lead to the second part of this thesis, where we propose a minimal latent geometry model that forgoes the motion component, yet it is capable of reproducing the same properties as the first model. We also demonstrate that the models can be used to simulate epidemic spreading on synthetic human proximity networks realistically. Finally, in the last part of this thesis, based on our latent geometry model, we solve the inverse problem of mapping real human proximity networks into their latent geometry and explore several applications.

1.1 Contributions

In this thesis we make several contributions towards understanding the emergence of the structural and dynamical properties observed in real human proximity networks, in particular the elusive recurrent formation of groups that consists of the same people. In this regard, our contributions are the following:

- In Chapter 4, we propose a model of mobile agents where the social dynamics responsible for the formation of recurrent components in human proximity networks, find a natural explanation in the assumption that the agents of the temporal network reside in a hidden similarity space. Distances between the agents in this space act as similarity forces directing their motion towards other agents in the physical space and determining the duration of their interactions. By contrast, if such forces are ignored in the motion of the agents recurrent components do not form, although other main properties of such networks can still be reproduced. This work has been published in Physical Review Letters [87].

  Interestingly, without enforcing it into the model, the per-snapshot connection probability resembles qualitatively the connection probability of the known $\mathbb{S}^1$ of traditional (non-mobile) complex networks, which is isomorphic to random hyperbolic graphs [55, 56]. Our next contribution originates from this result.

- In Chapter 5, we propose a minimal latent space model where the main observed properties of human proximity networks, including the elusive recurrent components, emerge naturally and simultaneously. This model does not model node mobility directly, but captures the connectivity in each snapshot–each snapshot in the model is a realization of the $\mathbb{S}^1$ model. By forgoing the motion component the model facilitates mathematical analysis, allowing us to prove the contact, inter-contact and weight distributions. Further, we show that paradigmatic epidemic and rumor spreading processes perform similarly in real and modeled networks. This work has been published in Physical Review E [78].

  This model also simplifies our final research question: Can we embed real human proximity networks into hyperbolic spaces and obtain meaningful results?

- In Chapter 6, we propose a methodology to embed real human proximity networks into hyperbolic spaces according to our proposed latent space model. Network snapshots are often very sparse in human proximity networks, consisting of a small number of interacting (i.e., non-zero degree) nodes. Yet, we show that the time-aggregated representation of such systems over sufficiently large periods can be meaningfully embedded into the hyperbolic space, using methods developed for traditional (non-mobile) complex networks. We justify this compatibility theoretically and validate it experimentally. We produce hyperbolic maps of six different real systems, and show that the maps can be used to identify communities, facilitate efficient greedy routing on the temporal network, and predict future links with significant precision. Further, we show that epidemic arrival times are positively correlated with the hyperbolic distance from the infection sources in the maps. This work has been accepted in Scientific Reports. The pre-print is available in [88].

Chapter 2

Methodology

In this thesis we study human proximity networks through geometric approaches. We propose generative models where a latent similarity space is the main mechanism that explains the observed properties of this networks, including the formation of recurrent components that previous models from the literature do not capture. Further, we develop a geometric framework for the embedding of real human proximity networks into their latent geometry based on one of the generative models we propose, which allows the efficient use of the embeddings for several important applications: community detection, greedy routing, link prediction and the prediction of epidemic arrival times. Here we provide an overview of the methodology followed to achieve these goals.

2.1 Data

We started by analyzing real human proximity networks from two popular sources: SocioPatterns [97] and the MIT Human Dynamics Lab [85]. The networks from SocioPatterns that we consider, capture the face-to-face interactions among people in closed settings such as a Hospital [107], a Primary School [102], a High School [64], a Scientific Conference [45] and an Office Building [35]. All the networks have a proximity range among individuals of up to $\sim$ 1.5m and time slot durations of 20 seconds. From the MIT Human Dynamics Lab, we consider the Friends & Family network [1], corresponding to a residential community; and the MIT Social Evolution network [24], corresponding to a university dormitory. In these networks the proximity range among individuals is up to $\sim$ 10m and time slot durations of $\sim$ 5 minutes. For further details see Section 3.1.

Then, we investigated the following properties of the networks: contact and inter-contact duration distributions, weight and strength distributions, group size and group interaction duration distributions, and the distribution of the shortest time-respecting paths. We observed that these properties are similar in all the networks, regardless of the setting or the devices used to collect the data, as reported in the literature [101]. Inspired by the results of Sune, et. al. [94], we also studied group dynamics in the networks. We used the Disjoint Set Union algorithm [30] to find the connected components formed in each snapshot of the networks. We observed that connected components formed by the exact same nodes appear recurrently and abundantly through out the network snapshots in all real networks considered. We did the same with synthetic networks generated with a popular model of human proximity networks, known as the *attractiveness model* [99, 100]. In these networks recurrent components are almost non-existent because the social dynamics responsible for their formation is far from random [94].

2.2 Modeling Approach

In the first and second parts of this thesis we give answer to the research question of whether we can model human proximity networks with geometric approaches and if the assumption of a hidden similarity space underlying the network could be the main mechanism from which recurrent components emerge. Specifically, in the first part of the thesis, we propose a model of mobile agents where the distances between the agents in a latent similarity space act as similarity forces dictating their motion in the physical space as well as the duration of their interactions. This model is called *force-directed motion* (FDM) model and draws inspiration from Langevin dynamics, a known approach from Physics used to model the dynamics of molecular systems. Each agent moves towards other agents according to the summation of the similarity forces exerted on the agent by the other agents but the agent's motion is also affected by a random force accounting omitted degrees of freedom. The intuitive idea behind this motion hypothesis is that in reality, we do not interact with random people but with people that are similar to us. However, human interactions are not deterministic and can also occur randomly, hence the introduction of random forces in the motion akin to Langevin dynamics. We validated the model comparing properties of real human proximity networks with the properties of their synthetic counterparts generated with the FDM, including the formation of recurrent components. We also simulated epidemic spreading in real networks and their synthetic counterparts using a known compartmental model known as the Susceptible Infected Susceptible (SIS) model [46]. In all cases we observe similar properties and epidemic spreading behavior between the real networks and their synthetic counterparts.

For the second part of the thesis, we observed that the per-snapshot connection probability in synthetic networks generated with the FDM qualitatively resembles the Fermi-Dirac connection probability of a popular latent space model for traditional (non-mobile) complex networks known as the $\mathbb{S}^1$ model [55, 95], although this is not enforced into the model. This is a quite interesting observation because the $\mathbb{S}^1$ model is equivalent to random hyperbolic graphs or the $\mathbb{H}^2$ model [55]. Thus, in the second part of the thesis we give affirmative answer to the question "Is the underlying geometry of human proximity networks hyperbolic, like in the case of traditional (non-mobile) complex networks?". We developed a minimal latent space model named dynamic-$\mathbb{S}^1$ that forgoes the motion component in favor of simplicity. The model assumes that each network snapshot of a human proximity network is a realization of the $\mathbb{S}^1$ model. As with the FDM, we validated the model comparing properties of real networks with properties of their synthetic counterparts generated with the dynamic-$\mathbb{S}^1$, as well as diffusion process behavior with the SIS epidemic spreading model and a rumor spreading model known as DK model [22]. Further, we used mathematical analysis to prove that the contact, inter-contact and weight distributions in the model are power laws with exponents $2 + T$, $2 - T$ and $1 + T$, respectively.

2.3 Hyperbolic embedding method

Given the results obtained with the dynamic-$\mathbb{S}^1$ model, in the last part of the thesis we are interested in embedding real human proximity networks into hyperbolic spaces. We answer our two final research questions: We showed that we can embed real human proximity networks into hyperbolic spaces by embedding their time-aggregated representation using methods developed for the $\mathbb{S}^1/\mathbb{H}^2$ models, and that the resulting embeddings can be efficiently used for a variety of applications.

First we showed that the connection probability of the time-aggregated networks in the dynamic-$\mathbb{S}^1$ model is similar to the Fermi-Dirac connection probability in the $\mathbb{S}^1$ model. Then we validated this experimentally, using a state-of-the-art embedding method known as Mercator [32]. We showed that the quality of the inferred embeddings obtained with the original version of the method is quantitatively similar to the inferred embeddings obtained with a modified version adapted to the time-aggregated connection probability of the dynamic-$\mathbb{S}^1$ model.

Finally, we visualized the hyperbolic maps of six different real networks and showed that these maps can be used to visually detect communities. We also implemented a simple greedy routing algorithm to forward packets between pairs of nodes using their inferred hyperbolic distances and showed that high success ratios can be achieved (close to 100% for nodes at smaller hyperbolic distances). We also showed that whether two nodes will interact in a day or not can be predicted with significant precision if we know their hyperbolic distance inferred from a previous day. Regarding epidemic spreading, we also showed that the time slot when a node becomes infected in Susceptible Infected (SI) simulations is significantly correlated with the inferred hyperbolic distance between the node and the source of the infection.

Chapter 3

Related Work

3.1 Real human proximity networks

In the literature, the most widely studied real human proximity networks are from *SocioPatterns* [97] and from the *MIT Human Dynamics Lab* [85]. The SocioPatterns data correspond to human proximity networks in diverse settings such as: a Hospital ward in Lyon [107]; a Primary School in Lyon [102]; a High School in Marseilles [64]; a scientific Conference (Hypertext 2009) in Turin [45]; and an office building in Saint Maurice [36]. The data were collected through the use of Radio-Frequency Identification (RFID) badges worn by individuals. Interactions were detected only if the badges were within 1-1.5 meters in front of each other and exchanged at least 1 radio packet in a 20 seconds interval. Therefore each time slot in the data has duration 20 seconds and corresponds to a network snapshot, whereas the proximity range implies face-to-face interaction. The data from the MIT Human Dynamics Lab correspond to human proximity networks in settings such as: a residential community [1], a university campus [26] and a university dormitory [62]. The data were collected with the Bluetooth capabilities of mobile phones carried by the participants. The phones detect the proximity of other phones within a radius of $\sim$ 10 meters in all directions, including different floors. Thus proximity in these networks does not imply face-to-face interaction. The resolution of these datasets is $\sim$ 5 minutes, which is the frequency by which the phones emitted a Bluetooth signal to be detected by other phones nearby.

Below we describe each real network considered in this thesis. Starting with face-to-face interaction networks of SocioPatterns.

(i) Hospital. The data were collected during a period of 5 days (December 6-10, 2010) and involve $N = 75$ nodes (29 patients and 46 health-care workers) in a hospital ward. There are two working shifts, a morning-afternoon shift and an afternoon-night shift. Health-care workers that are present in one shift are usually not present in the other shift. Each day corresponds to a cycle of recorded activity beginning at the earliest recorded interaction and ending at the last interaction recorded in the day. There are 43-46 nodes present and 2177-3889 time slots in each activity cycle. The network has a total number of 17376 time slots, including the time slots during the periods of inactivity between different days.

(ii) Primary School. The data were collected during a period of 2 days (October 1st, 2nd, 2009) and involve $N = 242$ nodes (232 children and 10 teachers) in a primary school. Each day corresponds to a cycle of recorded activity during working hours from 8:30am to 4:30pm [102]. Cycle 1 has duration of 1555 slots and consists of 238 nodes, while cycle 2 has duration of 1545 slots and consists of 236 nodes. The network has a total number of 5846 time slots, including the time slots during the periods of inactivity between different days.

(iii) High School. The data were collected during a period of 5 days (December 2-6, 2013) and involve $N = 327$ nodes (students) in a high school. Each day corresponds to a cycle of recorded activity beginning at the earliest recorded interaction and ending at the latest recorded interaction during working hours. Activity cycle 1 has duration 899 time slots, while each of the activity cycles 2-5 has duration 1619 time slots. There are 295-312 nodes present in each activity cycle. The network has a total of 18179 time slots, including the time slots during the periods of inactivity between different days.

(iv) Conference. The data were collected during a period of 2.5 days (June 29th to July 1st, 2009) and involve $N = 113$ nodes (participants) in a conference. Each day corresponds to a cycle of recorded activity beginning at the earliest and ending at the latest recorded interaction during the hours of the conference. Activity cycles 1, 2, 3 have durations $2874, 2210, 1946$ time slots, respectively. There are 97-102 nodes present in each activity cycle. The network has a total of 10618 time slots, including the time slots during the periods of inactivity between different days.

(v) Office Building. The data were collected during a period of ~ 2 weeks in 2015 and involve $N = 217$ nodes (employees) in an office building. Including only working days there are 10 days. Each day corresponds to a cycle of recorded activity beginning at the earliest recorded interaction and ending at the latest recorded interaction during working hours. There are 209-215 nodes present and 1973-2159 time slots in each activity cycle. The network has a total number of 49678 time slots, including the time slots during the periods of inactivity between different days and the weekend period.

The following human proximity networks are from the MIT Human Dynamics Lab.

(vi) MIT Social Evolution. The data were collected during a period of 8 months (October 2008 - May 2009) and involve $N = 74$ nodes (students) in a dormitory of a major university in the United States. Each day in this network corresponds to interactions recorded during all the day. The network has a total of 60905 time slots.

(vii) Friends & Family. The data were collected during a period of 8 months (October 2010 - May 2011) and involve $N = 131$ nodes (residents) in a community adjacent to a major university in the United States. Each day in this network corresponds to interactions recorded during all the day. The network has a total of 57961 time slots.

3.2 Properties of human proximity networks

The main properties of human proximity networks include properties measured on the temporal network itself but also on its time-aggregated representation. The time-aggregated network consists of the aggregation of network snapshots into a static weighted network. In this network two nodes are connected if they were within proximity in at least one network snapshot and the weight of the edge is the total number of snapshots where the nodes remained within proximity [40].

The main properties of human proximity networks can be classified into three categories: the *individual* or *microscopic properties*, the *group* or *mesoscopic* properties and the *collective* or *macroscopic* properties [100].

Here we describe the properties that we consider in this thesis, which have also been considered in previous related works [99, 100, 101]).

- Microscopic

    (a) *Distribution of contact durations.* This is the distribution of the time duration (in number of time slots) that two nodes remain in contact (interact).

    (b) *Distribution of intercontact durations.* This is the distribution of time (in number of time slots) that elapses between the last time that a pair of nodes interacted till the time that the same pair of nodes interacts again.

    (c) *Weight distribution.* The weight distribution is the distribution of the edge weights of the time-aggregated network.

    (d) *Strength distribution.* This is the distribution of node strengths in the time-aggregated network. The strength of a node is the sum of the weights of all edges attached to the node.

    (e) *Average node strength as a function of node degree.* From the time-aggregated network of contacts we also compute the degree of each node (sum of edges attached to the node) and for each degree we compute the average strength among nodes with that specific degree.

- Mesoscopic

    (f) *Distribution of component sizes.* This is the distribution of the number of nodes in the connected components formed throughout the observation time, including components of size 2.

    (g) *Average total interaction duration of a group as a function of its size.* The total interaction duration of a group of nodes is the total number of time slots throughout the observation time where the exact same group of nodes formed a connected component. For each group size we compute the average of this duration among groups with that specific size.

- Macroscopic

    (h) *Distribution of shortest time-respecting paths.* Consider three nodes i, k and j, where i and k interact at slot t and k and j interact at slot $t' > t$. In this example, the time-respecting path between i and j is $i \rightarrow k \rightarrow j$ and has length 2. The shortest time-respecting path between i and j is the shortest such path throughout the observation time. We consider the distribution of lengths of the shortest time-respecting paths among all pairs of nodes [40, 100].

3.3 Generative models of human proximity networks

Many generative models for temporal networks have been proposed [40]. However, generative models that specifically model the main characteristics of human proximity networks are few [99, 100, 103, 112]. Here we discuss two popular models, which are minimal, yet are capable of reproducing many of the main properties observed in real systems.

The agent-based model proposed in [103], models the characteristic distributions of contact durations, intercontact durations, weight and strength observed in reality. The main idea of the model is to form groups of interacting agents according to a simple mechanism, which assumes that the longer an agent is interacting in a group, the less likely it is to leave it, while the longer an agent is isolated (not interacting) the less likely it is to form a new group. The model assigns a *sociability* value η_i to each agent $i = 1, \dots, N$, sampled uniformly at random from $[0, 1]$. Each node i also has a *coordination* value n_i, which is the current degree of the node and a value t_i, which is the time slot at which n_i last changed. The model begins with random initial conditions and proceeds in a time slotted manner. In each time step t, the following steps are performed:

1. Chose a random agent i
2. Update the node's current degree n_i as follows:

    a) If the node is isolated ($n_i = 0$), the node initiates an interaction with probability $p_0(t, t_i) = \frac{\eta_i}{1+(t-t_i)/N}$ with an isolated node j chosen with probability $p_0(t, t_j)$. Set $n_i = n_j = 1$.

    b) If the node is interacting in a group of size n ($n_i = n$), with probability $p_n(t, t_i) = \frac{1-\eta_i}{1+(t-t_i)/N}$, the node either leaves the group or introduces another isolated node to the group:

        i. With probability λ, the node leaves the group. Set $n_i = 0$, $n_k = n - 1$ for all k nodes remaining in the group.

        ii. With probability $1 - \lambda$, the node introduces another isolated node j into the group, chosen with probability $p_0(t, t_j)$. Set $n_j = n + 1$, $n_k = n + 1$ for all k nodes in the group.

    c) With probability $1 - p_n(t, t_i)$ (if $n_i = n$) or $1 - p_0(t, t_i)$ (if $n_i = 0$), node i does not change its current state (n_i).

The mechanism of the model that forms groups (components) favors the most active (sociable) agents to form groups. However, the agent selected and the group (or agent) to join are random. Thus, recurrent components do not form as abundantly as in reality.

The model of mobile agents proposed in [99, 100] is capable of reproducing all properties described in Section 3.2. The main idea in this model is that the agents have an intrinsic social attractiveness and they perform random walks in a two dimensional space, abstracting a physical location. When an agent is within proximity of another agent they may stop to interact with each other. Agents have a higher probability to remain interacting with more attractive agents, while they are more likely to resume mobility if the agents within proximity are less attractive. Specifically, the model assigns an *attractiveness* value s_i and an activation probability a_i to each agent $i = 1, \dots, N$, both sampled uniformly at random from $[0, 1]$. Initially, the agents are distributed uniformly at random in a closed box of linear size L where they perform random walks. The agents that are within distance d from each other are considered as interacting and do not move, while the rest of the agents are considered as inactive. Time in the model is slotted and each time step t consists of the following steps:

1. Each inactive agent i becomes active with probability a_i.
2. Each interacting agent i escapes from interactions with probability $p_i(t) = 1 - \max_{j \in \mathcal{N}_i(t)}\{s_j\}$, where $\mathcal{N}_i(t)$ is the set agents interacting with i at time t.
3. Each active and escaped agent i moves towards a direction ϕ sampled uniformly at random from $[0, 2\pi]$ with a displacement magnitude v.
4. All agents that are within distance d from each other are considered as interacting and stop moving, while the rest of the agents are set as inactive.

The model assumes $d = v = 1$ and the size of the box L is used to tune the resulting properties. A small box produces a denser space where larger groups form and the average degree in the time-aggregated network is larger. Although simple, the model reproduces all properties described in section 3.2. However, as it has been shown in [94], the random motion of the agents cannot reproduce the formation of recurrent components observed in real systems. This also means that any generative model where groups are formed by completely random mechanisms will have the same limitation.

It is then important to develop generative models capable of reproducing the behavior of the formation of recurrent components because they are fundamental structures of these networks that are crucial to better understand their dynamics and have predictive power [40, 94]. In this thesis we propose two generative models based on latent metric spaces. First, we propose a model of mobile agents where the distances among them in a latent metric space act as similarity forces that drive their motion and determine the duration of their interactions. This model offers new insights into the human behavior responsible for the formation of recurrent components, based on a known approach from Physics that models molecular dynamics. Second, we forgo the motion aspect in favor of simplicity and propose a minimal latent space model, which models the connectivity among the agents in each snapshot. The model assumes that each snapshot is a realization of a well-known latent space model for traditional (non-mobile) complex networks. The simplicity of the model facilitates the implementation of inference methods by using the model as a base to embed real human proximity networks into their latent geometry. The realism of the model yields meaningful embeddings that can be efficiently used for several applications, as shown in the last chapter of this thesis.

Chapter 4

Similarity forces and recurrent components in human face-to-face interaction networks

This chapter has been published, with some modifications, in “Physical Review Letters” [87].

Understanding the mechanisms that drive the dynamics of face-to-face interaction networks is crucial for better analyses of spreading phenomena. In particular, phenomena that evolve as fast as real-time face-to-face interactions, such as respiratory transmitted diseases, word-of-mouth information transfer and viruses in mobile networks [10, 40, 65]. Furthermore, deriving efficient epidemic control strategies requires an accurate description of fast-evolving contagions [10, 41, 42, 49, 71]. However, a complete understanding of the processes responsible for the structural and dynamical properties of face-to-face interaction networks has been an elusive task [8, 40, 94].

Face-to-face interaction networks portray social interactions in closed settings such as schools, hospitals, offices, etc. A typical representation consists of a series of network snapshots. Each snapshot corresponds to an observation interval, which can span from a few seconds to several minutes depending on the devices used to collect the data [97, 101]. The agents (nodes) in each snapshot are individuals and an edge between any two agents represents a direct face-to-face interaction.

Analyses of such networks have uncovered universal properties, such as the heavy-tailed distributions of the interaction duration and time between consecutive interactions, cf. [45]. Previous results point to the idea of social attractiveness as a mechanism responsible for these universal properties and for other structural characteristics of the time-aggregated network of contacts, like its degree, weight and strength distributions [99, 100, 101]. Specifically, in the attractiveness model [99, 100] agents have an activation probability a_i and a *global attractiveness* value s_i that are sampled uniformly at random from $[0, 1]$. Time is slotted and in each slot each non-interacting agent i is active with probability a_i. Active agents perform *random walks* in a closed Euclidean space moving towards a random direction every slot with a constant velocity (displacement) v. Agents stop moving to interact whenever they encounter another agent within a threshold distance d. The activation probability represents the activeness of each agent in the social event. The global attractiveness of the agents defines an *escaping probability* from the interactions. For instance, an agent i that has stopped moving in order to interact with other agents within distance d, can resume mobility with probability $1 - \max_{j \in \mathcal{N}_i}\{s_j\}$, where $\mathcal{N}_i$ is the set of agents interacting with i [100]. Therefore, longer interactions occur when an individual with a high global attractiveness s_j is involved.

However, it has been recently revealed that face-to-face interaction networks exhibit structural and dynamical properties such as community formation, which originate from motion patterns that are far from random [94]. In a temporal setting, communities are dynamic, meaning that their structure and size change over time. A common strategy to track dynamic communities is to construct their evolution timelines by aggregating connected components of at least three nodes in different time slots, according to some similarity measure [37, 94]. In other words, the building blocks of dynamic communities are connected components that appear recurrently. If we extract the connected components in each time slot of a real face-to-face interaction network, we can see that many of the exact same components appear several times throughout the observation period. Indeed, in Figs. 4.1a-c we have extracted and assigned IDs, in order of appearance, to the unique components found in three real-world datasets from SocioPatterns [97]: a Hospital, a Primary School and a High School [64, 102, 107] (see Table 4.1, Section 3.1, Sec. 4.1 and Appendix A.1, where we also consider a fourth dataset from a conference [45]). The blue lines in Figs. 4.1a-c represent *recurrent components*, i.e., components that appeared at least once in a previous time interval. By contrast, in the attractiveness model we observe very few recurrent components (Fig. 4.1d, Sec. 4.1 and Appendix A.1), even though the model accurately reproduces the broad distributions of contact durations and of times between consecutive contacts (Figs. 4.1f,g). This is because in the model nodes drift according to their own random trajectories and the probability for a group of at least three nodes to meet again is vanishing. In other words, components form in this model purely based on chance.

Dataset	N	τ	$\bar{n}$	$\bar{l}$	Cycles	μ_1	F_0	μ_2
Hospital	70	4400	7.09	4.7	4	0.8	0.12	0.9
Primary School	242	3100	56.38	40.57	2	0.35	0.2	0.78
High School	327	7375	41.89	25.56	5	1.2	0.11	0.86
Conference	113	7030	4.98	2.96	3	2.65	0.02	3.6

Table 4.1: Analyzed datasets. N is the total number of agents; τ is the total duration of the dataset in slots of 20 seconds after removing the periods without interactions between consecutive days; $\bar{n}$, $\bar{l}$ are the average numbers of interacting agents and links (interactions) per slot. The activity cycles correspond to observation periods in different days (see Section 3.1). μ_1, F_0, μ_2 are the FDM parameters used in the simulated counterpart of each real network (see text).

Here we present a model of mobile agents where their motion is not totally random, but instead it is also directed by pairwise similarity forces. We show that this model can capture the most distinctive features of face-to-face interaction networks including their observed recurrent component patterns. In addition to the two-dimensional Euclidean space where agents move and interact (an $L \times L$ square), agents in the model also reside in a hidden similarity space, where coordinates abstract their similarity attributes. Distances between the agents in this space act as similarity forces directing their motion towards other agents in the physical space and determining the duration of their interactions. We consider the simplest metric space as the similarity space, which is a circle of radius $R = N/2\pi$ where each agent $i = 1, 2, \ldots, N$ is assigned a random angular coordinate $\theta_i \in [0, 2\pi]$.

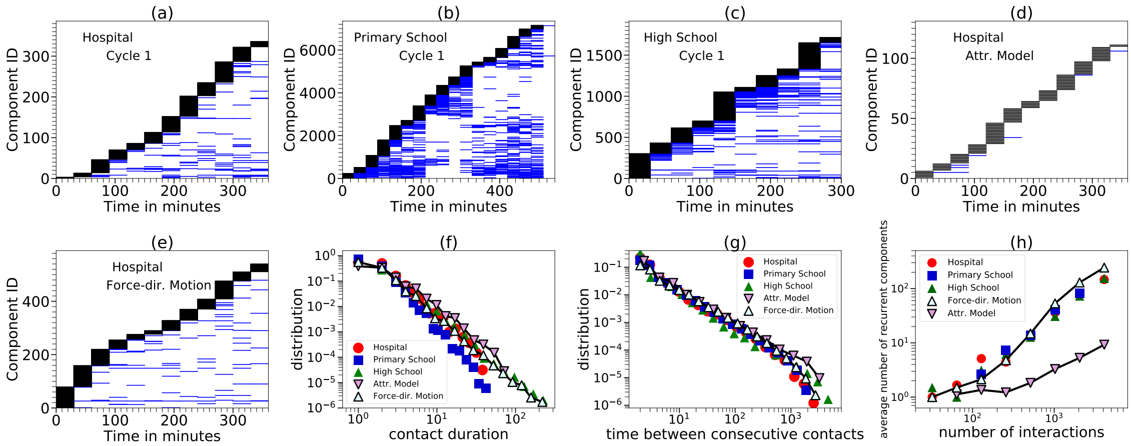

Figure 4.1: Recurrent component patterns and distributions of contact durations and of times between consecutive contacts in three real-world datasets and in simulated networks. **(a-c)** Components found in the first activity cycle of the Hospital, Primary School and High School (6, 8.6 and 5 hours, respectively). **(d)** Components found in a simulation of the attractiveness model with the same duration as in (a). **(e)** Same as (d) but with the FDM (Force-dir. Motion) model. **(f, g)** Distribution of contact duration and of time between consecutive contacts in real and simulated networks. **(h)** Average number of recurrent components where an agent participates as a function of its total number of interactions in real and simulated networks. The blue lines in (a-e) correspond to *recurrent* components while the black lines to components appearing for the first time, i.e., to the *unique* components. The x-axis is binned into 30 minute intervals, while the y-axis shows the component IDs observed in each bin; all components consist of at least three nodes. The simulations with the models use the parameters of the Hospital (Table 4.1 and Sec. 4.3). In (f-h) the results with the models are averages over 10 simulation runs. Results for all activity cycles, the Conference dataset, and for the simulated counterparts of the rest of the real networks are found in Secs. 4.1, 4.2 and Appendices A.1, A.2.

Therefore the similarity distance between two agents i, j is $s_{ij} = R\Delta\theta_{ij}$, where $\Delta\theta_{ij} = \pi - |\pi - |\theta_i - \theta_j||$ is the angular distance between the agents. (We also consider non-uniformly distributed coordinates in Appendix 4.6, obtaining similar results.)

Time in the model is slotted and at the beginning of each slot agents can be in one of two states: *inactive* or *interacting*. Inactive agents move in the slot only if they become active, while interacting agents move only if they escape their interactions. At the beginning of each slot t, each inactive agent i is activated with a preassigned probability a_i. Furthermore, each interacting agent i escapes its interactions with probability

$$P_i^e(t) = 1 - \frac{1}{|\mathcal{N}_i(t)|} \sum_{j \in \mathcal{N}_i(t)} e^{-s_{ij}/\mu_1}, \tag{4.1}$$

where $\mathcal{N}_i(t)$ is the set of agents that i is currently interacting with and s_{ij} is the similarity distance between agents i and j. The summands in Eq. (4.1) can be seen as *bonding forces* that decrease exponentially with the similarity distance, while parameter $\mu_1 > 0$ is the decay constant controlling the importance of these forces as the similarity distance increases and allowing us to tune the average contact duration (Sec. 4.3). The model assumes that the contact duration in number of slots between two agents i, j is exponentially distributed with rate s_{ij}/μ_1. The discrete analogue of this distribution is the geometric distribution with success probability $p_{ij} = 1 - e^{-s_{ij}/\mu_1}$. Therefore, Eq. (4.1) is the average of $p_{ij}, j \in \mathcal{N}_i(t)$.

Each moving agent i in the slot updates its position $(x_i^t,\, y_i^t)$ according to the following motion equations

$$x_i^{t+1} = x_i^t + \sum_{j \in \mathcal{S}(t)} F_{ij} \frac{(x_j^t - x_i^t)}{\sqrt{(x_j^t - x_i^t)^2 + (y_j^t - y_i^t)^2}} + R_i^x, \tag{4.2}$$

$$y_i^{t+1} = y_i^t + \sum_{j \in \mathcal{S}(t)} F_{ij} \frac{(y_j^t - y_i^t)}{\sqrt{(x_j^t - x_i^t)^2 + (y_j^t - y_i^t)^2}} + R_i^y, \tag{4.3}$$

where $\mathcal{S}(t)$ is the set of all moving and interacting agents in the slot, while F_{ij} is the magnitude of the *attractive force* between agents i and j, which also decreases exponentially with their similarity distance,

$$F_{ij} = F_0 e^{-s_{ij}/\mu_2}. \tag{4.4}$$

Parameter $F_0 \geq 0$ is the force magnitude at the minimum similarity distance, $s_{ij} = 0$, while $\mu_2 > 0$ is the decay constant controlling the importance of the force magnitude as the similarity distance increases. Therefore, the sums in Eqs. (4.2), (4.3) are the total attractive forces exerted to agent i by the agents $j \in \mathcal{S}(t)$ along the x and y directions of the motion. The random motion components are $R_i^x = v \cos \phi_i$, $R_i^y = v \sin \phi_i$, where ϕ_i is sampled uniformly at random from $[0, 2\pi]$ and $v \geq 0$ is the magnitude of the random displacement. We can think of R_i^x, R_i^y as accounting for omitted degrees of freedom, akin to Langevin dynamics [93]. At $v = 0$ the motion becomes deterministic, while at $F_0 = 0$ it degenerates to random walks. Once the moving agents update their positions they either transition to the interacting state if they are within interaction range d from other non-inactive agents, or to the inactive state. We call the described model *Force-Directed Motion (FDM)* model.

To understand how the formation of components depends on F_0, μ_2, v, we first consider deterministic motion. In this case, the magnitude of the expected agent displacement is controlled by F_0 and μ_2. This magnitude can be kept fixed if, when F_0 decreases, μ_2 increases accordingly. As μ_2 increases, larger components form that involve agents at larger similarity distances, until the agents eventually collapse into a giant component. At the same time, the number of components initially increases and then decreases, see Fig. 4.2(a). The motion in Eqs. (4.2), (4.3) is deterministic motion with random noise. This noise decreases the chances for similar—close in the similarity space—agents to meet, which reduces the size of components. At the same time, it can either increase (if its magnitude v is sufficiently small) or decrease (if v is sufficiently large) the number of components (Fig. 4.2(b)).

To tune FDM's parameters in simulations of real networks we follow the procedure in Appendix 4.3. In a nutshell, we fix $v = d = 1$. The number of agents N and time slots τ are the same as in the real networks (Table 4.1). The activation probability a_i is either $a_i = 0.5$ for every agent i (Primary and High School), or sampled uniformly at random from $[0, 1]$. Parameters μ_1, F_0, μ_2 (Table 4.1) and the size of the Euclidean space L (Sec. 4.3, Table 4.2) are adjusted in order to approximately match the following quantities between simulated and real networks: (i) the average contact duration (using μ_1); (ii) the average number of recurrent components per interval of 10 minutes, while ensuring a similar size of the largest component formed (using F_0, μ_2); and (iii) the average agent degree in the time-aggregated network (using L).

In Fig. 4.1e we see that the FDM can reproduce a similar pattern of unique and recurrent components as in the Hospital (Fig. 4.1a), in stark contrast to the attractiveness model (Fig. 4.1d). Similar results hold for all cycles of activity and for

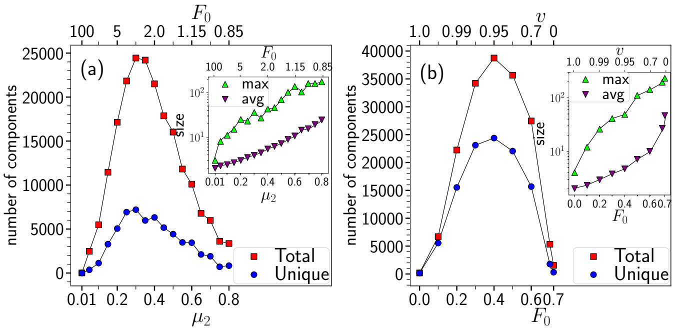

Figure 4.2: Formation of components in the FDM. **(a)** Number of components formed (total and unique) in deterministic motion ($v = 0$) for pairs of parameters μ_2 (bottom x-axis) and F_0 (top x-axis). **(b)** Same as (a) but for pairs of F_0 and $v \geq 0$. In both (a, b) as one parameter increases the other decreases so that the expected agent displacement per slot is always $\approx d = 1$. The insets show the maximum and average size across all components. In both plots $N = 242$, in (b) $\mu_2 = 1$. See also Sec. 4.3.

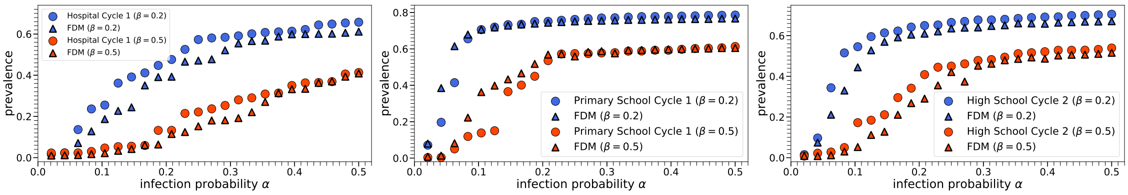

Figure 4.3: Average percentage of infected agents per time slot (prevalence) of the SIS model as a function of the infection probability α in real and simulated networks (circles and triangles respectively), for two recovery probabilities β. In the SIS each agent can be in one of two states, susceptible or infected. At any time slot an infected agent recovers with probability β and becomes susceptible again, whereas infected agents infect the susceptible agents with whom they interact, with probability α. To simulate the SIS process on temporal networks we use the dynamic SIS implementation of the Network Diffusion Library [90]. See Appendix A.3, for further details.

all considered datasets (Sec. 4.1 and Appendix A.1). In Fig. 4.1h we also see that the model can capture the correlations between the average number of recurrent components where an agent participated and the total number of interactions of the agent (see also Sec. 4.1.3). At the same time, the model reproduces the broad distributions of contact durations and of times between consecutive contacts (Figs. 4.1f,g). The model also adequately reproduces a range of other properties of the considered real networks, including weight distributions, distributions of component sizes and of shortest time-respecting paths, and group interaction durations (Sec. 4.2 and Appendix A.2). It is then not surprising that the susceptible-infected-susceptible (SIS) spreading process [51] behaves similarly in real and simulated networks (Fig. 4.3). Fig. 4.4 shows that agents close in the similarity space tend to stay closer to each other in the Euclidean space throughout the simulations and interact more often, as expected.

The exponential form of the attractive force in Eq. (4.4) promotes locality and the formation of small components, as observed in real data. This is also promoted by the metric property of the similarity space, i.e., the triangle inequality, which ensures that if an agent a is similar to an agent b and b is similar to a third agent c, then c is also similar to a. This means that these agents will tend to gather close to each other in the Euclidean space forming triangle abc. On the other hand, if similarity distances do not satisfy the triangle inequality, then agents a and c might be close to some other agents d and e, forming chain $dabce$ in the network. In other words,

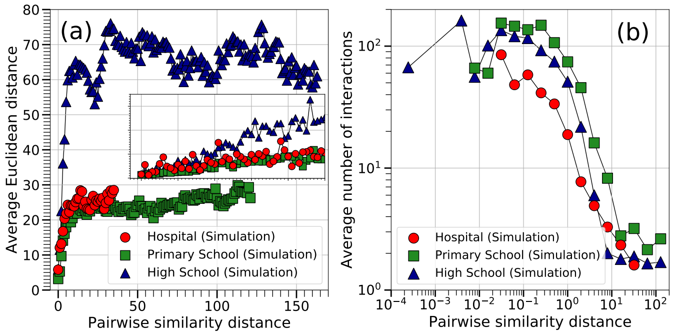

Figure 4.4: Average Euclidean distance and number of interactions between two agents as a function of their similarity distance, in simulated counterparts of the Hospital, Primary School and High School. The inset in (a) is a zoom in on similarity distances up to 5.

agents will tend to form larger components. We verify this argument in Appendix 4.5, where we break the triangle inequality by randomly assigning similarity distances to all pairs of agents instead of assigning to the agents similarity coordinates. In this way forces lose their localization effect and we see that a giant connected component, non-existent when the similarity space satisfies the metric property, forms in the middle of the Euclidean space.

In summary, forces emerging from similarity distances in metric spaces appear to provide a natural explanation for the observed recurrent component dynamics in face-to-face interaction networks. These forces direct the motion of the agents in the physical space and determine the agents' interaction durations. Motion based on these principles can still capture a wide range of other main properties of such networks, in addition to their recurrent component patterns. The interactions do not have to be exactly face-to-face or of few activity cycles. In Appendix A.1, A.2, we see that similar results hold in a longitudinal dataset from an MIT dormitory, where proximity was captured if mobile phones were within 10 meters from each other [24].

The modeling approach we consider bears similarities to N-body simulations and Langevin dynamics [93], suggesting that similar techniques and approaches from these well established areas of physics can be applicable to contemporary network science problems. Yet, we note that the similarity forces in our case only direct the motion of the agents in the physical space, and do not depend on the agents' distances in this space akin to gravity.

We also observe that *hyperbolic* spaces appear to underlie the topologies of traditional complex networks, whose degree distributions are heterogeneous [55]. In this case, the hidden distance between two nodes is not just the angular distance $R\Delta\theta$ but the effective distance $\chi = R\Delta\theta/(\kappa\kappa')$, where κ, κ' are the expected degrees of the nodes [55]. One can replace angular with effective distances in the FDM. However, in all datasets we considered, the distribution of κs was quite homogeneous to justify the need for this description [1]. Indeed, if we use effective distances in the FDM with the estimated κs from the real data we obtain very similar results (Sec. 4.7 and Appendix A.4).

[1] an agent's κ is its average degree per time slot

4.1 Recurrent components

4.1.1 Extraction process

Given a real or simulated network we first find all connected components in each time slot of the network using the Disjoint Set Union algorithm from [30]. Each identified component is a set of at least three nodes. We then go over all time slots from the beginning to the end and assign IDs $1, 2, \dots$, etc., to their components as follows. If a component is seen for the first time, i.e., it does not consist of exactly the same nodes as a component seen in a previous slot, it is assigned a new ID and it is marked as *unique*; if more than one unique components are found in a slot they are assigned new IDs arbitrarily. Otherwise, if a component consists of exactly the same nodes as a previously seen component, it is assigned the ID of that component and it is marked as *recurrent*.

In Figs. 4.1a-e, Fig. 4.5 below and in Appendix Figs. A.1- A.3, the observation period (x-axis) is binned into 30 minute intervals, while in Fig. A.4 the observation period is binned into 60 minute intervals. The black lines spanning each interval indicate in the y-axis the IDs of the unique components found in the slots of the interval. Similarly, the blue lines indicate the IDs of the recurrent components found in each interval.

4.1.2 Unique and recurrent components in real and modeled networks

Fig. 4.5 shows the unique and recurrent components in the Hospital and in corresponding simulated networks with the attractiveness and FDM models. Results are shown for each activity cycle. See Appendix A.1 for results with the other real networks considered in this chapter, as well as results with another generative model from the literature [103], described in Section 3.3.

We see that in simulated networks with the attractiveness model recurrent components are almost non-existent, while they are abundant in simulated networks with the FDM as in the real data. We note that if attraction forces are disabled in the FDM ($F_0 = 0$), agents perform random walks, and the results are similar to the attractiveness model.

4.1.3 Recurrent components and node interactions

Fig. 4.6 and Fig. 4.1h show the correlations between the average number of recurrent components where a node participates and its total number of interactions in the real datasets and in the corresponding simulated networks. The total number of interactions of a node i, I_i, is the total number of edges (interactions) between i and other nodes $j \neq i$ over the duration of the dataset. This metric is the same as the *strength* of the node in the time-aggregated network of contacts (Sec. 3.2). The number of recurrent components where a node participates is the total number of such components where the node is a member of over the duration of the dataset. We measure the recurrent components within intervals of 30 minutes in all cases, except for the MIT Social Evolution data where we use intervals of 60 minutes. A recurrent component appearing more than once in an interval is counted only once. We see that the FDM can better capture the behavior in the real networks compared

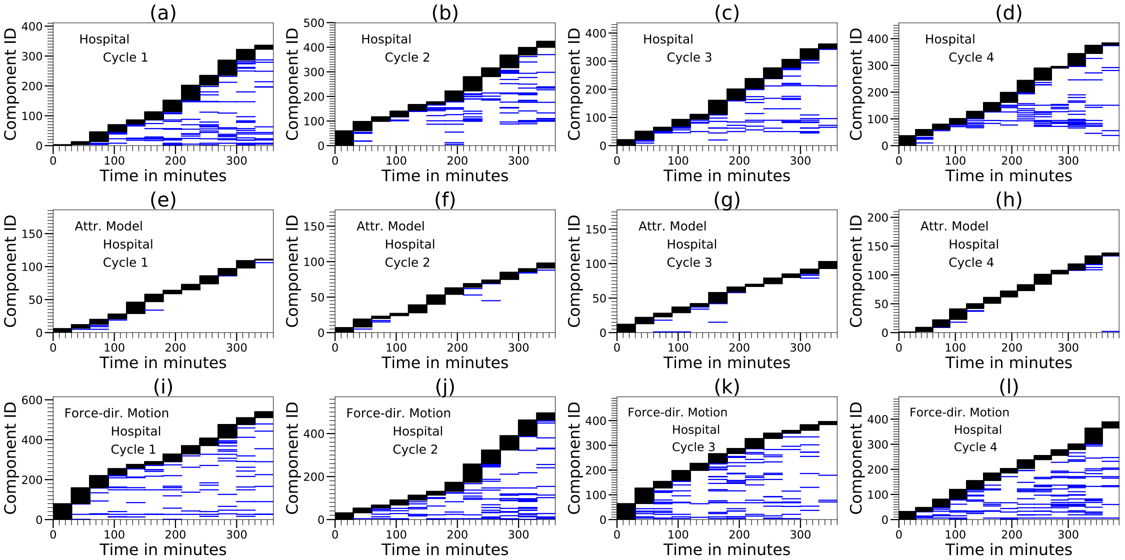

Figure 4.5: **(a-d)** Unique and recurrent components found in each cycle of activity in the Hospital. **(e-h)** Components found in a simulation run of the attractiveness model assuming activity cycles of the same durations as in (a-d). **(i-l)** Same as (e-h) but for the FDM (Force-dir. Motion) model. All simulations use the Hospital parameters (Table 4.2 in Sec. 4.3).

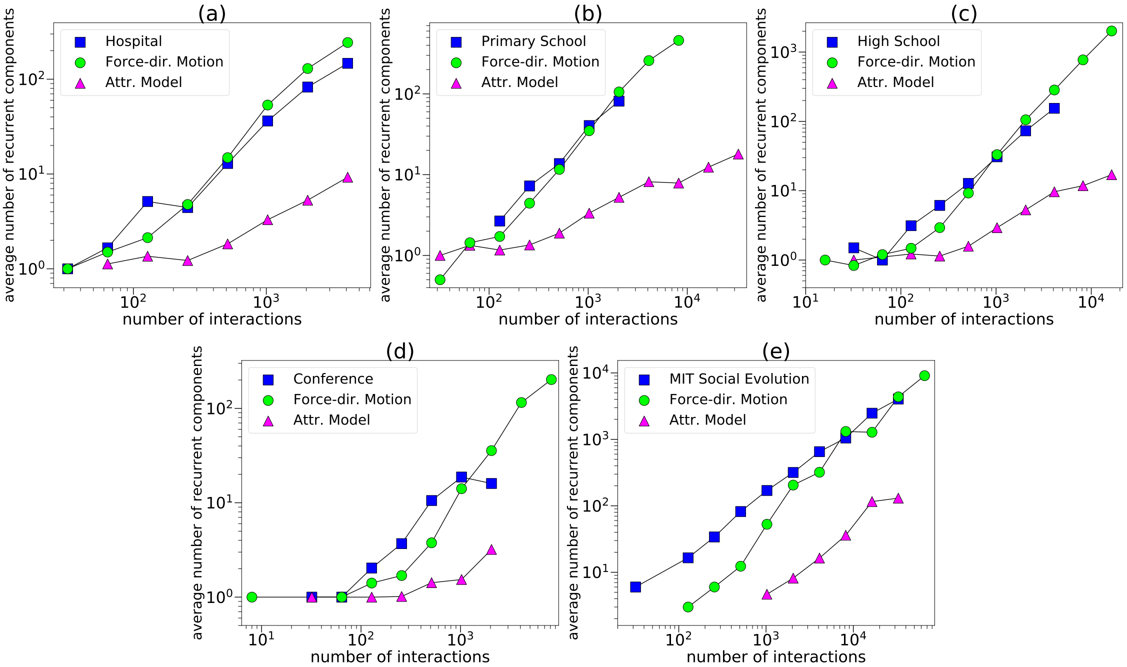

Figure 4.6: Average number of recurrent components where a node participates as a function of the total number of interactions of the node in the datasets and in simulated networks with the FDM (Force-dir. Motion) and attractiveness models. For each real dataset the corresponding simulations with the models use the dataset's parameters (Sec. 4.3). The results in (a-d) are averages over 10 simulation runs, while (e) shows results from one simulation run.

to the attractiveness model, as expected. We again note that if attraction forces are disabled in the FDM ($F_0 = 0$) the results are similar to the attractiveness model.

4.2 Other properties of real versus modeled networks

In Fig. 4.7 we compare a range of other properties between the Hospital network and the corresponding simulated networks with the FDM model. These properties are described in Sec. 3.2.

We see that the FDM can adequately capture the characteristics of the real-world networks. An exception is the distribution of the shortest time-respecting paths between the Conference and the model, where we observe a significant deviation (see Appendix A.2).

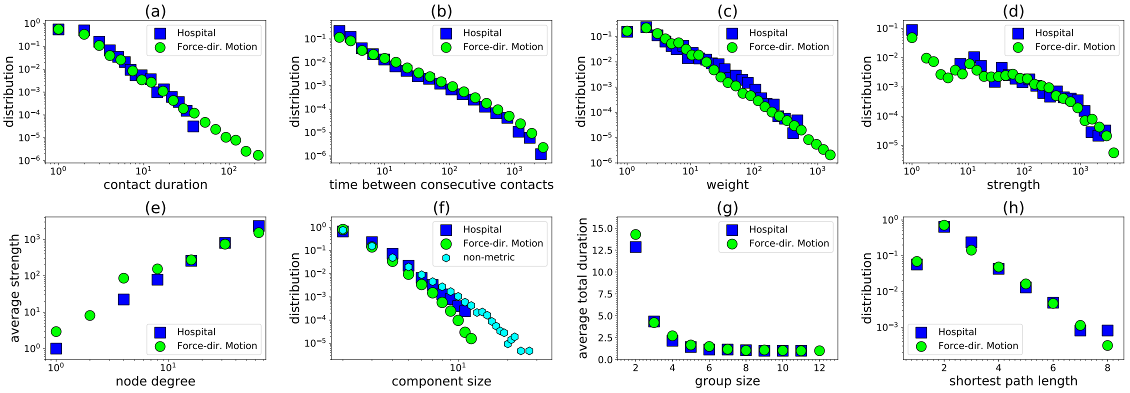

Figure 4.7: Properties of the Hospital face-to-face interaction network and of corresponding simulated networks with the FDM (Force-dir. Motion) model. **(a)** Distribution of contact duration. **(b)** Distribution of time between consecutive contacts. **(c)** Weight distribution. **(d)** Strength distribution. **(e)** Average node strength as a function of node degree. **(f)** Distribution of component sizes. **(g)** Average total interaction duration of a group as a function of its size. **(h)** Distribution of shortest time-respecting paths. In all cases the simulation results are averages over 10 runs. The distributions in (a)-(d) have been binned logarithmically; (e) also uses logarithmic binning. Plot (f) also shows the results if we randomly assign similarity distances to pairs of nodes (non-metric) instead of assigning to nodes similarity coordinates (see Sec. 4.5).

4.3 Model parameters

The FDM has the following six parameters: (i) N, which is the number of agents to simulate; (ii) τ, which is the number of time slots to simulate; (iii) L, which determines the area of the two-dimensional Euclidean space where agents move and interact (an $L \times L$ square); (iv) μ_1 in Eq. (4.1), which controls the average contact duration; and (v, vi) F_0, μ_2 in Eq. (4.4), which control the expected agent displacement due to attraction forces and the abundance and size of components (see Fig. 4.2 and the related discussion). The interaction radius d and magnitude of random displacement v are fixed to $v = d = 1$. One can fix d to any other value with $v = d$, which will result in a rescaling of the size of the Euclidean space L. We also note that the radius of the similarity space in the FDM, $R = N/2\pi$, is a dummy parameter in the sense that if R changes one can rescale μ_1, μ_2 such that the bonding and attraction forces (Eqs. (4.1), (4.4)) remain the same. Below we discuss how we tune the above parameters in the simulated counterparts of each real network—see Table 4.2 for their values.

Network	N	τ	τ_{warmup}	L	μ_1	F_0	μ_2
Hospital	70	4400	2500	95	0.8	0.12	0.9
Primary School	242	3100	2000	98	0.35	0.2	0.78
High School	327	7375	6500	295	1.2	0.11	0.86
Conference	113	7030	6000	340	2.65	0.02	3.6
MIT Social Evolution	62	60905	10000	2200	1.9	0.1	1.03

Table 4.2: FDM parameter values used in the simulated counterpart of each real network.

Parameters N, τ are set equal to the total number of agents and time slots in the real dataset. τ_{warmup} is a simulation warmup period until the average number of interacting agents per slot stabilizes. All properties of the simulated networks are measured after this period. This period is required in order to give time to the agents that are close in the similarity space to move close to each other in the Euclidean space, as agents are initially uniformly distributed in the Euclidean space. One can avoid using a warmup period by assigning to the agents initial positions in the Euclidean space not uniformly at random but from a snapshot of a previous simulation run after τ_{warmup}, along with the similarity coordinates that the agents had in the run. Figs. 4.8a,b show snapshots of the agents in the Euclidean space at times $t = 0$ and $t = \tau_{\text{warmup}} + 1$ in a simulated counterpart of the High School. Fig. 4.9b also visualizes the agents in the Euclidean space at time $t = 6108$ after τ_{warmup}, while Fig. 4.9a shows the agents in their similarity space. As expected, we see that the majority of agents that participate in interactions in the snapshot are very close to each other in the similarity space along the angular direction.

For setting L, μ_1, F_0, μ_2 we follow a two-stage procedure that consists of a parameter initialization and a parameter tuning phase. We describe these two phases below.

4.3.1 Parameter initialization

- Parameter L: We set the initial value of this parameter based on the average number of interacting agents per slot $\bar{n}$ and the total number of agents N in the dataset (see Table 4.1 for the values of $\bar{n}$). Specifically, assuming that there are no boundary effects, no inactive agents, a uniform spatial distribution of agents with density $\delta = N/L^2$, and an interaction radius $d = 1$, the expected degree of an agent is $\bar{k} \approx N\pi/L^2$. Therefore, the probability that a given agent interacts with another agent is $p_c \approx \pi/L^2$, while the probability that the agent does not interact with any other agent is $(1 - p_c)^N \approx e^{-\bar{k}}$. This means that the expected number of interacting agents per slot is $\bar{n} \approx N(1 - e^{-\bar{k}})$. Solving for L we get
$$L \approx \sqrt{-\frac{N\pi}{\ln(1 - \bar{n}/N)}}. \tag{4.5}$$

- Parameter $\mu_1 > 0$: We set the initial value of this parameter to $\mu_1 = 0.5$.

- Parameters $F_0 \geq 0$, $\mu_2 > 0$: From our experiments we observed that $0.1 \leq F_0 \leq 0.2$ with $\mu_2 = 0.8$ is a good initial configuration for these parameters.

4.3.2 Parameter tuning

We next tune the above parameters as described below in order to match the following quantities between simulated and real networks: (i) the average contact duration; (ii) the average number of recurrent components per 10 minute interval, while ensuring a similar size of the largest component formed; and (iii) the average agent degree in the time-aggregated network. We choose a 10 minute interval in (ii) in order to give some time to components to break apart (components appearing more than once in an interval are counted only once), but no more than 10 minutes to avoid losing the resolution of the components formation. While tuning a specific parameter all other parameters remain fixed, and we take the average of the corresponding metric over 10 simulation runs.

1. We tune μ_1 such that the average contact duration in simulations is approximately the same as in the real dataset. (The average contact duration increases with μ_1.)

2. We tune F_0 and μ_2 such that the average number of recurrent components per interval of 10 minutes is approximately the same as in the real dataset, while the size of the largest component formed is similar as in the dataset. (As μ_2 increases larger components form, until the agents eventually collapse into a giant connected component. At the same time, the number of components initially increases and then decreases, see Fig. 4.2(a) and Figs. 4.10a-c below. A similar behavior is observed as F_0 increases because the magnitude of the deterministic motion increases compared to the magnitude of the random motion, see Fig. 4.2(b) and Figs. 4.10d-f below. In this case, an eventual collapse into a giant component can occur if μ_2 is not sufficiently small (Fig. 4.2(b) and Figs. 4.10d-f). In general, to avoid collapses, as one of these parameters increases the other should decrease.)

3. We tune L such that the average agent degree in the time-aggregated network is approximately the same as in the real network. (Larger values of L result in a smaller average agent degree.)

4. We repeat steps (1)-(3) if needed until all considered metrics ((i)-(iii) above) are approximately the same as in the real dataset.

Finally, each agent i is also assigned an activity value a_i, which is the probability of the agent to become active at the beginning of each slot if the agent is inactive. In the simulated counterparts of the Hospital and Conference the a_is are sampled uniformly at random from $[0, 1]$. In the simulated counterparts of the Primary and High School we assign $a_i = 0.5$ for all agents i, as we have observed that the number of interactions per agent in the corresponding real datasets is somewhat more homogeneous than in the Hospital and Conference datasets. We also assign $a_i = 0.5$ for every agent i in the simulations of the MIT Social Evolution. We note that after tuning the model parameters as described, the resulting average number of interacting agents and links per slot are also similar as in the real networks ($\bar{n}, \bar{l}$ in Table 4.1).

In the attractiveness model [100] we also have $v = d = 1$ and the free parameters are the number of agents N, the number of time slots τ, and the size of the space L, while the a_is are sampled uniformly at random from $[0, 1]$. In our simulations with this model N and τ are equal to their counterparts in the real networks, while L is set such that the average number of interacting agents per slot is approximately the same as in the real networks. Specifically, the values of L for the simulated networks of the Hospital, Primary School, High School, Conference and MIT Social Evolution are $L = 44, 50, 80, 85, 45$, respectively. A warmup period is not required.

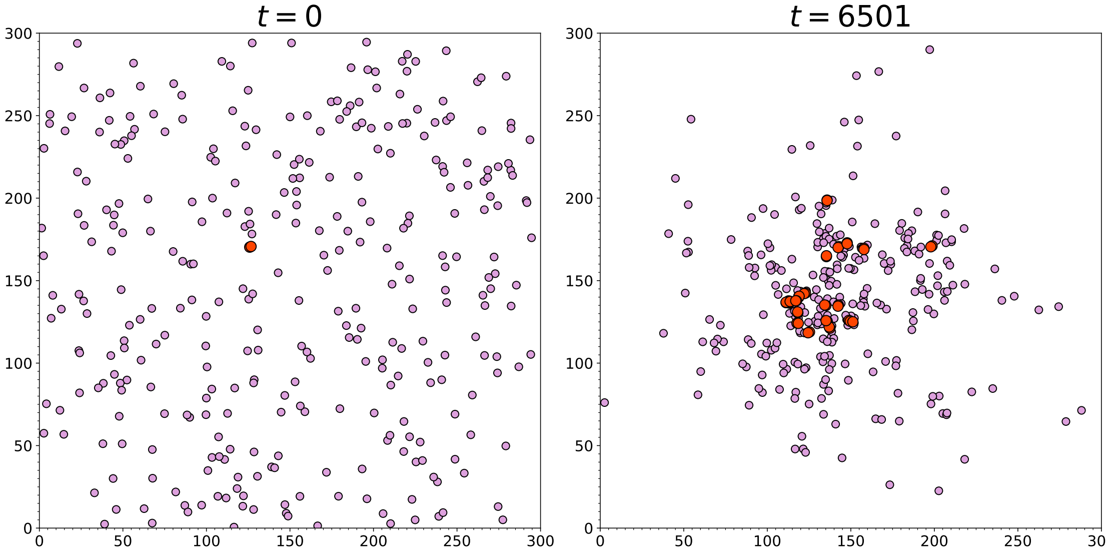

Figure 4.8: Distribution of the agents in the Euclidean space at $t = 0$ and $t = \tau_{\text{warmup}} + 1 = 6501$ in a simulated counterpart of the High School. Agents engaged in interactions are shown by red circles while the rest of the agents are shown by light purple circles.

4.4 Agent displacement

In this section we analyze the expected agent displacement in the FDM, $E[\Delta a]$. Let $\mathcal{S}(t)$ be the set of moving and interacting agents in slot t and $F_i^{t,x}, F_i^{t,y}$ the total attractive forces exerted to a moving agent i by all agents $j \in \mathcal{S}(t)$ along the x and y directions of the motion,

$$F_i^{t,x} = \sum_{j \in \mathcal{S}(t)} F_{ij} \cos \psi_{ij}^t, \quad \cos \psi_{ij}^t = \frac{(x_j^t - x_i^t)}{\sqrt{(x_j^t - x_i^t)^2 + (y_j^t - y_i^t)^2}}, \tag{4.6}$$

$$F_i^{t,y} = \sum_{j \in \mathcal{S}(t)} F_{ij} \sin \psi_{ij}^t, \quad \sin \psi_{ij}^t = \frac{(y_j^t - y_i^t)}{\sqrt{(x_j^t - x_i^t)^2 + (y_j^t - y_i^t)^2}}, \tag{4.7}$$

where $F_{ij} = F_0 e^{-\frac{s_{ij}}{\mu_2}}$ and $s_{ij} = N(\pi - |\pi - |\theta_i - \theta_j||)/2\pi$ is the similarity distance between i and j. Since the similarity coordinates are uniformly distributed we can set without loss of generality $\theta_i = 0$, and compute the second moment of F_{ij},

$$E[F_{ij}^2] = \frac{1}{2\pi} \int_0^{2\pi} F_{ij}^2 \mathrm{d}\theta_j = \frac{F_0^2 \mu_2}{N}(1 - e^{-\frac{N}{\mu_2}}) \approx \frac{F_0^2 \mu_2}{N}, \tag{4.8}$$

where the last approximation holds for large N/μ_2. Further, assuming that ψ_{ij}^t in Eqs. (4.6), (4.7) is uniformly distributed on $[0, 2\pi]$, i.e., assuming that the agents

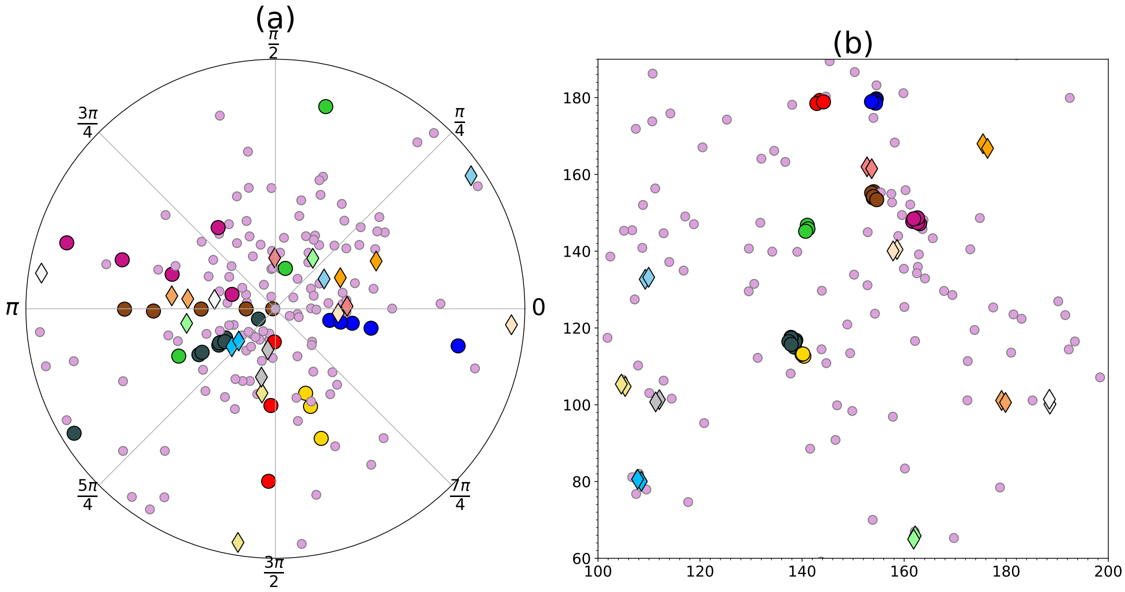

Figure 4.9: Hidden similarity and physical Euclidean space of the agents. **(a)** Similarity space of the agents in a simulated counterpart of the High School. The agents are colored, sized and marked as in the snapshot of their temporal network in (b) and placed according to their angular (similarity) coordinates. For visualization purposes, the agents also have radial coordinates assigned using the formula $r_i = \mathcal{R} - \log I_i$, where I_i is the total number of interactions of agent i in the simulation, while $\mathcal{R} = \log \max_i\{I_i\}$ is the radius of the circle. **(b)** Snapshot of the agents in the Euclidean space at time slot $t = 6108$ (total slots = 7375). The diamonds represent interactions involving only 2 agents, while the bigger circles represent interactions between at least 3 agents. The smallest grayed out circles are the moving agents that are not interacting, while inactive agents are not shown.

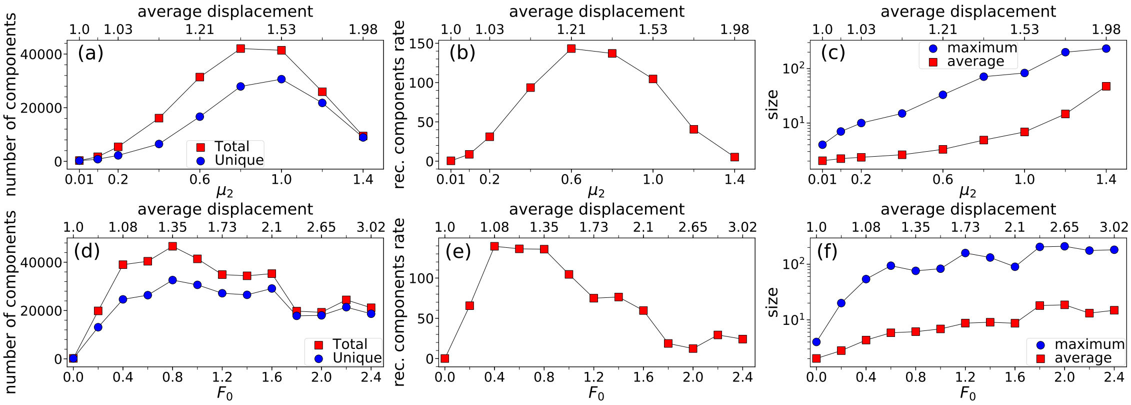

Figure 4.10: Formation and size of components as a function of μ_2 (top row) and F_0 (bottom row). In all cases $v = d = 1$ and the top x-axis indicates the corresponding average agent displacement. **(a, d)** Number of components of at least three agents. The squares show the number of all components formed (unique and recurrent), while the circles show the number of unique components. The recurrent components are measured within intervals of 10 minutes as described in the text. Each time slot is assumed to be 20 seconds as in the real face-to-face interaction networks, and thus the interval of 10 minutes corresponds to 30 time slots in the simulation. The recurrent components are also measured this way in Fig. 4.2. **(b, e)** Corresponding recurrent components rate, i.e., the average number of recurrent components observed in an interval of 10 minutes. **(c, f)** Maximum and average size across all components formed (including components of size 2). Other simulation parameters are $N = 242, \tau = 3100, \tau_{\text{warmup}} = 2000, L = 98, \mu_1 = 0.35$ and the activation probability a_i for each agent i is sampled uniformly at random from $[0, 1]$; these parameter values are also used in Fig. 4.2. In (a-c) $F_0 = 1$, while in (d-f) $\mu_2 = 1$.

$j \in \mathcal{S}(t)$ are uniformly distributed around agent i in the Euclidean space, we have $E[\cos\psi_{ij}^t] = E[\sin\psi_{ij}^t] = 0$, and

$$E[(\cos\psi_{ij}^t)^2] = E[(\sin\psi_{ij}^t)^2] = \frac{1}{2\pi}\int_0^{2\pi}(\sin\psi_{ij}^t)^2 \mathrm{d}\psi_{ij}^t = \frac{1}{2}. \tag{4.9}$$

Using Eqs. (4.6)-(4.9) we can write

$$E[(F_i^{t,x})^2|\mathcal{S}(t)] = E[(F_i^{t,y})^2|\mathcal{S}(t)] = E\left[\left(\sum_{j\in\mathcal{S}(t)} F_{ij}\sin\psi_{ij}^t\right)^2\right] =$$

$$\sum_{j\in\mathcal{S}(t)} E[F_{ij}^2]E[(\sin\psi_{ij}^t)^2] \approx \frac{F_0^2\mu_2}{2}\frac{|\mathcal{S}(t)|}{N}. \tag{4.10}$$

The above relation depends only on the number of moving and interacting agents, $|\mathcal{S}(t)|$, and not on the exact agent i or the agents $j \in \mathcal{S}(t)$. Furthermore, the average number of moving and interacting agents per slot is $\bar{a}N + (1-\bar{a})\bar{n}$, where $\bar{a}$ is the average agent activation probability [2]. Therefore, removing the condition on the index i and slot t we can write

$$E[(F^x)^2] = E[(F^y)^2] \approx \frac{F_0^2\mu_2}{2}\left(\bar{a} + (1-\bar{a})\frac{\bar{n}}{N}\right). \tag{4.11}$$

Now, from Eqs. 4.2, 4.3, the expected displacement of an agent i in slot t, $E[\Delta a_i^t|\mathcal{S}(t)]$, is

$$\begin{aligned}
E[\Delta a_i^t|\mathcal{S}(t)] &= E\left[\sqrt{(x_i^{t+1}-x_i^t)^2 + (y_i^{t+1}-y_i^t)^2}|\mathcal{S}(t)\right] \\
&= E\left[\sqrt{\left(F_i^{t,x}+R_i^x\right)^2 + \left(F_i^{t,y}+R_i^y\right)^2}|\mathcal{S}(t)\right] \\
&\leq \sqrt{E\left[\left(F_i^{t,x}+R_i^x\right)^2|\mathcal{S}(t)\right] + E\left[\left(F_i^{t,y}+R_i^y\right)^2|\mathcal{S}(t)\right]} \\
&= \sqrt{2E[(F_i^{t,x})^2|\mathcal{S}(t)] + 2E[(R_i^x)^2]} = \sqrt{F_0^2\mu_2\frac{|S(t)|}{N} + v^2}.
\end{aligned} \tag{4.12}$$

As above, we can remove the condition on the index i and slot t and write

$$E[\Delta a] \leq \sqrt{F_0^2\mu_2\left(\bar{a} + (1-\bar{a})\frac{\bar{n}}{N}\right) + v^2}. \tag{4.13}$$

The inequalities in Eqs. (4.12), (4.13) hold since $E[\sqrt{x}] \leq \sqrt{E[x]}$ for $x \geq 0$ (Jensen's inequality for concave functions). Further, since $R_i^x = v\cos\phi_i$, $R_i^y = v\sin\phi_i$, where ϕ_i is sampled uniformly at random from $[0, 2\pi]$, we also use in Eq. (4.12) the facts $E[R_i^x] = E[R_i^y] = 0$ and $E[(R_i^x)^2] = E[(R_i^y)^2] = v^2/2$. Finally, we note that the magnitude of the random displacement is always $\sqrt{(R_i^x)^2 + (R_i^y)^2} = v$.

Table 4.2 shows the values of F_0, μ_2, N, while Table 4.1 shows the values of $\bar{n}$. Parameter v is fixed to $v = d = 1$ and in all of our simulated networks $\bar{a} = 1/2$ (Sec. 4.3). The corresponding average displacement per slot in the simulations of

[2] Given that the expected number of interacting agents per slot is $\bar{n}$ and that we activate each of the $N - \bar{n}$ inactive agents with an average probability $\bar{a}$, the expected number of moving and interacting agents per slot is $\bar{n} + (N-\bar{n})\bar{a} = \bar{a}N + (1-\bar{a})\bar{n}$.

the Hospital, Primary School, High School, Conference and MIT Social Evolution is $1.00356, 1.0090, 1.0023, 1.0005, 1.00271$ while the corresponding upper bounds predicted by Eq. (4.13) are $1.00359, 1.0096, 1.0029, 1.0004, 1.00278$. We can see that the random displacement is significantly larger than the displacement due to the attraction forces. Specifically, with the Hospital, Primary School, High School, Conference and MIT Social Evolution parameters we have $F_0\sqrt{\mu_2\left(\bar{a}+(1-\bar{a})\frac{\bar{n}}{N}\right)} = 0.084, 0.139, 0.077, 0.028, 0.075$ vs. $v = 1$. We note that in general a natural choice for the total expected displacement is to be in the order of the interaction range d. This will give the chance to escaping agents to move away from their interactions in one time slot, without drifting far away.

4.5 Similarity forces in spaces with broken triangle inequality

The key metric property of the similarity space, i.e., the triangle inequality, ensures that if an agent a is close to an agent b and b is close to a third agent c, then c is also close to a. This means that the forces between all the three agents are strong and these agents will tend to gather close to each other in the Euclidean space forming triangle abc. In other words, the triangle inequality in the similarity space imposes a localization effect on the forces, which attract similar agents to form clusters in the observed network. If the forces decrease fast enough with the similarity distance, then we indeed expect to see an abundance of small connected components as in the real datasets and the model (Figs. 4.7f and Appendix A.2). On the other hand, if the similarity distances do not satisfy the triangle inequality, then agents a and c might not be close to each other, but instead close to some other agents d and e, forming chain $dabce$ in the observed network. That is, if the similarity space does not have a metric structure, forces loose their localization, and agents tend to form larger components.

To verify these arguments we break the triangle inequality in the similarity space by assigning similarity distances sampled uniformly from $[0, \pi R]$ to all pairs of agents (non-metric case), instead of assigning to the agents similarity coordinates on the circle (metric case). We see in Figs. 4.7f and Appendix A.2 that indeed in the non-metric case larger components form even though the values of the simulation parameters are set exactly as in the metric case.

Furthermore, in Fig. 4.11 we consider simulation runs with the FDM and the Primary School parameters in Table 4.2, except that in Figs. 4.11a-d we gradually increase μ_2 from 0.1 to 1, while in Figs. 4.11e-h we gradually increase F_0 from 0.1 to 1 with $\mu_2 = 0.4$. We see that in the non-metric case, as μ_2 or F_0 increases, the average number of interacting agents per slot increases much faster than in the metric case (Figs. 4.11a,e). This is also the case for the average agent degree in the time-aggregated network (Figs. 4.11b,f) and the size of the largest component (Figs. 4.11c,g). We also see that in the non-metric case neither μ_2 nor F_0 can increase the average number of recurrent components per interval of 10 minutes beyond a certain value (Figs. 4.11d,h), as most agents collapse into a giant connected component. Specifically, in the non-metric case agents start forming significantly larger and larger components after $\mu_2 = 0.5$ and $F_0 = 0.2$ (Figs 4.11c,g). For $\mu_2 \geq 0.7$ and $F_0 \geq 0.4$ the agents collapse into a giant component in the middle of the Euclidean space (cf. Fig. 4.12) and remain collapsed until the end of the simulation. By contrast, in the metric case the corresponding increases in Figs. 4.11a-c,e-g are much more gradual and agents do not collapse into a giant component. Furthermore, the average number of recurrent components per interval of 10 minutes increases with μ_2 and F_0 (Figs. 4.11d,h). Therefore, the metric structure of the similarity space promotes the formation of sparse network snapshots without giant connected components, as in the real networks. However, this would be a limitation of the model if one wishes to model a real network where giant connected components or sufficiently large connected components form. Tuning parameters F_0 and μ_2 to form large connected components, approximating the size of a giant connected component, would cause a collapse of the agents in the middle of the Euclidean space. The agents would remain collapsed and the other properties reproduced by the model would break down.

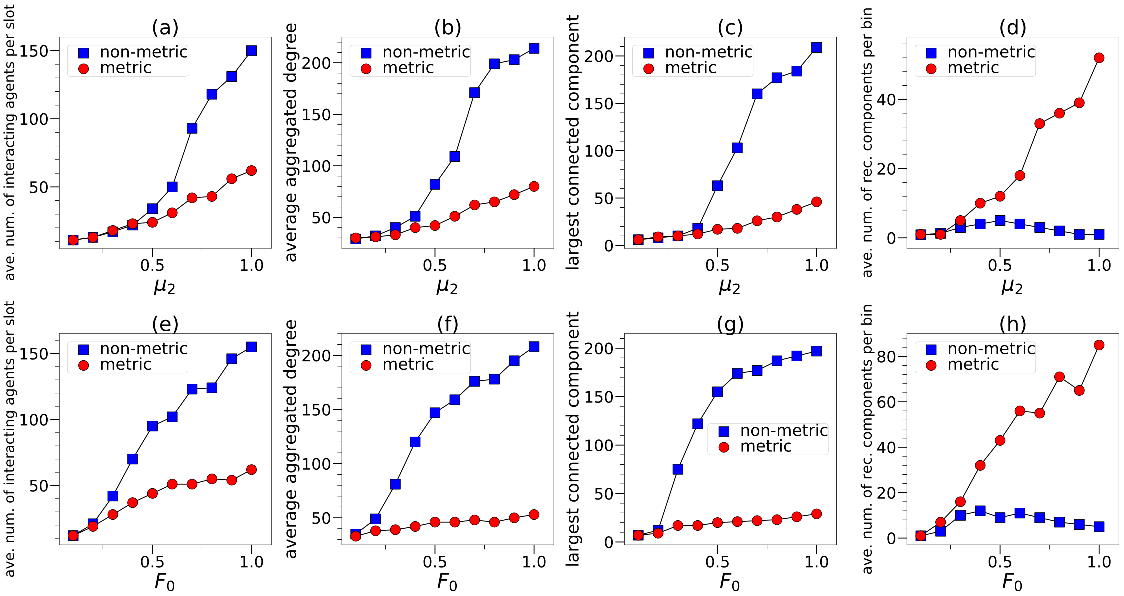

Figure 4.11: Similarity spaces with metric vs. non-metric structure. **(a, e)** Average number of interacting agents per slot in FDM simulated networks with and without metric structure in the similarity space, as a function of μ_2 and F_0. In (a) $F_0 = 0.2$ and in (e) $\mu_2 = 0.4$. **(b, f)** Average agent degree in the time-aggregated network of contacts for the networks in (a, e). **(c, g)** Size of the largest component formed in the networks of (a, e). **(d, h)** Average number of recurrent components per interval (bin) of 10 minutes in the networks of (a, e). Each point in the plots is an average over 10 simulation runs.

4.6 Non-uniform similarity coordinates

In this section we consider a non-uniform distribution of the similarity coordinates corresponding to the organization of agents into communities. To this end, we sample the angular coordinates of nodes from a Gaussian mixture distribution (GMD) as in [69]. The GMD is a mixture of multiple Gaussian distribution components, characterized by the following parameters [67, 69]: (i) $C > 0$, which is the number of components, each one representative of a community; (ii) $\mu_{1...C} \in [0, 2\pi]$, which are the means of every component, representing the central locations of the communities in the angular space; (iii) $\sigma_{1...C} > 0$, which are the standard deviations of every

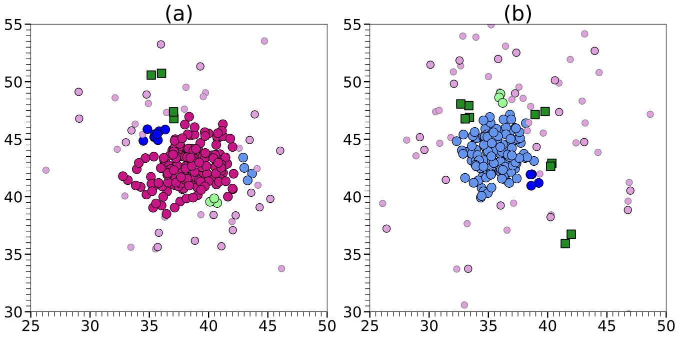

Figure 4.12: Snapshots of collapsing agents in the middle of the Euclidean space. **(a)** Snapshot of the agents in the Euclidean space at time slot $t = 2677$ (total slots $= 3100$) in a simulated network of Fig. 4.11 with non-metric similarity space and $\mu_2 = 0.8, F_0 = 0.2$. **(b)** Same as (a) but for $F_0 = 0.7, \mu_2 = 0.4$ at slot $t = 2981$. In both (a, b) the smallest grayed-out circles are inactive agents, the second smallest circles are moving agents that are not interacting, the dark green squares are connected components of 2 agents, and the largest multicolored circles are connected components of least 3 agents. In (a) the agents in the giant connected component are colored purple, while in (b) they are colored light blue.

component, determining how much the communities are spread in the angular space; and (iv) $\rho_{1\ldots C}$ ($\sum_i \rho_i = 1$), which are the mixing proportions of every component, determining the relative sizes of the communities.

We consider simulations of the Primary School. Since in the Primary school students are divided into 10 classes [102], we assume that there are 10 communities and sample the angular coordinates of the agents from a GMD with parameters $C = 10$, $\mu_i = 2\pi(i-1)/C$, $\sigma_i = 2\pi/(8C)$, $\rho_i = 1/C$, $i = 1\ldots C$. Fig. 4.13(b) visualizes the distribution of the agents' coordinates and juxtaposes it against the uniform distribution (Fig. 4.13(a)). We can see that the agents are divided into 10 distinct communities in the similarity space with each community having a similar number of agents.

We tune the model parameters L, μ_1, F_0, μ_2 as described in Sec. 4.3, obtaining: $L = 83$, $\mu_1 = 0.1$, $F_0 = 0.2$, $\mu_2 = 0.32$. The rest of the parameters are as in the simulations of the Primary School in Sec. 4.3. In Figs. 4.14, 4.15 we see that the results are very similar to the ones in Appendix. A.1, A.2 where the similarity coordinates were uniformly distributed. In other words, the organization of agents into communities does not affect the results.

4.7 Hyperbolic space considerations

Hyperbolic spaces appear as the most natural geometric spaces underlying the observed topologies of traditional complex networks, whose degree distributions are heterogeneous [55]. In addition to similarity coordinates θs, nodes in these spaces also have popularity coordinates r, and the hidden distance between two nodes is not just the angular distance $R\Delta\theta$ but the effective distance $\chi = R\Delta\theta/(\kappa\kappa')$, where κ, κ' are the expected degrees of the nodes, $\kappa \sim e^{-r}$ [55, 79].

One can replace the angular distances $s_{ij} = R\Delta\theta_{ij}$ with effective distances $\chi_{ij} = s_{ij}/(\kappa_i\kappa_j)$ in the bonding and attractive forces of the FDM (Eqs. 4.1, 4.4). However, in all datasets we considered the distribution of κs was in general quite homogeneous to justify the need for this description—see Fig. 4.16, where the expected degree κ of each agent is its average degree per time slot. Indeed, in

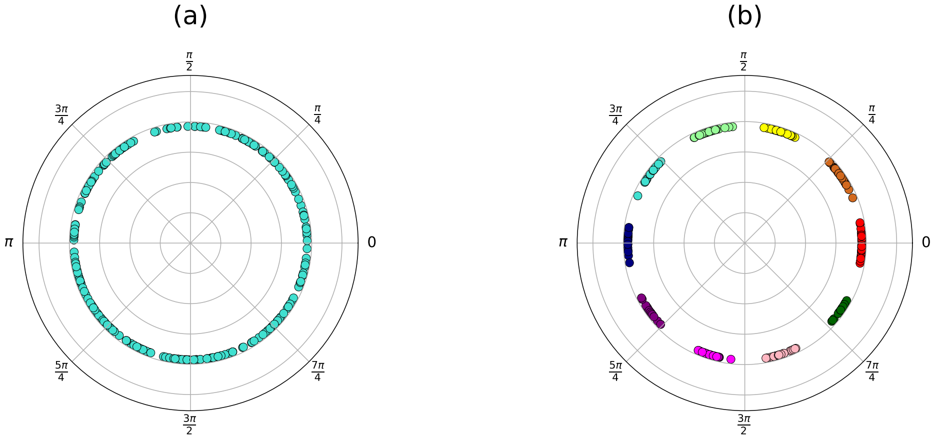

Figure 4.13: **(a)** Uniform distribution of the similarity coordinates. **(b)** Non-uniform distribution of the similarity coordinates corresponding to the separation of agents into 10 communities, each indicated by a different color.

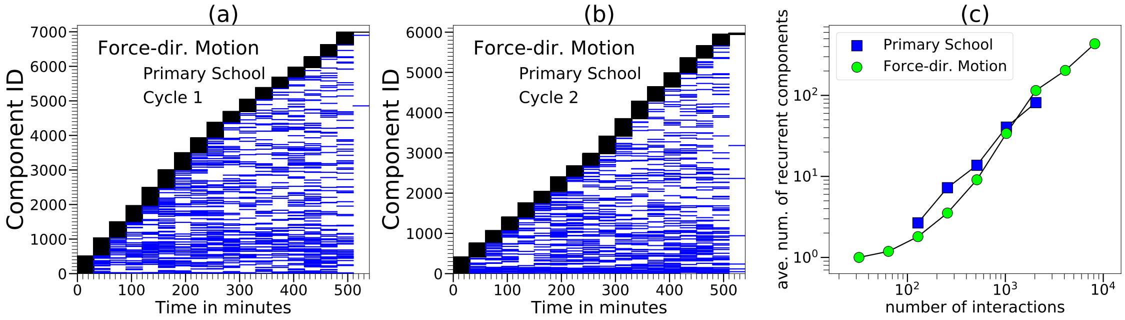

Figure 4.14: **(a, b)** Unique and recurrent components found in a simulation run of the FDM (Force-dir. Motion) model with the non-uniform similarity coordinates in Fig. 4.13(b), assuming activity cycles of the same durations as in the Primary School. **(c)** Average number of recurrent components where a node participates as a function of the total number of interactions of the node in the Primary School and in simulated networks with the non-uniform similarity coordinates in Fig. 4.13(b). The result in (c) is an average over 10 simulation runs.

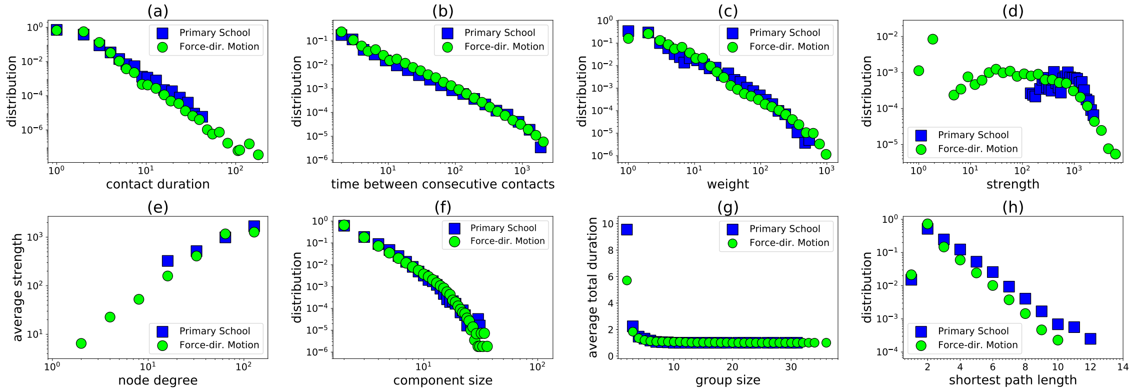

Figure 4.15: Properties of the Primary School face-to-face interaction network and of corresponding simulated networks with the FDM (Force-dir. Motion) model with the non-uniform similarity coordinates in Fig. 4.13(b). In all cases the simulation results are averages over 10 runs.

Figs. 4.17, 4.18 we see that if we assign to agents the estimated κs from the real Hospital data and use effective distances in the FDM, we obtain very similar results as in Secs. 4.1, 4.2 where we use only angular distances. For the simulations in

Figs. 4.17, 4.18 we tune again the model parameters L, μ_1, F_0, μ_2 as described in Sec. 4.3, see Table 4.3 for their values. The rest of the simulation parameters are as in Sec. 4.3.

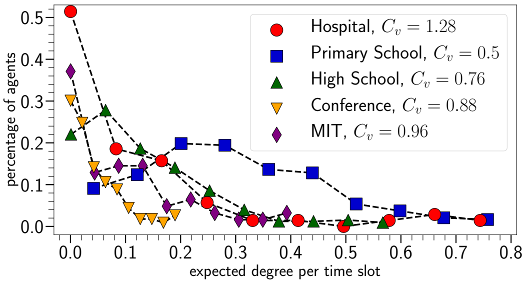

Figure 4.16: Distribution of the expected agent degree per time slot in the real data. The C_vs in the legend indicate the coefficient of variation (ratio of the standard deviation to the mean) of each distribution.

Network	τ_{warmup}	L	μ_1	F_0	μ_2
Hospital	2500	95	29	0.12	33
Primary School	2000	105	2.7	0.2	6.1
High School	6500	240	23	0.2	16
Conference	1800	75	145	0.05	85

Table 4.3: Parameter values in the simulations with the FDM model that uses effective distances.

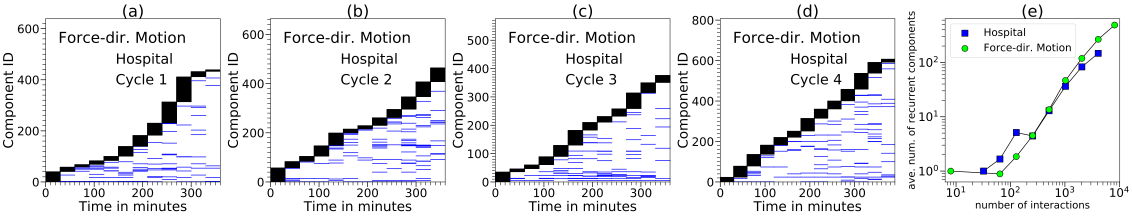

Figure 4.17: **(a-d)** Unique and recurrent components found in a simulation run of the FDM (Force-dir. Motion) model that uses effective distances, in activity cycles of the same durations as in the Hospital. **(e)** Average number of recurrent components where a node participates as a function of the total number of interactions of the node in the Hospital and in simulated networks with the FDM that uses effective distances. The result in (e) is an average over 10 simulation runs.

In Appendix A.4, we show results for the other real networks from SocioPatterns considered in this chapter.

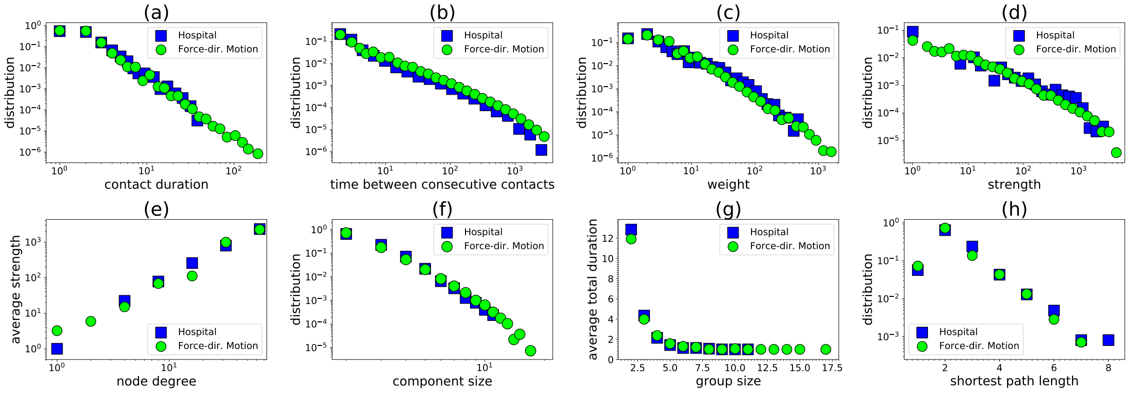

Figure 4.18: Properties of the Hospital face-to-face interaction network and of corresponding simulated networks with the FDM (Force-dir. Motion) model that uses effective distances. The results are averages over 10 simulation runs.

Chapter 5

Latent geometry and dynamics of proximity networks

This chapter has been published, with some modifications, in "Physical Review E" [78].

Understanding the time-varying proximity patterns among humans in a physical space is important in various contexts. These include the analysis and containment of spreading phenomena, like respiratory transmitted diseases, the design of routing algorithms for mobile networks, and the understanding of social relationships and influence [1, 9, 16, 24, 41, 42, 44, 48]. To this end, proximity networks have been captured in different environments [1, 16, 24, 36, 45, 64, 102, 107]. Each snapshot in these networks corresponds to an observation interval, which typically spans a few seconds to several minutes depending on the devices used to collect the data. The agents (nodes) in each snapshot are individuals and an edge between two agents means that they are within proximity range.

At the finest granularity level an edge between two agents represents a close-range face-to-face proximity (up to 1.5 m, detected using wearable sensors). Such networks have been captured over the period of few days or weeks in different closed settings, such as hospitals, schools, scientific conferences and workplaces [36, 45, 64, 102, 107]. The main motivation for obtaining these data has emerged in epidemiological studies of infectious diseases. Other proximity networks have been captured for longer periods of time (months) and over larger areas, such as university campuses, using Bluetooth sensing or WiFi tracking [1, 16, 24]. These methods yield information only on proximity at a range, e.g., up to 10 m using Bluetooth devices and up to 40 m or more using WiFi tracking [1, 24, 39]. Thus, proximity in these networks does not imply face-to-face interaction. The collection of these data has been motivated by research in mobile networking [16, 44, 48] and social studies [1, 24].

Irrespectively of the context, measurement period, and measurement method, different proximity networks have been shown to exhibit similar statistical properties [9, 16, 48, 101]. The most widely studied properties are the aggregated—obtained by considering the samples from all pairs of nodes together—distributions of contact and inter-contact durations. The former is the distribution of time that a pair of nodes spends in contact, i.e., remains within proximity range, while the latter is the distribution of time separating two contacts between the same pair of nodes. These metrics are important in determining the capacity and delay of a network, and the dynamics of spreading processes [20, 34, 60, 96, 108]. It has been found that both of these distributions are broad in real data and compatible with power laws, $P(t) \propto t^{-\gamma}$, with or without exponential cutoffs [16, 44, 48, 101]. Studies have reported exponents $\gamma \geq 2$ for contact durations [19, 92] and $\gamma \in (1, 2)$ for inter-contact durations [16, 29, 44, 104]. Further, it has been shown that aggregated power laws can emerge from pairwise distributions that are either power-laws, exponentials or log-normals, with the latter two better fitting most pairwise inter-contact durations in real data [18, 31, 81]. Another property of interest is the distribution of the total duration of contacts between two agents throughout the observation period, called weight distribution [34, 101, 109]. The aggregated weight distribution is also roughly compatible with power laws [101], while an exponent $\gamma = 1.4$ has been reported for this distribution in the contact network of high school students [29].

These and other distinctive features of real proximity networks can be well reproduced by minimal models of mobile interacting agents [87, 100, 101]. Minimal models, i.e., models that reproduce many of the observed properties under minimal assumptions, are crucial for generating realistic synthetic networks and understanding the mechanisms that are responsible for the observed behaviors. In particular, the Force-Directed Motion (FDM) model presented in Chapter 4, utilizes the idea of a latent metric space where the agents reside, and where the distance d between two agents abstracts their *similarity*. Attractive forces that decrease exponentially with the similarity distance direct the agents' motion towards other agents in the physical space, and determine the duration of their interactions. One can also consider the *effective distance* between two agents, $\chi = d/(\kappa\kappa')$, where κ and κ' are the agents' expected degrees per snapshot, abstracting their *popularity* [79]. In this case, dissimilar agents can still be attracted by strong forces if their popularities are high. The FDM casts the problem of modeling proximity networks as an N-body problem akin to molecular dynamics [93]. However, mathematically proving the properties of generated networks by the FDM is not straightforward, and the model has been so far studied only in simulations.

The FDM has been inspired by the $\mathbb{S}^1$ model of traditional (non-mobile) complex networks [55, 95]. In the $\mathbb{S}^1$, nodes are also separated by effective distances χ, and are connected with the *Fermi-Dirac* connection probability $p(\chi) = 1/(1 + \chi^{1/T})$, where $T \in (0, 1)$ is the *network temperature*, controlling clustering [25] in the network. The $\mathbb{S}^1$ is isomorphic to hyperbolic geometric graphs [55]. It can generate network snapshots that possess many of the common structural properties of real networks, including heterogeneous or homogeneous degree distributions, strong clustering, and the small-world property [55, 79, 95]. Fig. 5.1 shows the probability that two agents are connected in a snapshot of FDM-simulated networks as a function of their effective distance. Interestingly, we see that this probability resembles qualitatively the Fermi-Dirac connection probability in the $\mathbb{S}^1$ model, even though this form of connection probability is *not* enforced into the FDM. Specifically, we see in Fig. 5.1 that the connection probability in the FDM has a smooth step-like form, where connection probabilities at small distances are orders of magnitude larger than connection probabilities at large distances.

Motivated by the observation in Fig. 5.1, here we consider a simple latent space model for human proximity networks, where each snapshot is a realization of the $\mathbb{S}^1$ model. We call this model *dynamic*-$\mathbb{S}^1$ and show that it *simultaneously* reproduces many of the observed properties of real systems. The dynamic-$\mathbb{S}^1$ does not model node mobility directly, but captures the connectivity in each snapshot. By forgoing the motion component it facilitates mathematical analysis, allowing us to prove the contact, inter-contact and weight distributions. We show that these distributions are power laws in the thermodynamic limit, with exponents $2 + T$,

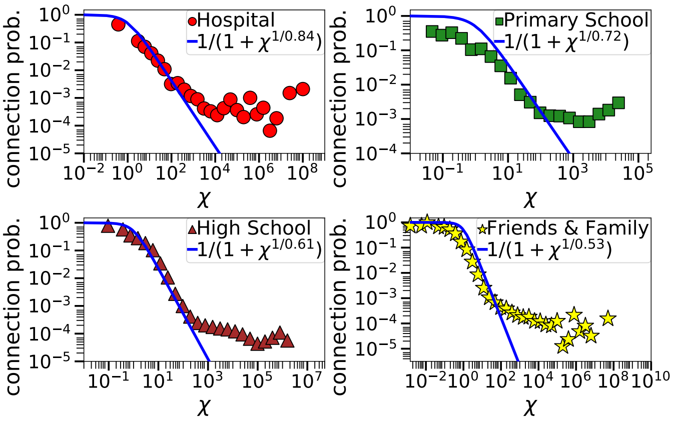

Figure 5.1: Probability that two agents are connected in a snapshot as a function of their effective distance χ in FDM-simulated counterparts of the hospital, primary school and high school face-to-face interaction networks [64, 102, 107]; and of the Friends & Family proximity network [1]. The simulations are performed as in Chapter 4, while the connection probabilities are computed excluding agents that are inactive in each snapshot. The solid lines are Fermi-Dirac connection probabilities with temperatures $T = 0.84, 0.72, 0.61, 0.53$, corresponding respectively to the temperatures of the hospital, primary school, high school and Friends & Family (Sec. 5.3.2).

$2 - T$ and $1 + T$, respectively, where $T \in (0, 1)$ is the temperature in the Fermi-Dirac connection probability. These exponents are within the ranges observed in real systems. We also show that temperature controls the agents' time-aggregated degrees and the formation of unique and recurrent components. Additionally, we consider paradigmatic epidemic and rumor spreading processes [22, 51] and find that they perform remarkably similar in real and modeled networks.

The rest of the chapter is organized as follows. In Sec. 5.1 we review the $\mathbb{S}^1$ model. In Sec. 5.2 we introduce the dynamic-$\mathbb{S}^1$. In Sec. 5.3 we juxtapose the properties of modeled and real networks. In Sec. 5.4 we compare the performance of epidemic and rumor spreading processes running on them. In Sec. 5.5 we mathematically analyze the main properties of the model. In Sec. 5.6 we elucidate the crucial role of temperature in the formation of components. Finally, in Sec. 5.8 we conclude this chapter.

5.1 $\mathbb{S}^1$ model

In the $\mathbb{S}^1$ model [55] each node has latent (or hidden) variables κ, θ. The latent variable κ is proportional to the node's expected degree in the resulting network. The latent variable θ is the angular similarity coordinate of the node on a circle of radius $R = N/2\pi$, where N is the total number of nodes. To construct a network with the model that has size N, average node degree $\bar{k}$, and temperature $T \in (0, 1)$, we perform the following steps:

(1) coordinate assignment: for each node $i = 1, 2, \ldots, N$, sample its angular coordinate θ_i uniformly at random from $[0, 2\pi]$, and its degree variable κ_i from a probability density function (PDF) $\rho(\kappa)$;

(2) creation of edges: connect every pair of nodes i, j with the Fermi-Dirac

connection probability

$$p(\chi_{ij}) = \frac{1}{1+\chi_{ij}^{1/T}}. \tag{5.1}$$

In the last expression, χ_{ij} is the effective distance between nodes i and j,

$$\chi_{ij} = \frac{R\Delta\theta_{ij}}{\mu\kappa_i\kappa_j}, \tag{5.2}$$

where $\Delta\theta_{ij} = \pi - |\pi - |\theta_i - \theta_j||$. Parameter μ in (5.2) is derived from the condition that the expected degree in the network is indeed $\bar{k}$, yielding

$$\mu = \frac{\bar{k}\sin(T\pi)}{2\bar{\kappa}^2 T\pi}, \tag{5.3}$$

where $\bar{\kappa} = \int \kappa\rho(\kappa)\mathrm{d}\kappa$. The expected degree of a node with latent variable κ is [55]

$$\bar{k}(\kappa) = \frac{\bar{k}}{\bar{\kappa}}\kappa. \tag{5.4}$$

For sparse networks ($\bar{k} \ll N$) the resulting degree distribution $P(k)$ has a similar functional form as $\rho(\kappa)$ [13]. For instance, a power law degree distribution with exponent $\gamma > 2$ is obtained if $\rho(\kappa) \propto \kappa^{-\gamma}$, while a Poisson degree distribution with mean $\bar{k}$ is obtained if $\rho(\kappa) = \delta(\kappa - \bar{k})$, where $\delta(x)$ is the Dirac delta function [13, 95]. Smaller values of the temperature T favor connections at smaller effective distances and increase the average clustering [25] in the network, which is maximized at $T = 0$, and nearly linearly decreases to zero with $T \in [0, 1)$. At $T \to 0$ the connection probability in (5.1) becomes the step function $p(\chi_{ij}) \to 1$ if $\chi_{ij} < 1$, and $p(\chi_{ij}) \to 0$ if $\chi_{ij} > 1$.

5.2 Dynamic-$\mathbb{S}^1$

The dynamic-$\mathbb{S}^1$ models a sequence of network snapshots, G_t, $t = 1, \ldots, \tau$, where τ is the total number of time slots. Each snapshot is a realization of the $\mathbb{S}^1$ model. Therefore, there are N agents that are assigned latent variables κ, θ as in the $\mathbb{S}^1$ model, which remain fixed in all time slots. The temperature T is also fixed, while each snapshot G_t is allowed to have a different average degree $\bar{k}_t$. Thus, the model parameters are $N, \tau, \rho(\kappa)$, T, and $\bar{k}_t, t = 1, \ldots, \tau$. The snapshots are generated according to the following simple rules:

(1) at each time step $t = 1, \ldots, \tau$, snapshot G_t starts with N disconnected nodes, while $\bar{k}$ in Eq. (5.3) is set equal to $\bar{k}_t$;

(2) each pair of nodes i, j connects with probability given by Eq. (5.1);

(3) at time $t+1$, all the edges in snapshot G_t are deleted and the process starts over again to generate snapshot G_{t+1}.

We note that the snapshots are conditionally independent given the agents' latent variables $\kappa_1, \theta_1, \ldots, \kappa_N, \theta_N$, but *not* independent. In other words, even though each snapshot G_t is constructed anew, there are correlations among the snapshots that are induced by the nodes' effective distances χ_{ij}. In particular, nodes at smaller effective distances have higher chances of being connected in each snapshot, as dictated by the connection probability in (5.1). Fig. 5.2 provides a visualization of snapshots generated by the model, where we see that agents at smaller similarity distances tend to stay connected in consecutive time slots and form recurrent components. Next, we compare the properties of synthetic networks generated by the model and real networks.

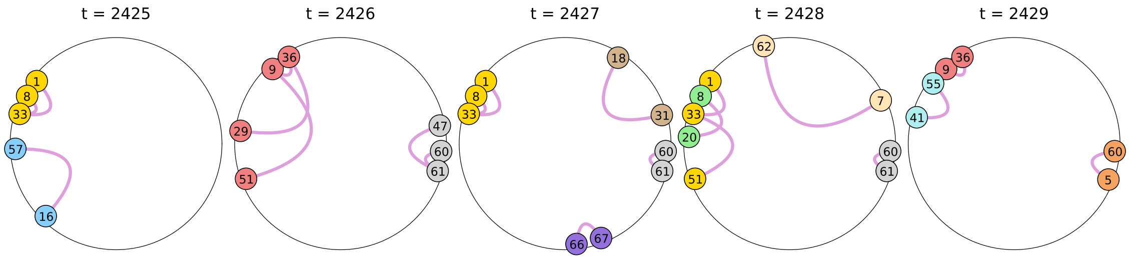

Figure 5.2: Snapshots from the simulated counterpart of the hospital face-to-face interaction network generated by the dynamic-$\mathbb{S}^1$ (Sec. 5.3). The snapshots correspond to time slots $t = 2425$-2429. Each snapshot shows the interacting agents in their similarity space and the connections between them. The agents are colored according to the connected component where they belong, while the non-interacting agents in each snapshot, i.e., the agents with zero degree, are not shown to avoid clutter. The contact duration between agents 60 and 61 is three slots (2426-2428), while the inter-contact duration between agents 9 and 36 is two slots (2427, 2428). Agents 1, 8 and 33 belong to a component forming both at $t = 2425$ and $t = 2427$ (recurrent component).

5.3 Modeled vs. real networks

5.3.1 Overview of real networks

We consider four face-to-face interaction networks from SocioPatterns [97], which correspond to: (i) a hospital ward in Lyon [107]; (ii) a primary school in Lyon [102]; (iii) a high school in Marseilles [64]; and (iv) a scientific conference in Turin [45]. These networks were captured over a period of 5, 2, 5 and 2.5 days, respectively. Each of their snapshots corresponds to a time slot of 20 sec. We also consider the Bluetooth-based proximity network of the members of a residential community adjacent to a research university in North America, taken from the Friends and Family dataset [1]. The snapshots here correspond to slots of 5 min, spanning the period October 2010 to May 2011. In all cases we number the slots and assign node IDs sequentially, $t = 1, 2, \ldots, \tau$ and $i = 1, 2, \ldots, N$. Table 5.1 gives an overview of the data.

We define the average degree per slot of agent i as

$$\bar{d}_i = \frac{1}{\tau} \sum_{t=1}^{\tau} d_{i,t}, \tag{5.5}$$

where $d_{i,t} \geq 0$ is agent's i degree in slot t, while the average agent (snapshot) degree in slot t is

$$\bar{k}_t = \frac{1}{N} \sum_{i=1}^{N} d_{i,t}. \tag{5.6}$$

Network	N	τ	$\bar{n}$	$\bar{d}$	$\bar{k}_{\text{aggr}}$
Hospital	75	17376	2.9	0.05	30
Primary school	242	5846	30	0.18	69
High school	327	18179	17	0.06	36
Conference	113	10618	3.3	0.03	39
Friends & Family	131	57961	52	1.1	97

Table 5.1: Overview of the considered real networks. N is the number of agents; τ is the total number of time slots; $\bar{n}$ is the average number of interacting agents per slot; $\bar{d}$ is the average agent degree per slot; and $\bar{k}_{\text{aggr}}$ is the average degree in the time-aggregated network (defined in Sec. 5.3.3). Average values above 10 have been rounded to the nearest integer.

Fig. 5.3 shows the distribution of $\bar{d}_i$ and $\bar{k}_t$ in the considered networks. The average agent degree per slot is

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N}\bar{d}_i = \frac{1}{\tau}\sum_{t=1}^{\tau}\bar{k}_t. \tag{5.7}$$

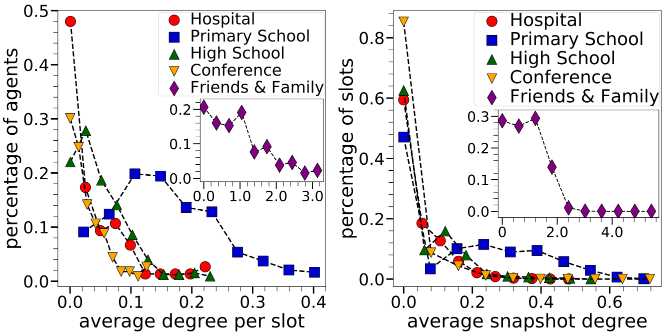

Figure 5.3: Distribution of the average agent degree per slot (left) and of the average snapshot degree (right) in the considered networks.

5.3.2 Modeled networks

For each real network we construct its synthetic counterpart using the dynamic-$\mathbb{S}^1$. Each counterpart has the same number of nodes N and duration τ as the corresponding real network, while the latent variable κ_i of each agent $i = 1, \ldots, N$ is set equal to the agent's average degree per slot in the real network,

$$\kappa_i = \bar{d}_i. \tag{5.8}$$

Thus, the distribution of κ_i is the corresponding empirical distribution in Fig. 5.3 (left). The target average degree $\bar{k}_t$ in each snapshot G_t, $t = 1, \ldots, \tau$, is set equal to the average degree in the corresponding real snapshot at slot t. Finally, the temperature T is set such that the resulting average time-aggregated degree, $\bar{k}_{\text{aggr}}$, is similar to the one in the real network—we analyze the dependence of $\bar{k}_{\text{aggr}}$ on T in Sec. 5.5.4.

In the counterparts, the expected degree of agent i in slot t is [Eq. (5.4)]

$$\bar{k}_t(\kappa_i) = \frac{\bar{k}_t}{\bar{d}}\kappa_i, \tag{5.9}$$

while agent's i expected degree per slot is $\sum_{t=1}^{\tau} \bar{k}_t(\kappa_i)/\tau = \kappa_i$. The counterparts aim at capturing the variability in the number of interacting agents per slot since the probability that an agent i interacts with at least one other agent in slot t is

$$I_{i,t} = 1 - \left[1 - \frac{\bar{k}_t(\kappa_i)}{N-1}\right]^{N-1}, \tag{5.10}$$

while $\bar{k}_t(\kappa_i) \propto \bar{k}_t \kappa_i$.

5.3.3 Properties of modeled vs. real networks

Table 5.2 gives an overview of the counterparts. We see that their characteristics are overall very similar to the ones of the real networks (Table 5.1). Further, Fig. 5.4 shows that the counterparts indeed capture the variability in the number of interacting agents per slot.

Modeled network	N	τ	$\bar{n}$	$\bar{d}$	$\bar{k}_{\text{aggr}}$	T
Hospital	75	17376	2.5	0.04	30	0.84
Primary school	242	5846	33	0.17	69	0.72
High school	327	18179	18	0.06	35	0.61
Conference	113	10618	2.9	0.03	30	0.85
Friends & Family	131	57961	67	1.1	96	0.53

Table 5.2: Modeled counterparts. The values of $\bar{n}$, $\bar{d}$ and $\bar{k}_{\text{aggr}}$ are averages over 20 simulation runs except from the Friends & Family where the averages are over 5 runs. Average values above 10 have been rounded to the nearest integer.

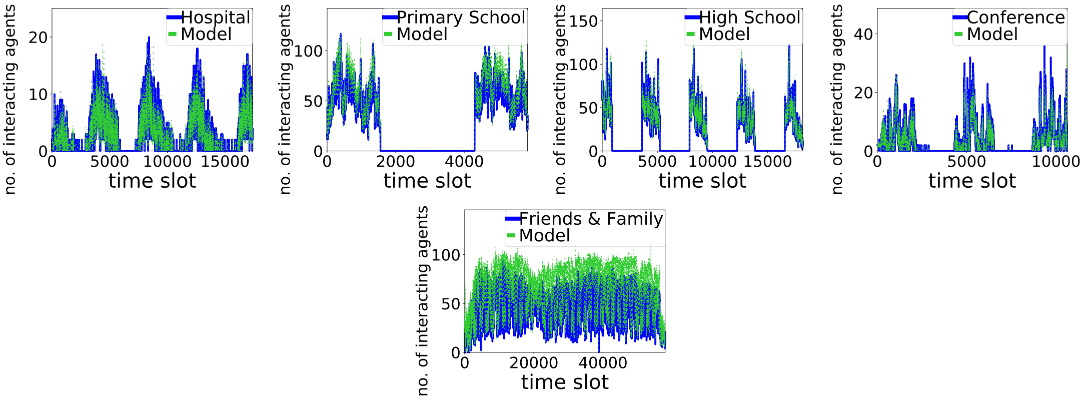

Figure 5.4: Number of interacting agents per slot in real and modeled networks. In the first four plots the cycles of activity, i.e., the periods with high numbers of interacting agents, correspond to the consecutive observation days where the agents were present in the corresponding premises (5, 2, 5 and 2.5 days, respectively.) There is a single activity cycle in the last plot, spanning the whole observation period—proximity in the Friends & Family was constantly captured using mobile phones.

In Figs. 5.5 and 5.6 we compare a range of other properties between real and modeled networks, considered also in [99, 100] and in Chapter 4, Sec. 4.1.3.

Figs. 5.5 and 5.6 show that the dynamic-$\mathbb{S}^1$ reproduces all the properties considered remarkably well. A main exception are the longer paths in the conference [Fig. 5.5(f)], which can not be captured by the model. We also note that $\bar{k}_{\text{aggr}}$ in conference's counterpart could not exceed ≈ 30 (vs. 39 in the real network). Thus, the dynamic-$\mathbb{S}^1$ does not totally capture the characteristics of this network. Interestingly, this was also the case with the FDM (Appendix A.2). Finally, we note that the ability of the model to capture the properties of the considered networks is not due to mere calibration of expected node degrees. In Appendix B.2, we show that the configuration model [17, 80] with the same calibration of expected node degrees, Eqs. (5.8, 5.9), cannot reproduce the abundance of recurrent components, nor the broad contact, inter-contact and weight distributions observed in the real systems. Further, in Sec. 5.5 we prove these distributions in the dynamic-$\mathbb{S}^1$ and show that they do not depend on the distribution of the degree variables $\rho(\kappa)$. Below, we also investigate the *pairwise* contact and inter-contact distributions in modeled and real networks.

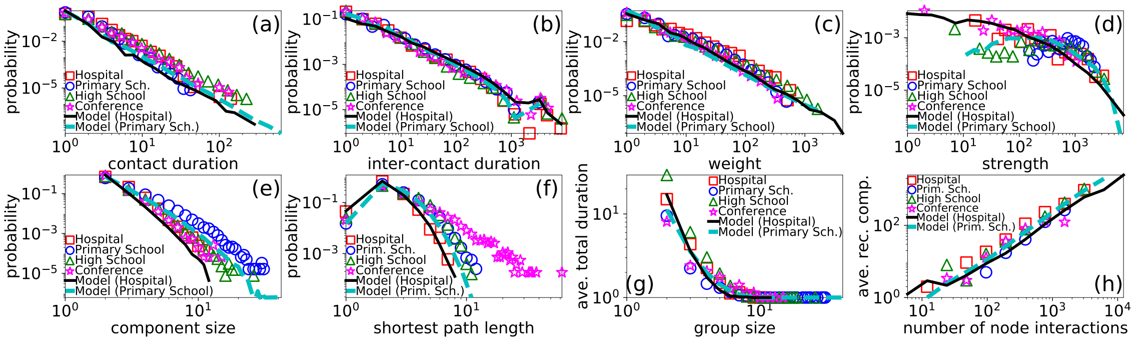

Figure 5.5: Real face-to-face interaction networks vs. simulated networks with the dynamic-$\mathbb{S}^1$. **(a)** Contact distribution. **(b)** Inter-contact distribution. **(c)** Weight distribution. **(d)** Strength distribution. **(e)** Distribution of component sizes. **(f)** Distribution of shortest time-respecting path lengths. **(g)** Average total duration of a group as a function of its size. **(h)** Average number of recurrent components where an agent participates as a function of the total number of interactions of the agent. The results with the model are averages over 20 simulation runs and correspond to the counterparts of the hospital and primary school. Similar results hold for the rest of the counterparts, not shown to avoid clutter. The probabilities in (a)-(f) represent relative frequencies, i.e., they are computed as $n_i / \sum_j n_j$, where n_i is the number of samples that have value i. (a)-(d) have been binned logarithmically. Durations are measured in numbers of time slots.

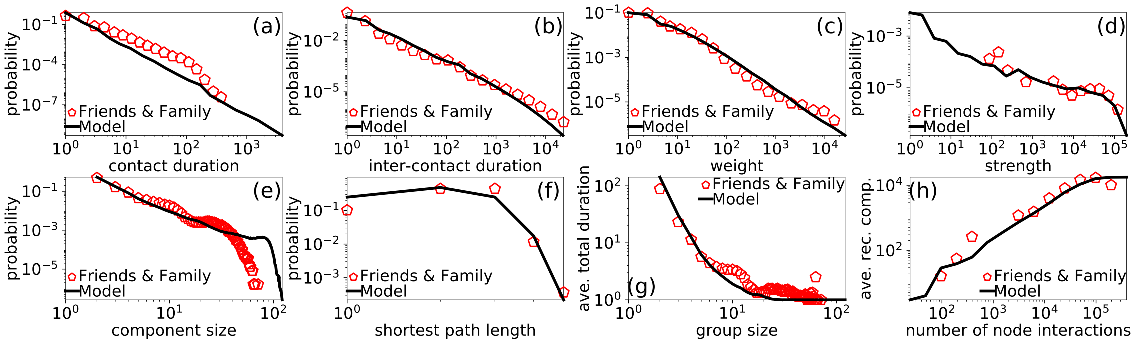

Figure 5.6: Same as Fig. 5.5 but for the Friends & Family proximity network and its modeled counterpart. The results with the model are averages over 5 simulation runs.

5.3.4 Pairwise contact and inter-contact distributions

If the expected snapshot degrees, $\bar{k}_t, t = 1, \ldots, \tau$, are independent and identically distributed, the pairwise contact and inter-contact distributions in the dynamic-$\mathbb{S}^1$

are geometric at $\tau \to \infty$ [1]. Indeed, in this case the probability for two nodes i, j with latent variables κ_i, κ_j and angular distance $\Delta\theta_{ij}$ to remain connected for $t = 1, 2, \ldots$ slots, is

$$P_{\rm c}(t; \kappa_i, \kappa_j, \Delta\theta_{ij}) = \bar{p}_{ij}^{t-1} \left(1 - \bar{p}_{ij}\right), \qquad (5.11)$$
$$\bar{p}_{ij} \equiv \int p[\chi_{ij}(\bar{k})] f(\bar{k}) \mathrm{d}\bar{k},$$

where $p[\cdot]$ is the connection probability in Eq. (5.1), while $\chi_{ij}(\bar{k})$ is the effective distance between the two nodes, which depends on the average snapshot degree $\bar{k}$ [Eqs. (5.2, 5.3)], whose PDF is denoted by $f(\cdot)$. Similarly, the probability that the two nodes remain disconnected for $t = 1, 2, \ldots$ slots, is

$$P_{\rm ic}(t; \kappa_i, \kappa_j, \Delta\theta_{ij}) = \left(1 - \bar{p}_{ij}\right)^{t-1} \bar{p}_{ij}. \qquad (5.12)$$

In general, these distributions are not geometric in the model as they depend on the stochastic process that describes the time evolution of the expected snapshot degrees.

Previous studies have reported that a significant portion of pairwise inter-contact durations in real data can be fitted with exponential distributions [18, 31]. Since the geometric distribution is the discrete analogue of the exponential distribution, these studies are in line with Eq. (5.12). Given these results, we check below how well the geometric distribution captures the pairwise contact and inter-contact distributions in the considered real systems and their modeled counterparts.

For each pair of nodes we consider the sets of its contact and inter-contact durations in each of the activity cycles shown in Fig. 5.4. We consider sets with at least three distinct duration values. For each set we estimate the parameter of the geometric distribution, i.e., the success probability $p = 1/m$, where m is the mean of the durations in the set. Then, we draw the same number of samples as the number of durations in the set from a geometric distribution with parameter p. Subsequently, we use the two-sample Kolmogorov-Smirnov (KS) goodness of fit test [6, 63] to test the hypothesis that the values in the set and the sampled values have the same distribution. We recall that such a statistical test can only *reject* or *fail to reject* a given hypothesis for a given significance level α. This level corresponds to the probability of incorrectly rejecting the hypothesis, while if the test fails to reject the hypothesis, we only know that this is true to a confidence level $1 - \alpha$. We use $\alpha = 0.01$, and find for each activity cycle the percentage of pairs for which the test failed to reject the hypothesis. Table 5.3 shows the average of this percentage across the activity cycles in each network, averaged across ten repetitions of the above procedure. The results for each counterpart are also averaged across ten different temporal network realizations.

We see in Table 5.3 that the geometric distribution fits a high percentage of contact durations in both modeled and real networks. It also fits a high percentage of inter-contact durations in modeled networks, and a significant percentage of inter-contact durations in the real systems, which however is not as high as in the modeled networks. These results suggest that the model captures the variability of the contact durations in the real systems. However, it does not totally capture the variability of the inter-contact durations.

[1] For finite τ they are truncated geometric.

Network	**Contact dist.**	**Inter-contact dist.**	
	geometric	**geometric**	**log-normal**
HP (model)	98%	97%	99%
HP (real)	97%	69%	100%
PS (model)	100%	100%	99%
PS (real)	98%	69%	100%
HS (model)	98%	98%	98%
HS (real)	94%	65%	100%
CF (model)	95%	92%	99%
CF (real)	97%	64%	100%
F & F (model)	80%	85%	87%
F & F (real)	77%	60%	78%

Table 5.3: Percentage of pairs (rounded to the nearest integer) where the KS test failed to reject the hypothesis that their contact/inter-contact distribution is geometric. The table also shows the results where a log-normal distribution is assumed for the inter-contact durations; samples from the log-normal are rounded to the nearest integer before applying the KS test. (HP: Hospital; PS: Primary school; HS: High school; CF: Conference; F & F: Friends and Family.)

To verify the last statement we also consider a log-normal distribution for the inter-contact durations, which offers a more versatile model to capture the variability in the distributions [18]. We recall that the PDF of the log-normal is $f(x) = 1/(x\sigma\sqrt{2\pi})e^{-(\ln x-\mu)^2/(2\sigma^2)}$, while its skewness is $(e^{\sigma^2}+2)\sqrt{e^{\sigma^2}-1}$. For each pair of nodes, the parameters μ and σ^2 are the mean and variance of the logarithms of its inter-contact durations. We see in Table 5.3 that the log-normal better fits the inter-contact durations, especially in the real systems, as also observed in [18]. Further, Fig. 5.7 shows that the inter-contact distributions in the real networks are indeed more skewed on average than in their counterparts. Nevertheless, the aggregated inter-contact distributions are very similar in real and synthetic systems [Figs. 5.5(b), 5.6(b)]. In the next section we also see that paradigmatic dynamical processes perform similarly in the two.

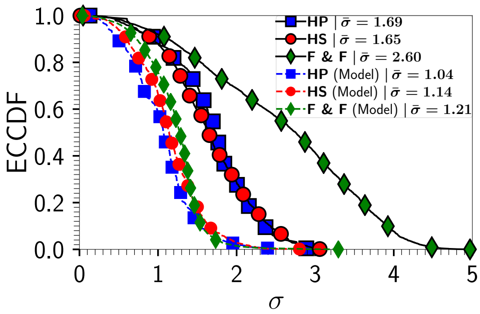

Figure 5.7: Empirical complementary cumulative distribution function (ECCDF) of the estimated log-normal's σ in real and modeled networks. The average ($\bar{\sigma}$) of each distribution is indicated in the legend.

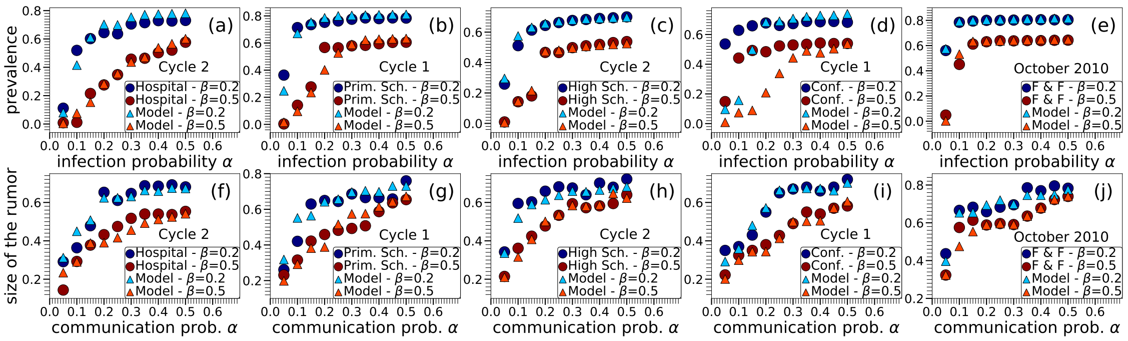

Figure 5.8: Performance of the SIS and DK processes in real and modeled networks. Top row: prevalence of the SIS process as a function of the infection probability α for two recovery probabilities β. Bottom row: size of the rumor in the DK process as a function of the probability to communicate the rumor α for two stifling probabilities β. The results are averages over ten runs of each process in the activity cycles indicated in the plots. Each run of the SIS/DK process starts with a random set of infected/spreader agents that consists of 10% of agents. The results for the modeled counterparts are also averaged across ten different temporal network realizations.

5.4 Dynamical processes on modeled vs. real networks

We consider the susceptible-infected-susceptible (SIS) epidemic spreading model [51] and the DK (Daley and Kendall) model for rumor spreading [22]. In the SIS each agent can be in one of two states, susceptible (S) or infected (I). At any time slot an infected agent recovers with probability β and becomes susceptible again, whereas infected agents infect the susceptible agents with whom they interact with probability α. Thus, the transition of states is S $\rightarrow$ I $\rightarrow$ S. In the DK model each agent can be in one of three states, ignorant (I), spreader (S) or stifler (R). An ignorant agent that interacts with a spreader receives the rumor with probability α and becomes a spreader, while a spreader that interacts with another spreader or a stifler becomes a stifler with probability β and no longer communicates the rumor. The transition of states is I $\rightarrow$ S $\rightarrow$ R.

To simulate the SIS process on temporal networks we use the dynamic SIS implementation of the Network Diffusion Library [90]. We have also modified this library to implement the DK model. For the SIS process we consider the average percentage of infected agents per slot (prevalence), while for the DK process we consider the percentage of stiflers at the final slot (size of the rumor). Fig. 5.8 shows that the two processes perform remarkably similar in real and modeled networks. The only exception is in the performance of the SIS in the conference and its counterpart at low infection probabilities [Fig. 5.8(d)]—a similar behavior has been observed in the FDM (Appendix A.3) and it may be due to the fact that the models do not totally capture the characteristics of this network, as noted in Sec. 5.3.3.

5.5 Mathematical analysis

Here we perform a detailed mathematical analysis of the main properties of the dynamic-$\mathbb{S}^1$. To facilitate the analysis, we assume that the expected snapshot degree is the same in all time slots, $\bar{k}_t = \bar{k}$, $\forall t$. This assumption renders the connection probability between two nodes [Eq. (5.1)] the same in all slots. However, we illustrate that the analytical results match closely the simulation results from the modeled counterparts of real systems, where this assumption does not hold.

We show that for sparse snapshots, $\bar{k} \ll N$, and large durations τ, the aggregated contact, inter-contact and weight distributions can be approximated by power laws with exponents $2+T$, $2-T$ and $1+T$, respectively, where $T \in (0,1)$ is the temperature in the connection probability. Technically, we consider these distributions in the thermodynamic limit, $N \to \infty$, and show that they are power-laws with the aforementioned exponents at $\tau \to \infty$. Interestingly, these results do not depend on the distribution of the latent degree variables $\rho(\kappa)$. Further, we analyze the expected degree in the time-aggregated network, and show that in finite networks the expected strength of a node grows super-linearly with its time-aggregated degree, as empirically observed in prior studies [100, 101]. We begin with the contact distribution.

5.5.1 Aggregated contact distribution

The probability $r_{\mathrm{c}}(t;\kappa_i,\kappa_j,\Delta\theta_{ij})$ to observe a sequence of exactly $t = 1, 2, \ldots, \tau-2$ consecutive slots where two nodes i, j with latent variables κ_i, κ_j and angular distance $\Delta\theta_{ij}$ are connected, is the percentage of time τ where we observe a slot where these two nodes are not connected, followed by t slots where they are connected, followed by a slot where they are not connected [2]. For each duration t, there are $\tau - t - 1$ possibilities where this duration can be realized. For instance, if $t = 2$ the two nodes can be disconnected in slot $i-1$, connected in slots $i, i+1$, and disconnected in slot $i+2$, where $i = 2, \ldots, \tau-2$. Therefore, the percentage of observation time where a duration of t slots can be realized is $(\tau-t-1)/\tau$. Since the two nodes are connected in each slot with probability $p(\chi_{ij})$ with χ_{ij} in Eq. (5.2), we have

$$r_{\mathrm{c}}(t;\kappa_i,\kappa_j,\Delta\theta_{ij}) = \left(\frac{\tau-t-1}{\tau}\right) p(\chi_{ij})^t [1-p(\chi_{ij})]^2. \tag{5.13}$$

Removing the condition on $\Delta\theta_{ij}$, which is uniform on $[0,\pi]$, yields

$$\begin{aligned} r_{\mathrm{c}}(t;\kappa_i,\kappa_j) &= \left(\frac{\tau-t-1}{\tau}\right) \frac{1}{\pi} \int_0^{\pi} p(\chi_{ij})^t [1-p(\chi_{ij})]^2 \mathrm{d}\Delta\theta_{ij} \\ &= \left(\frac{\tau-t-1}{\tau}\right) \frac{2\mu\kappa_i\kappa_j}{N} \\ &\times \int_0^{\frac{N}{2\mu\kappa_i\kappa_j}} p(\chi_{ij})^t [1-p(\chi_{ij})]^2 \mathrm{d}\chi_{ij} \\ &= \left(\frac{\tau-t-1}{\tau}\right) \left(\frac{N}{2\mu\kappa_i\kappa_j}\right)^{2/T} \left(\frac{T}{2+T}\right) \\ &\times {}_2F_1\left[t+2, 2+T, 3+T, -\left(\frac{N}{2\mu\kappa_i\kappa_j}\right)^{1/T}\right], \end{aligned} \tag{5.14}$$

[2] For brevity we ignore the cases where the first/last of the slots that two nodes can be connected starts/ends at the beginning/end of the observation period.

where ${}_2F_1[a, b, c; z]$ is the Gauss hypergeometric function [72]. At $N \to \infty$, the integral in (5.14) simplifies for $T \in (0, 1)$ and $t \geq 1$, to

$$\int_0^\infty p(\chi_{ij})^t[1 - p(\chi_{ij})]^2 \mathrm{d}\chi_{ij} = \frac{T\Gamma(2+T)\Gamma(t-T)}{\Gamma(t+2)}, \tag{5.15}$$

where $\Gamma(z)$ is the complete gamma function, $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\mathrm{d}x,\ z > 0$ [3]. From (5.14, 5.15), we have

$$\begin{aligned} Nr_\mathrm{c}(t; \kappa_i, \kappa_j) \xrightarrow{N\to\infty} & \left(\frac{\tau - t - 1}{\tau}\right) 2\mu\kappa_i\kappa_j \\ & \times \frac{T\Gamma(2+T)\Gamma(t-T)}{\Gamma(t+2)}. \end{aligned} \tag{5.16}$$

Removing the condition on κ_i and κ_j, gives

$$\begin{aligned} Nr_\mathrm{c}(t) &= N \int\int r_c(t; \kappa_i, \kappa_j)\rho(\kappa_i)\rho(\kappa_j)\mathrm{d}\kappa_i\mathrm{d}\kappa_j \\ & \xrightarrow{N\to\infty} \left(\frac{\tau - t - 1}{\tau}\right) \frac{2\mu\bar{\kappa}^2 T\Gamma(2+T)\Gamma(t-T)}{\Gamma(t+2)}. \end{aligned} \tag{5.17}$$

The aggregated contact distribution, $P_\mathrm{c}(t)$, is the probability that two nodes are connected for exactly t consecutive slots given that $t \geq 1$,

$$P_\mathrm{c}(t) = \frac{r_\mathrm{c}(t)}{\sum_{t=1}^{\tau-2} r_\mathrm{c}(t)}. \tag{5.18}$$

From (5.17, 5.18), we have

$$P_\mathrm{c}(t) \xrightarrow{N\to\infty} \frac{(\tau - t - 1)}{g(\tau)} \frac{\Gamma(t-T)}{\Gamma(t+2)} \approx \frac{(\tau - t - 1)}{g(\tau)} \frac{1}{t^{2+T}}, \tag{5.19}$$

where

$$g(\tau) \equiv \frac{[(\tau - 1)T - 1]\,\Gamma(1-T)}{T + T^2} + \frac{\Gamma(\tau - T)}{(T + T^2)\Gamma(\tau)}.$$

The approximation in (5.19) uses the facts $\Gamma(t - T) \approx t^{-T}\Gamma(t)$ and $\Gamma(t + 2) \approx t^2\Gamma(t)$, which hold for $t \gg 1$. We see from (5.19) that for $t \ll \tau$, $P_\mathrm{c}(t)$ is approximately a power law with exponent $2 + T$. At $\tau \to \infty$, we have a pure power law

$$P_\mathrm{c}(t) \xrightarrow[\tau\to\infty]{N\to\infty} \frac{1+T}{\Gamma(1-T)} \frac{\Gamma(t-T)}{\Gamma(t+2)} \approx \frac{1+T}{\Gamma(1-T)} \frac{1}{t^{2+T}}. \tag{5.20}$$

Fig. 5.9 shows that (5.20) provides an excellent approximation to simulation results.

From (5.19), the expected contact duration in the thermodynamic limit is

$$\begin{aligned} \bar{t}_\mathrm{c} \xrightarrow{N\to\infty} & \sum_{t=1}^{\tau-2} t \frac{(\tau - t - 1)}{g(\tau)} \frac{\Gamma(t-T)}{\Gamma(t+2)} \\ &= \frac{\Gamma(2-T)\Gamma(\tau+1) - \Gamma(\tau-T)[(1+T)\tau - 2T]}{\Gamma(2-T)[(\tau-1)T - 1]\Gamma(\tau) + \Gamma(\tau-T)(1-T)}. \end{aligned} \tag{5.21}$$

At $\tau \to \infty$, the last relation simplifies to

$$\bar{t}_\mathrm{c} \xrightarrow[\tau\to\infty]{N\to\infty} \frac{1}{T}. \tag{5.22}$$

Next, we derive the aggregated inter-contact distribution following the same steps.

[3] If z is a positive integer then $\Gamma(z) = (z - 1)!$.

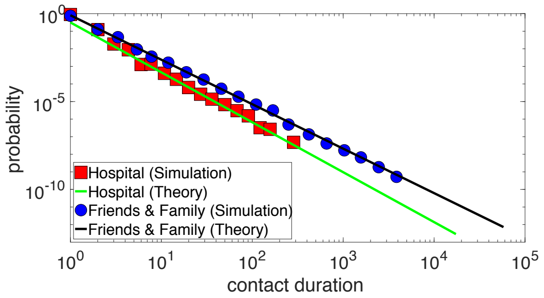

Figure 5.9: Aggregated contact distribution in the simulated counterparts of the hospital and Friends & Family (Sec. 5.3.2) vs. theoretical prediction in (5.20) with $T = 0.84, 0.53$. Similar results hold for the rest of the counterparts.

5.5.2 Aggregated inter-contact distribution

Let $r_{\text{ic}}(t; \kappa_i, \kappa_j, \Delta\theta_{ij})$ be the probability to observe a slot where two nodes i, j with latent variables κ_i, κ_j and angular distance $\Delta\theta_{ij}$ are connected, followed by t slots where they are not connected, followed by a slot where they are again connected. We have

$$r_{\text{ic}}(t; \kappa_i, \kappa_j, \Delta\theta_{ij}) = \left(\frac{\tau - t - 1}{\tau}\right) p(\chi_{ij})^2[1 - p(\chi_{ij})]^t. \tag{5.23}$$

Removing the condition on $\Delta\theta_{ij}$, yields

$$\begin{aligned} r_{\text{ic}}(t; \kappa_i, \kappa_j) &= \left(\frac{\tau - t - 1}{\tau}\right) \frac{1}{\pi} \int\limits_0^{\pi} p(\chi_{ij})^2[1 - p(\chi_{ij})]^t \mathrm{d}\Delta\theta_{ij} \\ &= \left(\frac{\tau - t - 1}{\tau}\right) \frac{2\mu\kappa_i\kappa_j}{N} \\ &\times \int\limits_0^{\frac{N}{2\mu\kappa_i\kappa_j}} p(\chi_{ij})^2[1 - p(\chi_{ij})]^t \mathrm{d}\chi_{ij} \\ &= \left(\frac{\tau - t - 1}{\tau}\right) \left(\frac{N}{2\mu\kappa_i\kappa_j}\right)^{t/T} \left(\frac{T}{t+T}\right) \\ &\times {}_2F_1\left[t + T, t + 2, t + T + 1, -\left(\frac{N}{2\mu\kappa_i\kappa_j}\right)^{1/T}\right]. \end{aligned} \tag{5.24}$$

At $N \to \infty$, the integral in (5.24) simplifies for $T \in (0, 1)$, to

$$\int\limits_0^{\infty} p(\chi_{ij})^2[1 - p(\chi_{ij})]^t \mathrm{d}\chi_{ij} = \frac{T\Gamma(2 - T)\Gamma(t + T)}{\Gamma(t + 2)}. \tag{5.25}$$

From (5.24, 5.25), and after removing the condition on κ_i and κ_j, we have

$$N r_{\text{ic}}(t) \xrightarrow{N\to\infty} \left(\frac{\tau - t - 1}{\tau}\right) \frac{2\mu\bar{\kappa}^2 T\Gamma(2 - T)\Gamma(t + T)}{\Gamma(t + 2)}. \tag{5.26}$$

The aggregated inter-contact distribution, $P_{\text{ic}}(t)$, is the probability that two nodes are disconnected for exactly t consecutive slots given that $t \geq 1$,

$$P_{\text{ic}}(t) = \frac{r_{\text{ic}}(t)}{\sum_{t=1}^{\tau-2} r_{\text{ic}}(t)}. \tag{5.27}$$

From (5.26, 5.27), we have

$$P_{\mathrm{ic}}(t) \xrightarrow{N\to\infty} \frac{(\tau-t-1)}{h(\tau)}\frac{\Gamma(t+T)}{\Gamma(t+2)} \approx \frac{(\tau-t-1)}{h(\tau)}\frac{1}{t^{2-T}}, \tag{5.28}$$

where

$$h(\tau) \equiv \frac{[(\tau-1)T+1]\Gamma(1+T)}{T-T^2} - \frac{\Gamma(\tau+T)}{(T-T^2)\Gamma(\tau)}.$$

The approximation in (5.28) holds for $t \gg 1$. For $t \ll \tau$, $P_{\mathrm{ic}}(t)$ is approximately a power law with exponent $2-T$. At $\tau \to \infty$, we have a pure power law

$$P_{\mathrm{ic}}(t) \xrightarrow[\tau\to\infty]{N\to\infty} \frac{1-T}{\Gamma(1+T)}\frac{\Gamma(t+T)}{\Gamma(t+2)} \approx \frac{1-T}{\Gamma(1+T)}\frac{1}{t^{2-T}}. \tag{5.29}$$

Fig. 5.10 juxtaposes (5.29) against simulation results.

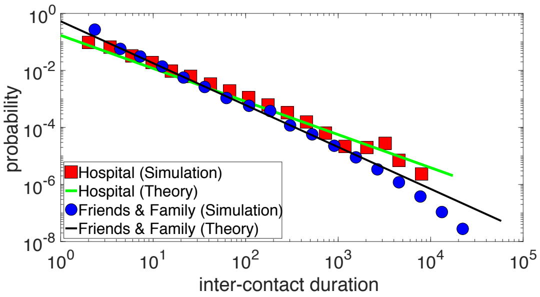

Figure 5.10: Aggregated inter-contact distribution in the simulated counterparts of the hospital and Friends & Family (Sec. 5.3.2) vs. theoretical prediction in (5.29) with $T = 0.84, 0.53$. Similar results hold for the rest of the counterparts.

From (5.28), the expected inter-contact duration in the thermodynamic limit is

$$\begin{aligned}\bar{t}_{\mathrm{ic}} \xrightarrow{N\to\infty} &\sum_{t=1}^{\tau-2} t\frac{(\tau-t-1)}{h(\tau)}\frac{\Gamma(t+T)}{\Gamma(t+2)}\\ =&\frac{\Gamma(\tau+T)[(1-T)\tau+2T]-\Gamma(2+T)\Gamma(\tau+1)}{\Gamma(2+T)[(\tau-1)T+1]\Gamma(\tau)-\Gamma(\tau+T)(1+T)}.\end{aligned} \tag{5.30}$$

The above relation increases approximately exponentially with $T \in (0,1)$, and diverges at $\tau \to \infty$,

$$\bar{t}_{\mathrm{ic}} \xrightarrow[\tau\to\infty]{N\to\infty} \infty. \tag{5.31}$$

We proceed with the weight distribution.

5.5.3 Aggregated weight distribution

The probability that two nodes i, j with latent variables κ_i, κ_j and angular distance $\Delta\theta_{ij}$ are connected in $t = 0, 1, \ldots, \tau$ slots, is given by the binomial distribution

$$r_{\mathrm{w}}(t; \kappa_i, \kappa_j, \Delta\theta_{ij}) = \binom{\tau}{t} p(\chi_{ij})^t [1-p(\chi_{ij})]^{\tau-t}. \tag{5.32}$$

Removing the condition on $\Delta\theta_{ij}$, yields

$$\begin{aligned}
r_{\mathrm{w}}(t;\kappa_i,\kappa_j) &= \frac{1}{\pi}\binom{\tau}{t}\int\limits_0^{\pi} p(\chi_{ij})^t[1-p(\chi_{ij})]^{\tau-t}\mathrm{d}\Delta\theta_{ij} \\
&= \frac{2\mu\kappa_i\kappa_j T}{N}\binom{\tau}{t}\int\limits_{u_{ij}^{\min}}^{1} u_{ij}^{t-T-1}(1-u_{ij})^{\tau-t+T-1}\mathrm{d}u_{ij} \\
&= \frac{2\mu\kappa_i\kappa_j}{N}\frac{T\Gamma(\tau+1)}{\Gamma(\tau-t+1)\Gamma(t+1)}\left[\frac{\Gamma(\tau-t+T)\Gamma(t-T)}{\Gamma(\tau)}\right. \\
&\left. -\frac{(u_{ij}^{\min})^{t-T}}{t-T}{}_2F_1(t-T,1-\tau-T+t,t-T+1,u_{ij}^{\min})\right],
\end{aligned} \tag{5.33}$$

where

$$u_{ij}^{\min} \equiv \frac{1}{1+\left(\frac{N}{2\mu\kappa_i\kappa_j}\right)^{1/T}}. \tag{5.34}$$

To reach (5.33), we perform the change of integration variable $u_{ij} \equiv p(\chi_{ij})$ and express the binomial coefficient in terms of gamma functions, $\binom{\tau}{t} = \Gamma(\tau+1)/[\Gamma(\tau-t+1)\Gamma(t+1)]$.

At $N\to\infty$, $u_{ij}^{\min}\to 0$, and the second term inside the brackets in (5.33) vanishes for $T\in(0,1)$ and $t\geq 1$. Removing the condition on κ_i and κ_j, we have

$$N r_{\mathrm{w}}(t) \xrightarrow{N\to\infty} \frac{2\mu\bar{\kappa}^2 T\tau\Gamma(\tau-t+T)\Gamma(t-T)}{\Gamma(\tau-t+1)\Gamma(t+1)}. \tag{5.35}$$

For $t=0$, we can write

$$N[1-r_{\mathrm{w}}(0)] = N\sum_{t=1}^{\tau} r_{\mathrm{w}}(t) \xrightarrow{N\to\infty} \frac{2\mu\bar{\kappa}^2\Gamma(1-T)\Gamma(\tau+T)}{\Gamma(\tau)}. \tag{5.36}$$

The aggregated weight distribution, $P_{\mathrm{w}}(t)$, is the probability that two nodes are connected in t slots given that $t\geq 1$,

$$P_{\mathrm{w}}(t) = \frac{r_{\mathrm{w}}(t)}{\sum_{t=1}^{\tau} r_{\mathrm{w}}(t)}. \tag{5.37}$$

From (5.35, 5.37), we have

$$P_{\mathrm{w}}(t) \xrightarrow{N\to\infty} \frac{1}{w(\tau)}\frac{\Gamma(\tau-t+T)\Gamma(t-T)}{\Gamma(\tau-t+1)\Gamma(t+1)} \tag{5.38}$$

$$\approx \frac{1}{w(\tau)(\tau-t)^{1-T}}\frac{1}{t^{1+T}}, \tag{5.39}$$

where

$$w(\tau) \equiv \frac{\Gamma(1-T)\Gamma(\tau+T)}{T\Gamma(\tau+1)}.$$

The approximation in (5.39) holds for $1 \ll t \ll \tau$. We see from (5.39) that for $t \ll \tau$, $P_{\mathrm{w}}(t)$ is approximately a power law with exponent $1+T$. At $\tau\to\infty$, we have a pure power law

$$P_{\mathrm{w}}(t) \xrightarrow[\tau\to\infty]{N\to\infty} \frac{T}{\Gamma(1-T)}\frac{\Gamma(t-T)}{\Gamma(t+1)} \approx \frac{T}{\Gamma(1-T)}\frac{1}{t^{1+T}}. \tag{5.40}$$

From (5.38), the expected weight in the thermodynamic limit is

$$\begin{aligned}\bar{t}_{\mathrm{w}} \xrightarrow{N\to\infty} & \sum_{t=1}^{\tau} \frac{t}{w(\tau)} \frac{\Gamma(\tau-t+T)\Gamma(t-T)}{\Gamma(\tau-t+1)\Gamma(t+1)} \\ = & \frac{\Gamma(1+T)\Gamma(\tau+1)}{\Gamma(\tau+T)} \approx \Gamma(1+T)\tau^{1-T}. \end{aligned} \tag{5.41}$$

The above relation decreases approximately exponentially with $T \in (0, 1)$, and diverges at $\tau \to \infty$,

$$\bar{t}_{\mathrm{w}} \xrightarrow[\tau\to\infty]{N\to\infty} \infty. \tag{5.42}$$

We next turn our attention to the expected degree in the time-aggregated network.

5.5.4 Time-aggregated degree and finite size effects

The probability that two agents i, j with latent variables κ_i, κ_j do not interact, is obtained by setting $t = 0$ in (5.33),

$$\begin{aligned} r_{\mathrm{w}}(0; \kappa_i, \kappa_j) = & \frac{2\mu\kappa_i\kappa_j}{N} \left[\frac{T\Gamma(\tau+T)\Gamma(-T)}{\Gamma(\tau)} \right. \\ & \left. +(u_{ij}^{\min})^{-T} {}_2F_1(-T, 1-\tau-T, 1-T, u_{ij}^{\min}) \right], \end{aligned} \tag{5.43}$$

where $u_{ij}^{\min}$ in (5.34). Removing the condition on κ_i and κ_j gives the probability that two agents do not interact

$$r_{\mathrm{w}}(0) = \int\int r_{\mathrm{w}}(0; \kappa_i, \kappa_j)\rho(\kappa_i)\rho(\kappa_j)\mathrm{d}\kappa_i\mathrm{d}\kappa_j. \tag{5.44}$$

The expected time-aggregated degree is

$$\bar{k}_{\mathrm{aggr}} = (N-1)\left[1 - r_{\mathrm{w}}(0)\right]. \tag{5.45}$$

At $N \to \infty$, $\bar{k}_{\mathrm{aggr}}$ is given by (5.36). Substituting μ in (5.36) with its expression in (5.3), gives

$$\bar{k}_{\mathrm{aggr}} \xrightarrow{N\to\infty} \frac{\Gamma(\tau+T)\bar{\kappa}}{\Gamma(1+T)\Gamma(\tau)} \approx \frac{\tau^T\bar{\kappa}}{\Gamma(1+T)}, \tag{5.46}$$

which increases exponentially with T and linearly with $\bar{\kappa}$. Fig. 5.11 juxtaposes simulation results against (5.44, 5.45) and the limit in (5.46). We see an excellent agreement between (5.44, 5.45) and simulations, while (5.46) is a good approximation only at sufficiently low temperatures.

Similarly, the expected time-aggregated degree of a node with latent variable κ_i, is

$$\bar{k}_{\mathrm{aggr}}(\kappa_i) = (N-1)\left[1 - \int r_{\mathrm{w}}(0; \kappa_i, \kappa_j)\rho(\kappa_j)\mathrm{d}\kappa_j\right] \tag{5.47}$$

$$\xrightarrow{N\to\infty} \frac{\Gamma(\tau+T)\kappa_i}{\Gamma(1+T)\Gamma(\tau)} \approx \frac{\tau^T\kappa_i}{\Gamma(1+T)}. \tag{5.48}$$

Fig. 5.12 juxtaposes simulation results against (5.47) and (5.48). We again see an excellent agreement between the exact prediction (5.47) and simulations, while (5.48)

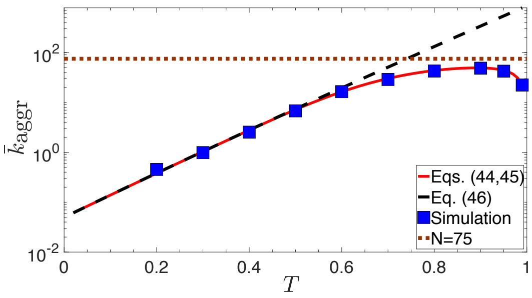

Figure 5.11: Average time-aggregated degree as a function of the temperature T in simulated networks vs. (5.44, 5.45) and (5.46). The simulation parameters are $N = 75, \bar{k} = 0.05$ and $\tau = 17376$ (as in the hospital), while $\kappa_i = \bar{k}$, $\forall i$, i.e., the PDF of κ is the Dirac delta function, $\rho(\kappa) = \delta(\kappa - \bar{k})$.

is a good approximation only for sufficiently small $\bar{k}_{\rm aggr}(\kappa)$. Therefore, one in general needs to use exact expressions [(5.44, 5.45), (5.47)] to accurately compute expected time-aggregated degrees. The thermodynamic limit approximations [(5.46), (5.48)] are accurate only at sufficiently low temperatures.

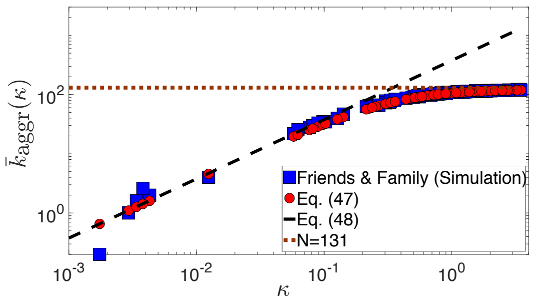

Figure 5.12: Average time-aggregated degree as a function of the latent degree variable κ in the simulated counterpart of the Friends & Family (Sec. 5.3.2) vs. (5.47) and (5.48). The simulation results are averages over 5 runs.

We also note that the normalization factor $w(\tau)$ of the weight distribution in (5.38) can be rewritten as

$$w(\tau) = \frac{\Gamma(1-T)\Gamma(T)\bar{k}_{\rm aggr}}{\tau\bar{\kappa}}, \tag{5.49}$$

where $\bar{k}_{\rm aggr}$ in (5.46). Fig. 5.13 juxtaposes (5.38) against simulation results, where in view of Fig. 5.11, we use in (5.49) the actual value of $\bar{k}_{\rm aggr}$ in the simulations instead of its limit in (5.46). We see again a very good agreement between theory and simulations.

5.5.5 Strength-degree correlations

We now analyze the strength-degree correlations in the time-aggregated network and justify previous empirical observations reporting a super-linear dependence between an individual's expected strength and its time-aggregated degree [100, 101].

The expected weight between two nodes i, j with latent variables κ_i, κ_j, is

$$\overline{w}(\kappa_i, \kappa_j) = \sum_{t=1}^{\tau} tr_{\rm w}(t; \kappa_i, \kappa_j), \tag{5.50}$$

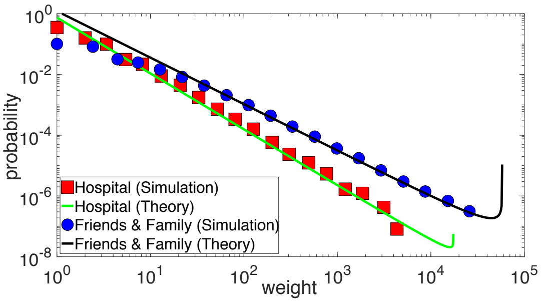

Figure 5.13: Aggregated weight distribution in the simulated counterparts of the hospital and Friends & Family (Sec. 5.3.2) vs. theoretical prediction given by (5.38, 5.49) with $\tau, T, \bar{k}_{\text{aggr}}$ and $\bar{\kappa} = \bar{d}$ as in Table 5.2. The upward bendings at the tails of the distributions are due to the finite observation time τ. Similar results hold for the rest of the counterparts.

where $r_{\text{w}}(t; \kappa_i, \kappa_j)$ in (5.33). At $N \to \infty$, the second term inside the brackets in (5.33) vanishes for $T \in (0, 1)$ and $t \geq 1$, yielding

$$N\overline{w}(\kappa_i, \kappa_j) \xrightarrow{N\to\infty} 2\mu\kappa_i\kappa_j T\tau \sum_{t=1}^{\tau} t \frac{\Gamma(\tau - t + T)\Gamma(t - T)}{\Gamma(\tau - t + 1)\Gamma(t + 1)}$$
$$= \frac{\tau\bar{k}\kappa_i\kappa_j}{\bar{\kappa}^2}. \qquad (5.51)$$

The expected strength of a node with latent variable κ_i, is

$$\bar{s}(\kappa_i) = N \int \overline{w}(\kappa_i, \kappa_j)\rho(\kappa_j)\mathrm{d}\kappa_j \xrightarrow{N\to\infty} \frac{\tau\bar{k}\kappa_i}{\bar{\kappa}}. \qquad (5.52)$$

Fig. 5.14 juxtaposes (5.52) against simulation results. We see that (5.52) can be a good approximation in finite networks. This is because the second term inside the brackets in (5.33) vanishes even for finite networks as t increases. The smaller the temperature the faster this term vanishes and the better the approximation in (5.52) is for finite networks.

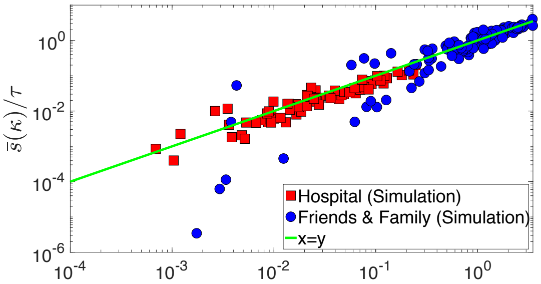

Figure 5.14: Normalized average strength $\bar{s}(\kappa)/\tau$ as a function of the latent degree variable κ in the simulated counterparts of the hospital and Friends & Family (Sec. 5.3.2). The results are averages over 20 and 5 runs, respectively. In the counterparts $\bar{k} = \bar{\kappa}$ $(= \bar{d})$, canceling out in (5.52).

We also see from (5.48, 5.52) that in the thermodynamic limit the expected strength of a node grows linearly with its expected time-aggregated degree,

$$\bar{s}(\kappa_i) \propto \bar{k}_{\text{aggr}}(\kappa_i). \qquad (5.53)$$

However, in the counterparts $\bar{k}_{\text{aggr}}(\kappa_i)$ grows sub-linearly with κ_i (Fig. 5.12), while $\bar{s}(\kappa_i)$ grows approximately linearly (Fig. 5.14). Thus, in the considered systems we expect the strength of a node to grow super-linearly with its time-aggregated degree, as verified in Fig. 5.15 and empirically observed in prior studies [100, 101].

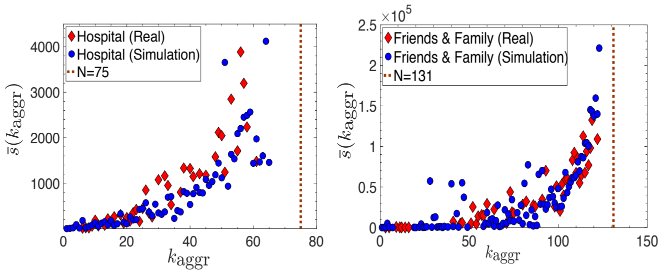

Figure 5.15: Average strength as a function of the time-aggregated degree in real and simulated networks. Similar results hold for the rest of the real networks and their counterparts from Sec. 5.3.2.

5.6 Component dynamics and temperature

Finally, we elucidate the important role of the temperature T in the formation of components. To this end, we consider the connected components formed in all time slots throughout the observation period τ, which consist of at least three nodes. We consider both unique and recurrent components. A component in a slot is called *unique* if it is seen for the first time, i.e., it is a component that does not consist of exactly the same nodes as a component seen in a previous slot. Otherwise, the component is recurrent. Fig. 5.16 shows that as T increases, the number of unique components increases almost exponentially up to a point and then decreases. This is because larger values of T increase the connection probability [Eq. (5.1)] at larger distances ($\chi_{ij} > 1$), while decreasing it at smaller distances ($\chi_{ij} < 1$). Since there are more pairs of nodes separated by larger distances, the number of unique components formed increases. However, at larger T closer to one, the probability of connections is relatively small at smaller and larger distances, which causes this number to decrease. The inset in Fig. 5.16 shows the size of the largest component formed.

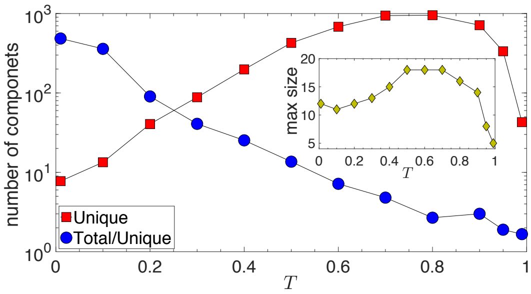

Figure 5.16: Number and size of components formed vs. temperature T. The simulation parameters are the same as in the counterpart of the hospital (Sec. 5.3.2) except that T varies in $(0, 1)$.

Further, Fig. 5.16 shows that the ratio of the total number of components formed to the number of unique components formed decreases with $T \in (0, 1)$. This means that as T increases fewer recurrent components are formed per unique component. This is expected since at larger T unique components consist of pairs separated by larger distances, and the probability to form again the same such components is vanishing.

5.7 Simulation of the spread of COVID-19 on synthetic networks generated from real survey data

The results in this section have not been published but we present interesting preliminary results on the applicability of the dynamic-$\mathbb{S}^1$ to model real life epidemic spreading scenarios to better inform decision makers on the implementation of containment measures.

Human proximity networks can be used to study the diseases that spread through contacts in a physical space. However, real human proximity data is not widely available and only few datasets exist for a small number of settings. On the other hand, survey data about the daily contacts among individuals is easy to generate and plenty of datasets exist. Here we show that realistic synthetic human proximity networks can be generated from such survey data using the dynamic-$\mathbb{S}^1$ model and simulate the spread of COVID-19 on the generated networks with a Susceptible Exposed Infected Removed (SEIR) model.

To generate synthetic human proximity networks with the dynamic-$\mathbb{S}1$ model, we consider a survey of 578 individuals is Cyprus [5] (unpublished data). In the data, the participants have declared their number of contacts in a single day at work, elsewhere and at home, before and during the lockdown in Cyprus during the COVID-19 pandemic in the summer of 2020. Although the reported contacts correspond to all contacts of the participants and not to their contacts with the other participants, to generate the synthetic networks we assume that the contacts in each period are among the 578 participants. The main idea is that the number of contacts reported by a participant at work/elsewhere/home is the participant's time-aggregated degree in a human proximity network corresponding to one day of observation at work/elsewhere/home. The dynamic-$\mathbb{S}1$ model can then generate a synthetic human proximity network for an average day at work/elsewhere/home, before or during the lockdown, using only the corresponding time-aggregated degrees of the nodes.

In Table 5.4 we categorize the 578 participants according to their age into different age groups. For each participant we consider her/his total number of daily contacts—the sum of her/his contacts at work, elsewhere and home—before and during the lockdown and compare the average number of daily contacts in each age group with the average number of daily contacts in the same age group from the POLYMOD dataset [68]. The POLYMOD dataset is a large-scale survey of the contact patterns of $7,290$ participants across eight European countries. The participants recorded their contacts in a diary for a period of one day including information such as age, sex, location, duration and frequency. Compared to the POLYMOD data, we observe a severe over-reporting of contacts in the Cyprus data. Most probably due to recall bias: the participants of the study were simply asked to report their number of contacts at work, elsewhere and at home in a regular day before the lockdown and a regular day during lockdown, instead of keeping a detailed diary of contacts like in the POLYMOD study. For this reason, we normalize the number of contacts at work, elsewhere and home of each participant so that the average number of daily contacts in each age group is approximately the same as in the POLYMOD data. For the contacts during lockdown, we normalize them so that the average number of daily contacts in each age group is approximately 3.1 as reported in a survey similar to POLYMOD conducted in the UK during the first COVID-19 lockdown [47].

age group	Participants		Average daily contacts		
	POLYMOD	Cyprus	POLYMOD	Cyprus before lockdown	Cyprus during lockdown
0-4	660	0	10.21	N/A	N/A
5-9	661	0	14.81	N/A	N/A
10-14	713	0	18.22	N/A	N/A
15-19	685	10	17.58	53.9	3.9
20-29	879	129	13.57	78.81	14.36
30-39	815	184	14.14	55.65	8.03
40-49	908	141	13.83	60.43	11.13
50-59	906	71	12.30	61.54	6.15
60-69	728	32	9.21	29.72	4.5
70+	270	11	6.89	9.27	3.36

Table 5.4: Average number of participants and average number of daily contacts per age group in the POLYMOD and the Cyprus datasets. For the Cyprus data we consider contacts before and during the lockdown. Note that the participants in the Cyprus dataset are 18 years old and above. We assume that the 18 and 19 years old participants have similar contacts as the 15-19 years old age group in the POLYMOD dataset.

To generate a synthetic human proximity network with the dynamic-$\mathbb{S}1$ corresponding to a single period (workplace, elsewhere or home) before or during the lockdown, we configure the model as follows. First, we assume that all 578 participants are present in all periods, therefore we set the number of nodes $N = 578$ in all cases. The number of time slots τ is set according to the duration of the period, which is the average of the hours that the participants reported for the period. We assume that the network time-slots have a duration of five minutes, thus $\tau = 84\,(7\,\text{hours})$ at work, $\tau = 12\,(1\,\text{hour})$ elsewhere and $\tau = 144\,(12\,\text{hours})$ at home before the lockdown, while $\tau = 24\,(2\,\text{hours})$ at work, $\tau = 0$ elsewhere and $\tau = 252\,(21\,\text{hours})$ at home during the lockdown. The average degree is assumed the same in all snapshots in the period $\bar{k}_t = \bar{\kappa}, t = 1, \ldots, \tau$. The hidden degrees per slot κ_i of each node $i = 1, \ldots, N$ and the temperature T are tuned simultaneously for each period. From Eq. (5.48), the hidden degree per slot κ_i of node i can be estimated as $\kappa_i = \bar{k}_{\text{aggr}}(\kappa_i)/\alpha$, where $\alpha = \tau^T/\Gamma(1+T)$ and $\bar{k}_{\text{aggr}}(\kappa_i)$ is the node's expected time-aggregated degree, i.e., the node's reported number of contacts in the period in our case. Hence, we tune the hidden degrees per slot κ_i and T simultaneously (as T changes the values of κ_i change) such that the average time-aggregated degree $\bar{k}_{\text{aggr}}$ in the resulting network is similar to the real average number of contacts in the period across the participants (see Table 5.5 for the values of T found for each period). The hidden similarity coordinate θ_i of each node i is sampled uniformly at random from 0 to 2π for each period, however, we assume they remain the same before and during the lockdown.

5.7.0.1 Properties of the synthetic human proximity networks

In this section we compare properties of the synthetic human proximity networks of each period before the lockdown versus during the lockdown using normalized

Period	T before lockdown	T during lockdown
Workplace	0.3	0.2
Elsewhere	0.5	N/A
Home	0.5	0.5

Table 5.5: Temperature T found for each period before and during the lockdown. The procedure to find these values is described in the text.

contacts. The properties considered in Fig. 5.17 are: i) the time-aggregated degree distribution, ii) contact duration distribution, iii) intercontact duration distribution and iv) group size distribution.

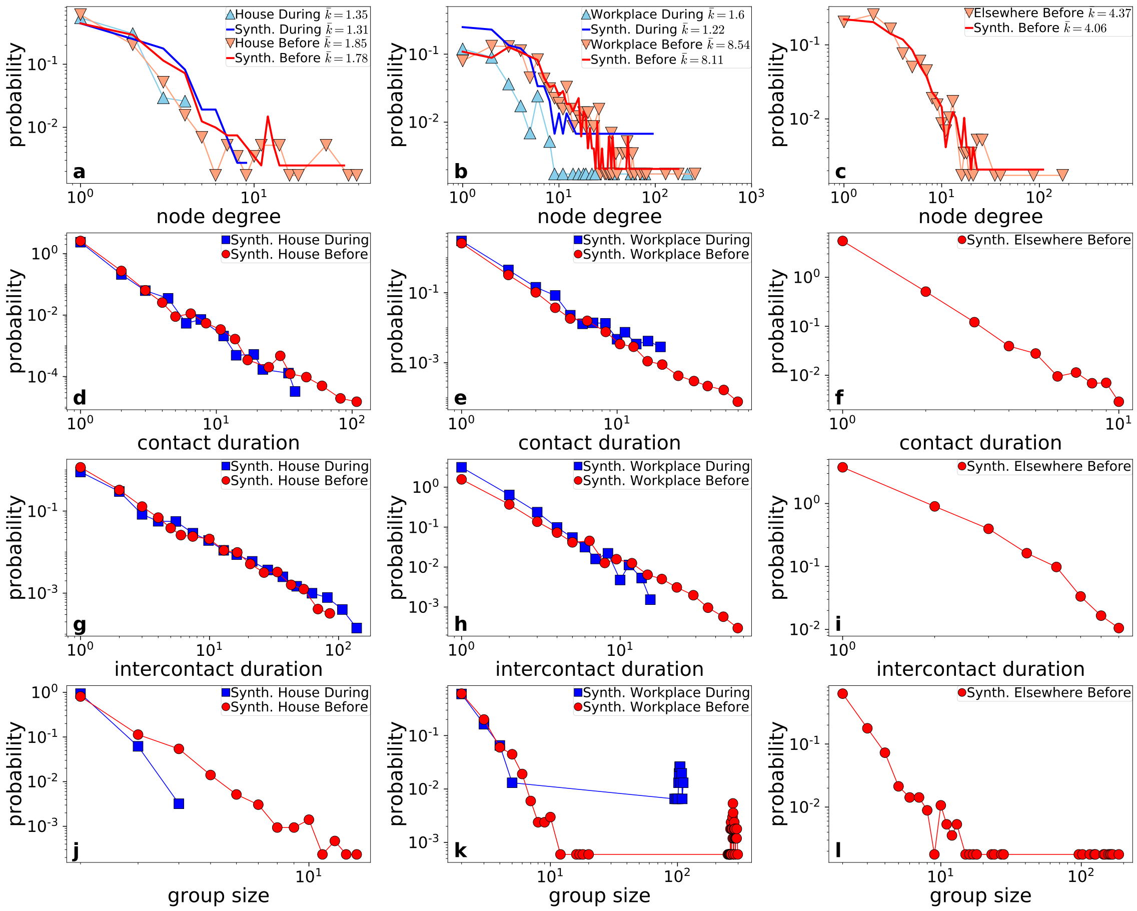

Figure 5.17: Properties of the synthetic human proximity network of normalized contacts for the home, workplace and elsewhere periods, before and during the lockdown. (**a-c**) time-aggregated degree distribution. The sky blue and light salmon triangles correspond to the real time-aggregated degree distributions in the Cyprus data during and before the lockdown, respectively. The blue and red solid lines correspond to the synthetic human proximity networks during and before the lockdown, respectively. The legend in each plot also shows the average time-aggregated degree in reality and in the synthetic networks. (**d-f**) contact duration distribution. The blue markers correspond to the periods during lockdown and the red markers to the periods before the lockdown. (**g-i**) same as (**d-f**) but for the intercontact duration distribution. (**j-l**) same as (**d-f**) but for the group size distribution. See Appendix B.1 for the properties corresponding to the synthetic networks generated from non-normalized contacts.

In Fig. 5.17 we can see that the dynamic-$\mathbb{S}1$ model reproduces the real time-aggregated degree distribution of the Cyprus data and the other properties look as expected in real human proximity networks. However, in Appendix B.1 we see that in the networks generated from non-normalized contacts, giant connected components form in the elsewhere and workplace periods before lockdown. Whereas, with the normalized contacts only giant connected components form in the workplace period before the lockdown and smaller groups form in all periods. This means that the model can also be used to identify overestimation of contacts in the survey data, since real human proximity network snapshots are sparse without giant connected components [9, 94, 101].

5.7.1 SEIR model simulations

In this section we simulate the spread of COVID-19 among the 578 participants of the Cyprus data based on their reported contacts before and during the lockdown. To this end, we generate a synthetic human proximity network from the normalized contacts and also from non-normalized contacts with the dynamic-$\mathbb{S}1$ model. Each network consists of 15 days before lockdown (from March 9 when the first case was reported in Cyprus until the start of the lockdown in March 24) and 59 days during lockdown (from March 24 until the end of the lockdown in May 21). Each day in the network consists of a work, an elsewhere and a home period. Each period in each day is generated as described in Section 5.7.

Compartmental models, such as the SEIR model, simplify the modeling of disease spreading by categorizing a population into compartments. According to the disease one wishes to model, different compartments may be considered. The SEIR model consists of the following compartments: the *Susceptible* compartment represents individuals that can be infected, the *Exposed* compartment represents individuals that have been infected but are not yet infectious, the *Infected* compartment represents infected individuals that can infect susceptible individuals and the *Removed* compartment represents individuals that have been "removed" from the population and can no longer be infected or infect others. In the classic compartmental models, a system of differential equations governs the progress of individuals between compartments, which for the SEIR model is:

$$\begin{cases} \dfrac{dS}{dt} = -\alpha \dfrac{I}{N} S \\ \dfrac{dE}{dt} = \alpha \dfrac{I}{N} S - \epsilon E \\ \dfrac{dI}{dt} = \epsilon E - \beta I \\ \dfrac{dR}{dt} = \beta I \end{cases} \tag{5.54}$$

where $N = S + E + I + R$ is the total population, α is the infection probability by which infected individuals infect susceptible individuals in a unit of time, ϵ is the probability by which exposed individuals transition to the infected compartment in a unit of time and β is the recovery probability by which infected individuals transition to the removed compartment in a unit of time. In the recent literature of the modeling of the spread of COVID-19, compartmental models have been widely used as a system of differential equations [15, 27, 33, 83, 111] but also as a dynamic process that runs on top of a network structure [28, 105]. Network-based epidemic models are more realistic than the classic compartmental models of differential equations, in the sense that infections happen only between infected and susceptible individuals connected in the network. This is particularly useful to model control measures that imply a decrease in contacts (social distancing) or constrained contacts (lockdown) [105].

Here we consider a dynamic SEIR process that runs on top of the synthetic human proximity networks generated with the dynamic-$\mathbb{S}1$ model. The process begins with a single node in the infected compartment chosen at random. Then, in each network snapshot each susceptible node that is connected with an infected node, becomes exposed with probability α; each exposed node becomes infected with probability ϵ and each infected node transitions to the removed compartment with probability β.

We run 50 dynamic SEIR processes on each synthetic network. We choose a different source infected node for each run. In all runs we use the same parameters tuned according to real reported values of the incubation period, the infection duration and R_0 for COVID-19. ϵ is the inverse of the incubation period of the disease. Incubation periods between 3 and 6 days on average have been reported [7, 38, 58], we consider 3 days. β is the inverse of the duration of the infection, which has been reported as 14 days on average for mild cases [73]. Adjusting for the 5 minute time slot duration of the synthetic networks, $\epsilon = \frac{5\,\text{minutes}}{3\,\text{days}} = \frac{5\,\text{minutes}}{4320\,\text{minutes}}$ and $\beta = \frac{5\,\text{minutes}}{14\,\text{days}} = \frac{5\,\text{minutes}}{20160\,\text{minutes}}$. Finally we set the infection probability α according to real reported values of R_0 in Cyprus using the relation $R_0 = \alpha/\beta$ of the classic SEIR model [86]. We consider $R_0 = 2.58$, 3.26 and 4.01 [84].

In Fig. 5.18 we plot the total cases, the daily new cases and the active infected cases, averaged over 50 dynamic SEIR processes on the synthetic network of normalized contacts, for $R_0 = 2.58$, 3.26 and 4.01. We see that the results are qualitatively similar to the corresponding real trends in [110] from March 9 until May 21 in Cyprus. In Appendix B.1 we also show results for the synthetic network of non-normalized contacts, where the results are qualitatively the same, but the majority of the participants become infected.

In Fig. 5.19 we consider the daily new cases originated at work, elsewhere and at home, averaged over 50 dynamic SEIR process on the synthetic network of normalized contacts, for $R_0 = 2.58$, 3.26 and 4.01. We observe that the majority of the cases originate at work, which is in agreement with the large number of clusters of COVID-19 cases reported in the workplace throughout Europe and the UK [23], while the contribution from the periods at home and elsewhere is almost non-existent on average. In Appendix B.1, we also show similar results for the synthetic network of non-normalized contacts.

In general, we observe that our approach produces qualitatively similar results to the real spread of COVID-19 in Cyprus from March 9 until May 21. We have also observed that the workplace is the main place of infection, perhaps lockdowns or strict control measures should target the workplace rather than “everywhere”.

The main difference between our work and the studies in Refs. [28, 105] is that the synthetic human proximity networks that we generate are time-varying, not static, and are constructed according to a realistic model of human proximity networks. Further, it has been shown in a real human proximity network of a hospital ward, that a dynamic SEIR process overestimates the total number of cases when run on the time-aggregated network rather than on the human proximity network [61]. In Ref. [61], other representations of the human proximity network were considered, the

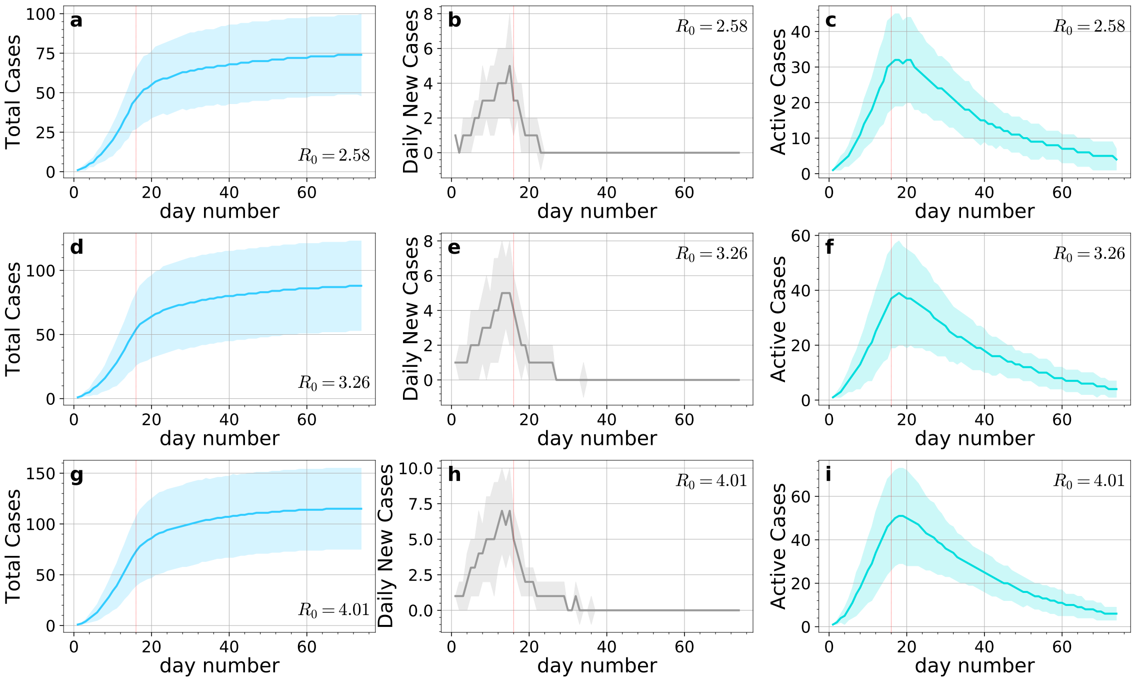

Figure 5.18: Total cases as a function of the number of days, new cases per day and active infected cases in each day. Results are averages over 50 dynamic SEIR processes on the synthetic network of normalized contacts. The shaded areas correspond to one standard deviation away from the average. The vertical red dashed line marks the beginning of the lockdown on day 16. (**a-c**) $R_0 = 2.58$. (**d-f**) $R_0 = 3.26$. (**g-i**) $R_0 = 4.01$. See Appendix B.1 for results with the synthetic network of non-normalized contacts.

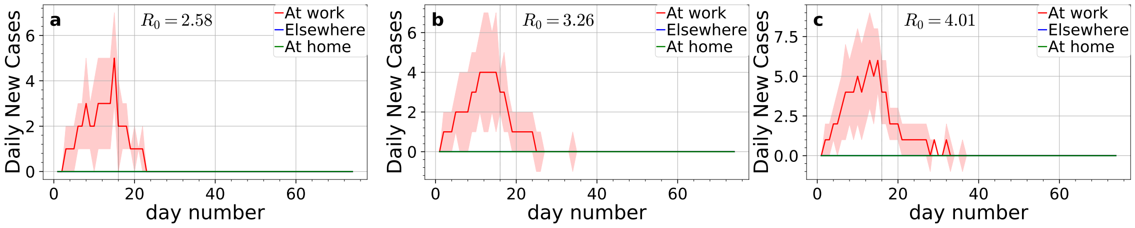

Figure 5.19: Daily new cases originated at work (red lines), elsewhere (blue lines) and at home (green lines). Results are averages over 50 dynamic SEIR processes on the synthetic network of normalized contacts. The shaded areas correspond to one standard deviation away from the average. The vertical black line marks the beginning of the lockdown on day 16. (**a**) $R_0 = 2.58$. (**b**) $R_0 = 3.26$. (**c**) $R_0 = 4.01$. See Appendix B.1 for results with the synthetic network of non-normalized contacts.

time-aggregated network is the second best approximation to the human proximity network. The best approximation is a network where each pair of nodes i, j is connected by an edge and the edge weight is sampled from a negative binomial distribution. This distribution is fitted to the empirical distribution of cumulative contact durations between individuals belonging to the same groups as i, j. In this case, the groups are the roles in the hospital: patients, physicians, nurses, ward assistants and caregivers. This could be a good network choice if one has access to detailed contact information among individuals and their durations, but this is not the case in the Cyprus data.

5.8 Discussion

Despite its simplicity the dynamic-$\mathbb{S}^1$ reproduces adequately many of the observed properties of real proximity networks. At the same time the model is amenable to mathematical analysis. We have proved here the model's main properties (Sec. 5.5). Other properties were studied only via simulations (Sec. 5.3.3) and it would be interesting in future work to prove those properties as well. We have seen that network temperature plays a central role in network dynamics, dictating the contact, inter-contact and weight distributions, the time-aggregated degrees, and the formation of unique and recurrent components.

The dynamic-$\mathbb{S}^1$ may not capture the properties of a real network *exactly*. For instance, the aggregated contact, inter-contact and weight distributions may deviate from pure power laws, may follow power laws with exponential cutoffs, may have different exponents than exactly $2+T, 2-T, 1+T$, etc., cf. Fig. 5.6(a). Further, we have seen that the pairwise inter-contact distributions are on average more skewed in real networks than in the model. As future work, it would be also interesting to investigate what mechanisms need to be introduced into the model in order to be able to capture such variations.

We also note that memory in the dynamic-$\mathbb{S}^1$ is induced only via the nodes' latent variables (κ, θ). Extensions to the model with *link persistence*, where connections/disconnections can also be copied from the previous to the next snapshot [66, 76], would allow additional control over the rate of dynamics, i.e., on how fast the topology changes from snapshot to snapshot. Further, generalizations of the model that would allow the nodes' latent variables (κ, θ) to change over time are desirable. However, for this purpose, one would first need to find the equations that realistically describe the motion of nodes in their latent spaces. The dynamic-$\mathbb{S}^1$ or extensions of it may apply to other types of time-varying networks, such as the ones considered in [50, 82], and constitute the basis of maximum likelihood estimation methods that infer the node coordinates and their evolution in the latent spaces of real systems [53]. Taken altogether, our results pave the way towards generative modeling of temporal networks that simultaneously satisfies simplicity, realism, and mathematical tractability.

Chapter 6

Hyperbolic mapping of human proximity networks

This chapter has been published, with some modifications, in “Scientific Reports” [88].

In Chapter 5, we have presented the dynamic-$\mathbb{S}^1$ model. The model assumes that each network snapshot is a realization of the $\mathbb{S}^1$ model of traditional (non-mobile) complex networks. The dynamic-$\mathbb{S}^1$ reproduces many of the observed characteristics of human proximity networks, while being mathematically tractable. We have proven several of the model's properties in Chapter 5.

In this chapter, we map human proximity networks into hyperbolic spaces founded on the dynamic-$\mathbb{S}^1$ model. Specifically, given that the dynamic-$\mathbb{S}^1$ can generate synthetic temporal networks that resemble human proximity networks across a wide range of structural and dynamical characteristics, can we reverse the synthesis and map (embed) human proximity networks into the hyperbolic space, in a way congruent with the model? Would the results of such mapping be meaningful? And could the obtained maps facilitate applications, such as community detection, routing on the temporal network, prediction of future links, and prediction of epidemic arrival times?

Here we provide the affirmative answers to these questions. Our approach is based on embedding the time-aggregated network of human proximity systems over an adequately large observation period, using methods developed for traditional complex networks that are based on the $\mathbb{S}^1$ model [32]. In the time-aggregated network, two nodes are connected if they are connected in at least one network snapshot during the observation period. We justify this approach theoretically by showing that the connection probability in the time-aggregated network in the dynamic-$\mathbb{S}^1$ model resembles the connection probability in the $\mathbb{S}^1$ model, and explicitly validate it in synthetic networks. Following this approach, we produce hyperbolic maps of six different real systems, and show that the obtained maps are meaningful: they can identify actual node communities, they can facilitate efficient greedy routing on the temporal network, and they can predict future links with significant precision. Further, we show that epidemic arrival times in the temporal network are positively correlated with the hyperbolic distance from the infection sources in the maps.

6.1 Results

6.1.1 Data

We consider the following face-to-face interaction networks from SocioPatterns [97]. (i) A hospital ward in Lyon [107], which corresponds to interactions involving patients and healthcare workers during five observation days. (ii) A primary school in Lyon [102], which corresponds to interactions involving children and teachers of ten different classes during two days. (iii) A scientific conference in Turin [45], which corresponds to interactions among conference attendees during two and a half days. (iv) A high school in Marseilles [64], which corresponds to interactions among students of nine different classes during five days. And (v) an office building in Saint Maurice [35], which corresponds to interactions among employees of 12 different departments during ten days. Each snapshot of these networks corresponds to an observation interval (time slot) of 20 s, while proximity was recorded if participants were within 1.5 m in front of each other.

We also consider the Friends & Family Bluetooth-based proximity network [1]. This network corresponds to the proximities among residents of a community adjacent to a major research university in the US during several observation months. We consider the data recorded in March 2011. Each snapshot corresponds to an observation interval of 5 min, while proximity was recorded if participants were within a radius of 10 m from each other. Thus proximity in this network does not imply face-to-face interaction. Table 6.1 gives an overview of the data.

Network	Days	N	τ	$\bar{n}$	$\bar{k}$	$\bar{k}_{\text{aggr}}$	T
Hospital	5	75	17376	2.9	0.05	30	0.84
Primary school	2	242	5846	30	0.18	69	0.72
Conference	2.5	113	10618	3.3	0.03	39	0.85
High school	5	327	18179	17	0.06	36	0.61
Office building	10	217	49678	2.8	0.01	39	0.74
Friends & Family	31	112	7317	58	1.5	57	0.48

Table 6.1: Overview of the considered real networks. N is the number of nodes, τ is the total number of time slots (snapshots), $\bar{n}$ is the average number of interacting (i.e., non-zero degree) nodes per snapshot, $\bar{k}$ is the average node degree per snapshot, $\bar{k}_{\text{aggr}}$ is the average degree in the time-aggregated network formed over the full observation duration τ, and parameter T is the network temperature used in the dynamic-$\mathbb{S}^1$ model to generate synthetic counterparts of the real systems (see Chapter 5). The table also shows the number of observation days for each network.

6.1.2 Preliminaries

We first provide an overview of the equivalence between the $\mathbb{S}^1$ model and random hyperbolic graphs or $\mathbb{H}^2$ model.Then we show that the connection probability in the time-aggregated network in the dynamic-$\mathbb{S}^1$ model, resembles the connection probability in the $\mathbb{S}^1$ model. Based on this equivalence, we then map the time-aggregated networks of the considered real data to the hyperbolic space using a recently developed method that is based on the $\mathbb{S}^1$ model.

6.1.2.1 $\mathbb{S}^1$ model

The $\mathbb{S}^1$ model described in Chapter 5, Sec. 5.1 is equivalent to random hyperbolic graphs, i.e., to the hyperbolic $\mathbb{H}^2$ model [55], after transforming the degree variables κ_i to radial coordinates r_i via

$$r_i = \hat{R} - 2 \ln \frac{\kappa_i}{\kappa_0}, \tag{6.1}$$

where κ_0 is the smallest κ_i and $\hat{R} = 2\ln\left[N/(\pi\mu\kappa_0^2)\right]$ is the radius of the hyperbolic disk where all nodes reside. After this change of variables, the effective distance in Chapter 5, Eq. (5.2) becomes $\chi_{ij} = e^{\frac{1}{2}(x_{ij}-\hat{R})}$, where $x_{ij} = r_i + r_j + 2\ln(\Delta\theta_{ij}/2)$ is approximately the hyperbolic distance between nodes i and j [55]. Therefore, we can refer to the degree variables κ_i as "coordinates" and use terms *effective distance* and *hyperbolic distance* interchangeably.

Given the ability of the $\mathbb{S}^1/\mathbb{H}^2$ model to construct synthetic networks that resemble real networks, several methods have been developed to map real networks into the hyperbolic plane, i.e., to infer the nodes' latent coordinates r (or κ) and θ, according to the model [2, 11, 12, 32, 75, 77]. The hyperbolic maps produced by these methods have been shown to be meaningful, and have been efficiently used in applications such as community detection, greedy routing and link prediction [2, 3, 4, 11, 12, 54, 74, 75, 77, 79]. Model-free mapping methods have also been developed [70]. Further, on a related note, there is a large body of work on embedding both static and temporal networks into Euclidean spaces, e.g., see Refs. [21, 53, 106], and references therein. However, no prior work has considered embedding temporal networks into hyperbolic spaces, which provide a more accurate reflection of the geometry of real networks [79].

6.1.3 Hyperbolic mapping of human proximity networks

6.1.3.1 Theoretical considerations

Assuming that a sequence of network snapshots G_t, $t = 1, \ldots, \tau$, has been generated by the dynamic-$\mathbb{S}^1$, we show below that we can accurately infer the nodes' latent coordinates κ, θ from the time-aggregated network, using existing methods that are based on the $\mathbb{S}^1$ model. This is justified by the fact that the connection probability in the time-aggregated network of the dynamic-$\mathbb{S}^1$ resembles the connection probability in the $\mathbb{S}^1$. Indeed, in the time-aggregated network two nodes are connected if they are connected in at least one of the snapshots. Assuming for simplicity that each snapshot has the same average degree $\bar{k}_t = \bar{k}$, the connection probability in the time-aggregated network of the dynamic-$\mathbb{S}^1$, is

$$P(\chi_{ij}) = 1 - \left[1 - p(\chi_{ij})\right]^{\tau}, \tag{6.2}$$

where $p(\chi_{ij})$ is given by Eq. (5.1) in Chapter 5. Further, as shown in Chapter 5, Eq. (5.5.4), the expected degree of a node in the time-aggregated network, $\tilde{\kappa}$, is related to the node's latent degree κ, via

$$\tilde{\kappa} = \alpha\kappa, \tag{6.3}$$

where $\alpha = \tau^T/\Gamma(1+T)$ for $\tau \gg 1$, and Γ is the gamma function. Eq. (6.3) is derived in the thermodynamic limit ($N \to \infty$), where there are no cutoffs imposed to node degrees by the network size. We can therefore rewrite (6.2) as

$$\begin{aligned} P(\tilde{\chi}_{ij}) &= 1 - \left[1 - p(\alpha\tilde{\chi}_{ij})\right]^{\tau} & (6.4)\\ &= 1 - \left\{1 + \frac{1}{\tau}\left[\frac{\Gamma(1+T)}{\tilde{\chi}_{ij}}\right]^{1/T}\right\}^{-\tau} \\ &\approx 1 - e^{-\left[\frac{\Gamma(1+T)}{\tilde{\chi}_{ij}}\right]^{1/T}}, & (6.5)\end{aligned}$$

where

$$\tilde{\chi}_{ij} = \frac{R\Delta\theta_{ij}}{\tilde{\mu}\tilde{\kappa}_i\tilde{\kappa}_j} = \frac{\chi_{ij}}{\alpha} \tag{6.6}$$

is the effective distance between nodes i and j in the time-aggregated network, while $\tilde{\mu} = \mu/\alpha$. The exponential approximation in (6.5) holds for sufficiently large τ. We also note that since $T \in (0, 1)$, $0.88 < \Gamma(1 + T) < 1$. At large distances, $\tilde{\chi}_{ij} \gg \Gamma(1 + T)$, we can use the approximation $e^{-x} \approx 1 - x$ in (6.5), to write

$$P(\tilde{\chi}_{ij}) \approx \frac{C}{\tilde{\chi}_{ij}^{1/T}} \propto \frac{1}{\tilde{\chi}_{ij}^{1/T}} \approx p(\tilde{\chi}_{ij}), \tag{6.7}$$

where $p(x)$ is given by (5.1), while $C = \Gamma(1 + T)^{1/T}$, $0.56 < C < 1$. At small distances, $\tilde{\chi}_{ij} \ll \Gamma(1 + T)$, the exponential in (6.5) is much smaller than one, and we can write $P(\tilde{\chi}_{ij}) \approx 1 \approx p(\tilde{\chi}_{ij})$. In other words, at both small and large effective distances $\tilde{\chi}_{ij}$, the connection probability in the time-aggregated network resembles the Fermi-Dirac connection probability in the $\mathbb{S}^1$ model. Fig. 6.1 illustrates this effect in the time-aggregated networks of synthetic counterparts of real systems, whose snapshots can also have different average degrees $\bar{k}_t, t = 1, \ldots, \tau$ (see Chapter 5, Fig. 5.3).

Given this equivalence, in Fig. 6.2 we apply Mercator, a recently developed embedding method based on the $\mathbb{S}^1$ model [32], to the time-aggregated network of the synthetic counterparts of the hospital and primary school. Mercator infers the nodes' coordinates $(\tilde{\kappa}, \theta)$ from the time-aggregated network (see Sec. 6.3.1), and from $\tilde{\kappa}$ we estimate κ using (6.3). We also modified Mercator to use the connection probability in Eq. (6.4) instead of the connection probability in Eq. (5.1) in Chapter 5 (see Appendix C.7). Fig. 6.2 shows that the two versions of Mercator perform similarly, inferring the nodes' latent coordinates remarkably well. Similar results hold for the synthetic counterparts of the rest of the real systems (Appendix C.2). In the rest of the chapter, we use the original version of Mercator as its implementation is simpler and does not require knowledge of parameter τ.

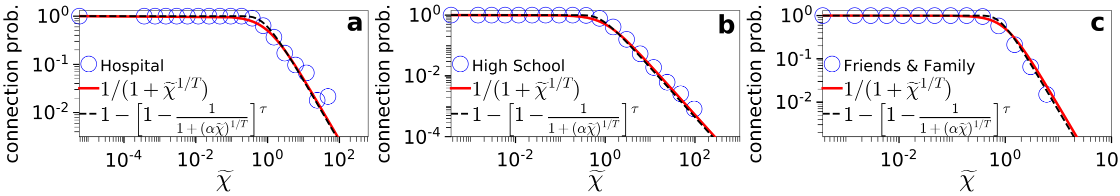

Figure 6.1: Connection probability in the time-aggregated network versus Fermi-Dirac connection probability. The results correspond to the synthetic counterparts of the hospital, high school and Friends & Family, constructed using the dynamic-$\mathbb{S}^1$ model as described in Chapter 5, Sec. 5.3.2. The blue circles show the empirical connection probabilities. The solid red and dashed black lines correspond to Eq. (5.1) in Chapter 5 and Eq. (6.4), respectively. The values of parameters T and τ in each case are as shown in Table 6.1, while $\alpha = \tau^T/\Gamma(1 + T)$. Similar results hold for the counterparts of the rest of the real systems (see Appendix C.1).

6.1.3.2 Aggregation interval

As the aggregation interval τ increases, the time-aggregated network becomes denser, eventually turning into a fully connected network. This can be seen in (6.2), where

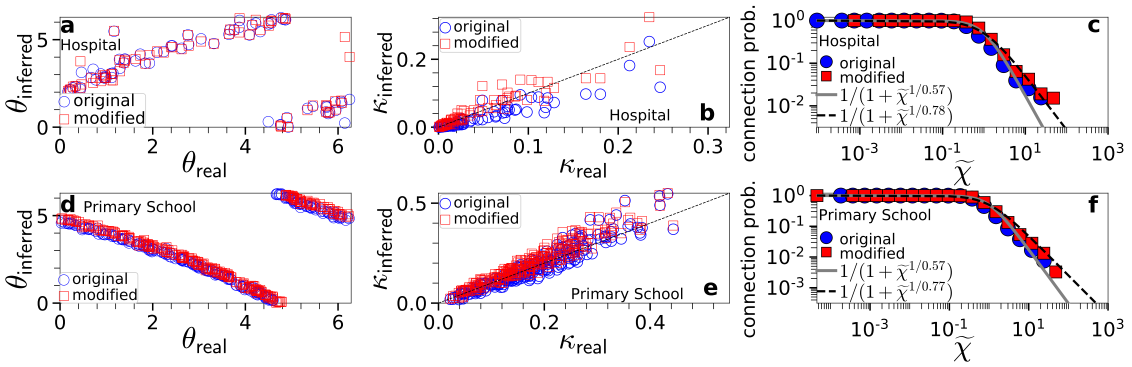

Figure 6.2: Inference of latent coordinates (κ, θ) with the original and modified versions of Mercator. The top row corresponds to a synthetic counterpart of the hospital, while the bottom row to a synthetic counterpart of the primary school. Both versions of Mercator are applied to the corresponding time-aggregated network formed over the full duration τ in Table 6.1. (**a** and **d**) Inferred versus real θ. (**b** and **e**) Inferred versus real κ. For each node, κ_{inferred} is estimated as $\kappa_{\text{inferred}} = \tilde{\kappa}/\alpha$, where $\tilde{\kappa}$ is the node's inferred latent degree in the time-aggregated network, while $\alpha = \tau^T/\Gamma(1+T)$, with τ as in Table 6.1 and T as inferred by each version of Mercator. (**c** and **f**) Connection probability as a function of the effective distance $\tilde{\chi}$ in the time-aggregated network computed using the inferred coordinates $(\tilde{\kappa}, \theta)$. The solid grey and dashed black lines correspond to Eq. (5.1) in Chapter 5 with temperature T as inferred by each version of Mercator. For the two networks, the original version estimates $T = 0.57$, the modified version estimates $T = 0.78$ and 0.77, while the actual values are $T = 0.84$ and 0.72. In general, the modified version estimates values of T closer to the actual values. However, both versions of Mercator perform remarkably well at estimating the nodes' latent coordinates (κ, θ). We note that due to rotational symmetry of the model, the inferred angles can be globally shifted compared to the real angles by any value in $[0, 2\pi]$.

irrespective of network size, at $\tau \to \infty$, $P(\chi_{ij}) \to 1, \forall i, j$. Further, at $\tau \to \infty$, $\alpha \to \infty$, and by (6.6) $\tilde{\chi}_{ij} \to 0$, $\forall i, j$. Clearly, no meaningful inference can be made in a fully connected network as all nodes "look the same". Thus for an accurate inference of the nodes' coordinates the interval τ has to be sufficiently small such that the corresponding time-aggregated network is not too dense. On the other hand, for intervals τ that are not sufficiently large there may not be enough data to allow accurate inference, as network snapshots are often very sparse in human proximity systems, consisting of only a fraction of nodes (Table 6.1). This effect is illustrated in Fig. 6.3, where we quantify the difference between real and inferred coordinates as a function of τ in a synthetic counterpart of the primary school. We see in Fig. 6.3 that there is a wide range of adequately large τ values, e.g., $500 < \tau < 10000$, where the accuracy of inference for both κ and θ is simultaneously high, while as τ becomes too large or too small accuracy deteriorates. Similar results hold for the counterparts of the rest of the considered real systems (Appendix C.8). The exact range of τ values where inference accuracy is high depends on the system's parameters, e.g., sparser networks (lower average snapshot degree) allow aggregation over longer intervals, as it takes longer for the time-aggregated network to become too dense. Further, our results with the synthetic counterparts suggest that daily aggregation intervals should be sufficient for accurate inference in most cases. Indeed, in this chapter we embed the time-aggregated networks of the considered real systems formed over the full observation durations τ in Table 6.1, as well as corresponding time-aggregated networks formed over individual observation days, obtaining in both cases meaningful results.

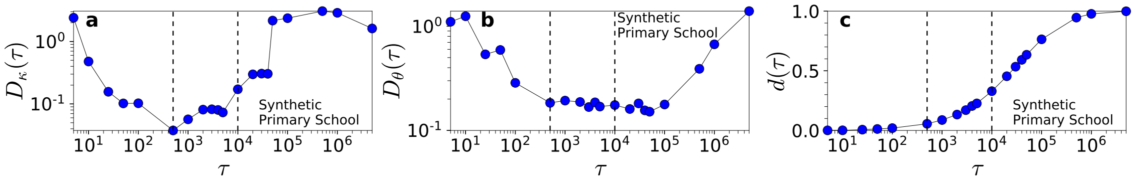

Figure 6.3: Inference accuracy vs. aggregation interval. The results correspond to a synthetic counterpart of the primary school constructed using the dynamic-$\mathbb{S}^1$ model. **(a)** Average difference between the inferred and real latent degrees as a function of the aggregation interval τ, $D_\kappa(\tau) = \sum_{i=1}^N |\kappa^i_{\text{inferred}} - \kappa^i_{\text{real}}|/N$, where $\kappa^i_{\text{inferred}}$ (κ^i_{real}) is the inferred (real) latent degree of node i. **(b)** Same as in (a) but for the average difference between the inferred and real angular coordinates, $D_\theta(\tau) = \sum_{i=1}^N |\theta^i_{\text{inferred}} - \theta^i_{\text{real}}|/N$. Before computing $D_\theta(\tau)$, the inferred angles are globally shifted such that the sum of the squared distances between real and inferred angles is minimized (to this end, we apply a Procrustean rotation [89], see Appendix C.8 for details). **(c)** Density of the time-aggregated network as a function of τ, $d(\tau) = 2L/[N(N-1)]$, where L is the number of links in the network. The vertical dashed lines indicate the interval $500 \leq \tau \leq 10000$. In this interval, $D_\kappa(\tau) < 0.2$, $D_\theta(\tau) < 0.2$, and $0.06 < d(\tau) < 0.33$.

6.1.3.3 Hyperbolic maps of real systems

In Fig. 6.4 we apply Mercator to the time-aggregated network of the real networks in Table 6.1 and visualize the obtained hyperbolic maps and the corresponding connection probabilities. We see that the embeddings are meaningful, as we can identify in them actual node communities that correspond to groups of nodes located close to each other in the angular similarity space. These communities reflect the organization of students and teachers into classes (Figs. 6.4b and 6.4c), employees into departments (Fig. 6.4d), while no communities can be identified in the hospital (Fig. 6.4a). In all cases, we see a good match between empirical and theoretical connection probabilities (Figs. 6.4e-h). Next, we turn our attention to greedy routing.

6.1.4 Human-to-human greedy routing

A problem of significant interest in mobile networking is how to efficiently route data in opportunistic networks, like human proximity systems, where the mobility of nodes creates contact opportunities among nodes that can be used to connect parts of the network that are otherwise disconnected [16, 20, 44, 48]. Motivated by this problem, and by the remarkable efficiency of hyperbolic greedy routing in traditional complex networks [4, 12, 74], we investigate here if hyperbolic greedy routing can facilitate navigation in human proximity systems. To this end, we consider the following simplest greedy routing process, which performs routing on the temporal network using the coordinates inferred from the time-aggregated network.

Human-to-human greedy routing (H2H-GR). In H2H-GR, a node's address is its coordinates $(\tilde{\kappa}, \theta)$, and each node knows its own address, the addresses of its neighbors (nodes currently within proximity range), and the destination address written in the packet. A node holding the packet (carrier) forwards the packet to its neighbor with the smallest effective distance to the destination, but only if that distance is smaller than the distance between the carrier and the destination. Otherwise, or if the carrier currently has no neighbors, the carrier keeps the packet. Clearly, a carrier delivers the packet to the destination if the latter is its neighbor. We note that there are no routing loops in H2H-GR, i.e., no node receives the same packet twice. Indeed, consider for instance a packet from a node i_0 to a node i_n, which has followed the

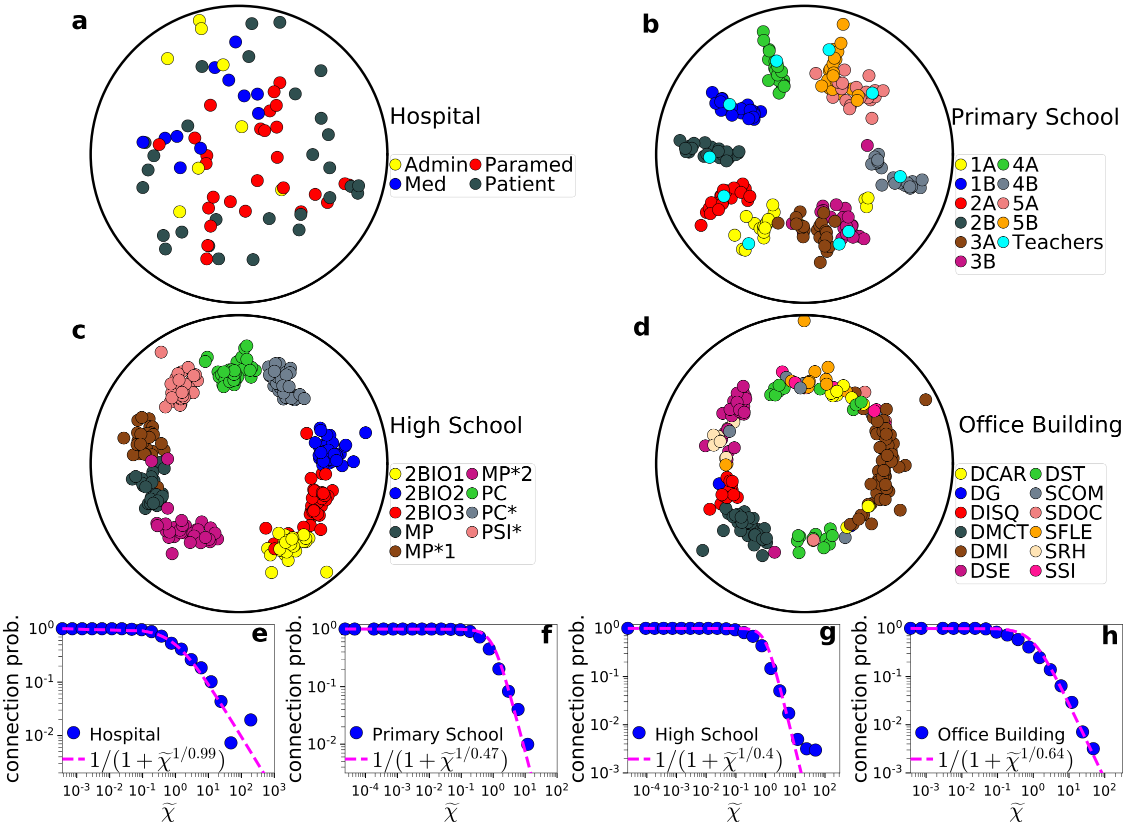

Figure 6.4: Hyperbolic embeddings of human proximity networks. (**a-d**) Hyperbolic maps of the time-aggregated networks of the hospital, primary school, high school and office building. In each case we consider the time-aggregated network formed over the full observation duration τ shown in Table 6.1. The nodes are positioned according to their inferred hyperbolic coordinates (r, θ) in the time-aggregated network [the radial coordinates r are computed using (6.1)]. The nodes are colored according to group membership information available in the metadata of each network. In the hospital, the nodes are administrative staff (Admin), medical doctors (Med), nurses and nurses' aides (Paramed), and patients (Patient). In the primary school, the nodes are teachers and students of the following classes: 1st grade (1A, 1B), 2nd grade (2A, 2B), 3rd grade (3A, 3B), 4th grade (4A, 4B), and 5th grade (5A, 5B). In the high school, the nodes are students of nine different classes with the following specializations: biology (2BIO1, 2BIO2, 2BIO3), mathematics and physics (MP, MP*1, MP*2), physics and chemistry (PC, PC*), and engineering studies (PSI*). In the office building, the nodes are employees working in different departments such as scientific direction (DISQ), chronic diseases and traumatisms (DMCT), department of health and environment (DSE), human resources (SRH), and logistics (SFLE). (**e-h**) Corresponding empirical connection probabilities as a function of the effective distance $\tilde{\chi}$. The pink dashed lines correspond to (5.1) with temperatures T as inferred by Mercator, $T = 0.99$, 0.47, 0.40 and 0.64, respectively. The maps for the conference and Friends & Family can be found in Appendix C.3. Daily hyperbolic maps for each real system can be found in Appendix C.5.

path $\{i_0, i_1, i_2, \ldots, i_{n-1}, i_n\}$. This means that $\tilde{\chi}_{i_0 i_n} > \tilde{\chi}_{i_1 i_n} > \tilde{\chi}_{i_2 i_n} > \ldots > \tilde{\chi}_{i_{n-1} i_n}$, where $\tilde{\chi}_{i_k i_n}$ is the effective distance between nodes i_k and i_n. A node i_k in the path never forwards the packet to a node i_l with $l < k$, i.e., to a node that has seen the packet before, because $\tilde{\chi}_{i_l i_n} > \tilde{\chi}_{i_k i_n}$.

For each network in Table 6.1, we simulate H2H-GR in one of its observation days. We consider the following two cases: i) H2H-GR that uses the nodes' coordinates inferred from the time-aggregated network of the considered day (current coordinates); and ii) H2H-GR that uses the nodes' coordinates inferred from the time-aggregated network of the previous day (previous coordinates). In the time-aggregated network of a day, two nodes are connected if they are connected in at least one network snapshot in the day. We compare these two cases to a baseline *random routing strategy (H2H-RR)*, where the carrier first determines the set of its neighbors that have never received the packet before, and then forwards the packet to one of these neighbors at random. If the destination is a neighbor the carrier forwards the packet to it. The carrier keeps the packet if it currently has no neighbors, or if all of its neighbors have received the packet before. Thus, there are no routing loops in H2H-RR either.

Performance metrics. We evaluate the performance of the algorithms according to the following two metrics: i) the percentage of successful paths, p_s, which is the proportion of paths that reach their destinations by the end of the considered day; and ii) the average stretch over the successful paths, $\bar{s}$. We define the stretch as the ratio of the hop-lengths of the paths found by the algorithms to the corresponding shortest time-respecting paths [43] in the network.

The results are shown in Table 6.2. We see that H2H-GR that uses the current coordinates significantly outperforms H2H-RR in both success ratio and stretch. The improvement can be quite significant. For instance, in the primary school the success ratio increases from 34% to 82%, while the average stretch decreases from 24.9 to 3.9. Similarly, in the hospital the success ratio increases from 38% to 80%, while the average stretch decreases from 7 to 2.2. These results show that hyperbolic greedy routing can significantly improve navigation. However, the success ratio decreases considerably if H2H-GR uses the previous coordinates. This suggests that the node coordinates change to a considerable extend from one day to the next. In Appendix C.5, we verify that this is indeed the case. Nevertheless, H2H-GR that uses the previous coordinates still outperforms H2H-RR with respect to success ratio, while achieving significantly lower stretch similar to the stretch with the current coordinates (Table 6.2).

Table 6.3 shows the same results for the synthetic counterparts of the real systems, where we can make qualitatively similar observations. Further, we see that H2H-GR achieves higher success ratios using the inferred coordinates in the counterparts compared to the real systems. This is not surprising as the counterparts are by construction maximally congruent with the assumed geometric model (dynamic-$\mathbb{S}^1$). Also, H2H-GR that uses the previous coordinates maintains high success ratios in the counterparts. This is expected, as the coordinates in the counterparts do not change over time. Thus the coordinates inferred from the time-aggregated network of the previous day are quite similar (but not exactly the same) to the ones inferred from the time-aggregated network of the day where routing is performed (see Appendix C.4).

The metrics in Tables 6.2 and 6.3 are computed across all source-destination pairs. In Figs. 6.5 and 6.6 we also compute these metrics as a function of the effective distance between the source-destination pairs. We see that H2H-GR that uses the current coordinates achieves high success ratios, approaching 100%, as the effective distance between the pairs decreases. As the effective distance between the pairs increases, the success ratio decreases. The average stretch for successful H2H-GR paths is always low.

H2H-RR also achieves considerably high success ratios for pairs separated by small distances (Fig. 6.5). This is because, even though packets in H2H-RR are forwarded to neighbors at random, the neighbors are not random nodes but nodes closer to the carriers in the hyperbolic space. Thus, packets between pairs separated by smaller distances have higher chances of finding their destinations. However, the stretch of

Real network	H2H-GR (current coordinates)	H2H-GR (previous coordinates)	H2H-RR
Hospital	$p_s = 0.80$, $\bar{s} = 2.2$	$p_s = 0.47$, $\bar{s} = 2.0$	$p_s = 0.38$, $\bar{s} = 7.0$
Primary school	$p_s = 0.82$, $\bar{s} = 3.9$	$p_s = 0.65$, $\bar{s} = 3.6$	$p_s = 0.34$, $\bar{s} = 24.9$
Conference	$p_s = 0.70$, $\bar{s} = 2.2$	$p_s = 0.35$, $\bar{s} = 2.0$	$p_s = 0.29$, $\bar{s} = 7.9$
High school	$p_s = 0.29$, $\bar{s} = 2.0$	$p_s = 0.13$, $\bar{s} = 1.9$	$p_s = 0.07$, $\bar{s} = 5.9$
Office building	$p_s = 0.15$, $\bar{s} = 1.4$	$p_s = 0.10$, $\bar{s} = 1.4$	$p_s = 0.06$, $\bar{s} = 2.5$
Friends & Family	$p_s = 0.45$, $\bar{s} = 1.8$	$p_s = 0.31$, $\bar{s} = 2.0$	$p_s = 0.21$, $\bar{s} = 5.3$

Table 6.2: Success ratio p_s and average stretch $\bar{s}$ of H2H-GR and H2H-RR in real networks. H2H-GR uses the coordinates inferred either from the time-aggregated network of the considered day where routing is performed (current coordinates); or from the time-aggregated network of the previous day (previous coordinates). The considered days in the hospital, primary school, conference, high school and office building are observation days $5, 2, 3, 5$ and 10, respectively. In Friends & Family, the considered day is the 31st of March 2011. For a fair comparison with H2H-GR that uses the previous coordinates, we ignore during all routing processes the nodes that exist in the considered day but not in the previous day, since for such nodes we cannot infer their coordinates from the previous day. The percentage of such nodes is $17\%, 3\%, 7\%, 6\%, 14\%$ and 3% for the hospital, primary school, conference, high school, office building and Friends & Family, respectively. In all cases, routing is performed among all possible source-destination pairs in the considered day that also exist in the previous day.

Synthetic network	H2H-GR (current coordinates)	H2H-GR (previous coordinates)	H2H-RR
Hospital	$p_s = 0.92$, $\bar{s} = 2.2$	$p_s = 0.78$, $\bar{s} = 2.2$	$p_s = 0.42$, $\bar{s} = 9.2$
Primary school	$p_s = 0.98$, $\bar{s} = 3.7$	$p_s = 0.97$, $\bar{s} = 3.8$	$p_s = 0.53$, $\bar{s} = 33.9$
Conference	$p_s = 0.85$, $\bar{s} = 2.4$	$p_s = 0.70$, $\bar{s} = 2.4$	$p_s = 0.31$, $\bar{s} = 9.8$
High school	$p_s = 0.72$, $\bar{s} = 2.7$	$p_s = 0.59$, $\bar{s} = 2.4$	$p_s = 0.11$, $\bar{s} = 7.8$
Office building	$p_s = 0.26$, $\bar{s} = 1.5$	$p_s = 0.17$, $\bar{s} = 1.5$	$p_s = 0.06$, $\bar{s} = 3.0$
Friends & Family	$p_s = 0.82$, $\bar{s} = 2.2$	$p_s = 0.70$, $\bar{s} = 2.3$	$p_s = 0.23$, $\bar{s} = 5.4$

Table 6.3: Same as in Table 6.2 but for the synthetic counterparts of the real systems constructed with the dynamic-$\mathbb{S}^1$ model. The results in each case correspond to one temporal network realization, while H2H-GR uses inferred coordinates as in Table 6.2.

successful paths in H2H-RR is quite high (Fig. 6.6). Further, we see that in real networks the success ratio of H2H-GR that uses the previous coordinates resembles in most cases the one of H2H-RR (Figs. 6.5a-c and Appendix C.4). However, the stretch in H2H-GR is always significantly lower than in H2H-RR (Figs. 6.6a-c and Appendix C.4).

Taken altogether, these results show that hyperbolic greedy routing can facilitate efficient navigation in human proximity networks. The success ratio for pairs separated by large effective distances can be low (Fig. 6.5). However, it is possible that more sophisticated algorithms than the one considered here could improve the success ratio for such pairs without significantly sacrificing stretch. Further, using coordinates from past embeddings decreases the success ratio. Even though the average stretch remains low, this observation suggests that the evolution of the nodes' coordinates should also be taken into account. Such investigations are beyond the scope of this thesis. Finally, we note that in Appendix C.4, we consider H2H-GR

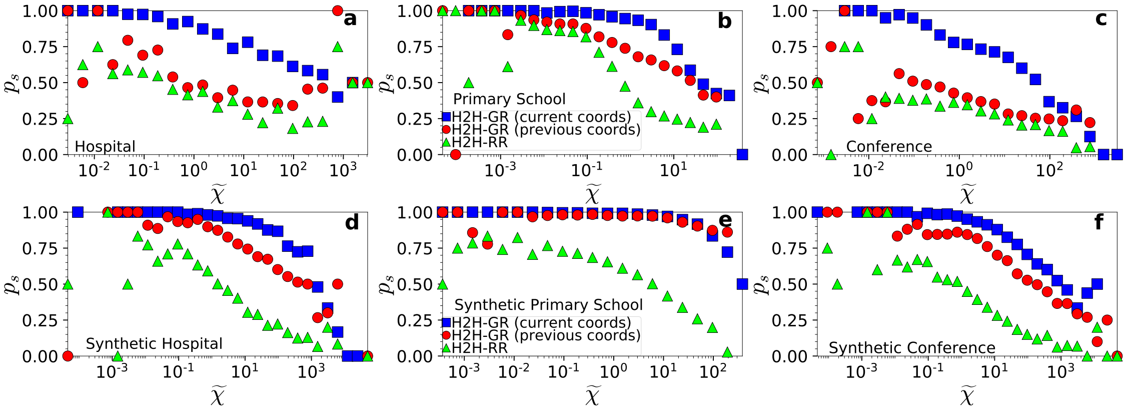

Figure 6.5: Success ratio p_s of H2H-GR and H2H-RR as a function of the effective distance $\tilde{\chi}$ between source-destination pairs. The top row corresponds to the results of the hospital, primary school and conference in Table 6.2, while the bottom row to the results of their synthetic counterparts in Table 6.3. The success ratio for H2H-RR and H2H-GR that uses the previous coordinates is shown as a function of the effective distance between the pairs in the previous day. Similar results hold for the other real networks and their synthetic counterparts (Appendix C.4).

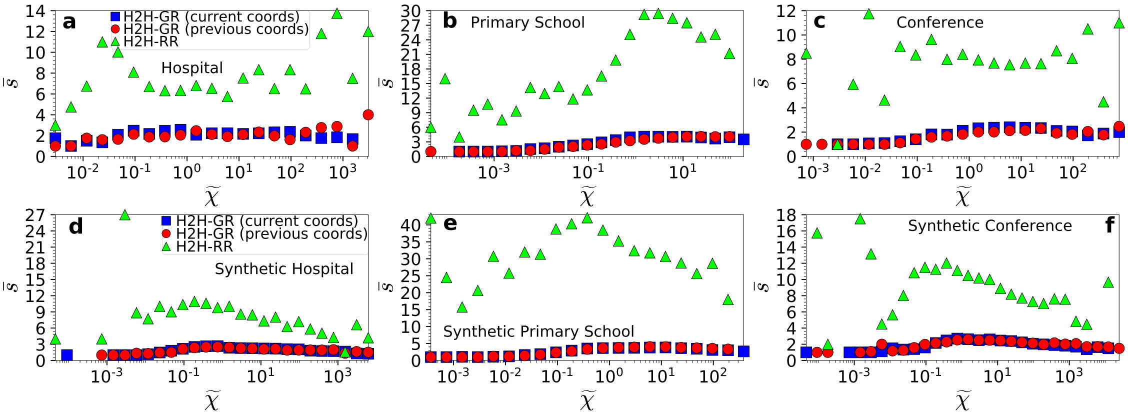

Figure 6.6: Same as in Fig. 6.5 but for the average stretch $\bar{s}$. Similar results hold for the other real networks and their synthetic counterparts (Appendix C.4).

that uses only the angular similarity distances among the nodes, and find that it performs worse than H2H-GR that uses the effective distances. This means that in addition to node similarities, node expected degrees (or popularities [79]) also matter in H2H-GR, even though the distribution of node degrees in human proximity systems is quite homogeneous [78].

6.1.5 Link prediction

In this section, we turn our attention to link prediction. We want to see how well we can predict if two nodes are connected in the time-aggregated network of a day, if we know the effective distances among the nodes in the previous day. To this end, for each pair of nodes i, j in the previous day that is also present in the day of interest, we assign a score $s_{ij} = 1/\tilde{\chi}_{ij}$, where $\tilde{\chi}_{ij}$ is the inferred effective distance between i and j in the time-aggregated network of the previous day. The higher the s_{ij}, the higher is the likelihood that i and j are connected in the day of interest. We call this approach *geometric*. To quantify the quality of link prediction, we use two standard metrics: (i) the *Area Under the Receiver Operating Characteristic curve (AUROC)*;

and (ii) the *Area Under the Precision-Recall curve (AUPR)* [91]. These metrics are described below.

The AUROC represents the probability that a randomly selected connected pair of nodes is given a higher score than a randomly selected disconnected pair of nodes in the day of interest. The degree to which the AUROC exceeds 0.5 indicates how much better the method performs than pure chance. As the name suggests, the AUROC is equal to the total area under the *Receiver Operating Characteristic (ROC)* curve. To compute the ROC curve, we order the pairs of nodes in the descending order of their scores, from the largest s_{ij} to the smallest s_{ij}, and consider each score to be a threshold. Then, for each threshold we calculate the fraction of connected pairs that are above the threshold (i.e., the True Positive Rate TPR) and the fraction of disconnected pairs that are above the threshold (i.e., the False Positive Rate FPR). Each point on the ROC curve gives the TPR and FPR for the corresponding threshold. When representing the TPR in front of the FPR, a totally random guess would result in a straight line along the diagonal $y = x$, while the degree by which the ROC curve lies above the diagonal indicates how much better the algorithm performs than pure chance. AUROC = 1 means a perfect classification (ordering) of the pairs, where the connected pairs are placed in the top of the ordered list.

The AUPR represents how accurately the method can classify pairs of nodes as connected and disconnected based on their scores. It is equal to the total area under the *Precision-Recall (PR)* curve. To compute the PR curve, we again order the pairs of nodes in the descending order of their scores, and consider each score to be a threshold. Then, for each threshold we calculate the TPR, which is called Recall, and the Precision, which is the fraction of pairs above the threshold that are connected. Each point on the PR curve gives the Precision and Recall for the corresponding threshold. A random guess corresponds to a straight line parallel to the Recall axis at the level where Precision equals the ratio of the number of connected pairs to the total number of pairs. The higher the AUPR the better the method is, while a perfect classifier yields AUPR = 1.

The results for the considered real networks and their synthetic counterparts are shown in Table 6.4. The corresponding ROC and PR curves are shown in Fig. 6.7. We see that geometric link prediction significantly outperforms chance in all cases. These results constitute another validation that the embeddings are meaningful, and illustrate that they have significant predictive power. As can be seen in Table 6.4 and Fig. 6.7, link prediction is more accurate in the synthetic counterparts. This is again expected since the counterparts are by construction maximally congruent with the underlying geometric space, while the node coordinates in them do not change over time.

We also compute the same metrics as in Table 6.4 but for a simple heuristic, where the score s_{ij} between two nodes i and j is the number of common neighbors they have in the time-aggregated network of the previous day (CN approach). The results are shown in Table 6.5. Interestingly, we see that the performance of the geometric and CN approaches is quite similar in real networks, suggesting that the latter is a good heuristic for link prediction in human proximity systems. The performance of the two approaches is also positively correlated in the synthetic counterparts (Tables 6.4 and 6.5). This is expected since the smaller the effective distance between two nodes the larger is the *expected* number of common neighbors the nodes have. However, as can be seen in Tables 6.4 and 6.5, in the counterparts the geometric approach performs better than the CN approach. This suggests that the performance of the former could be further improved in real systems, if more accurate predictions of the node coordinates in the period of interest could be made.

Network	AUROC real	AUPR real	AUROC chance	AUPR chance	AUROC synthetic	AUPR synthetic
Hospital	0.78	0.70	0.5	0.43	0.90	0.77
Primary school	0.81	0.62	0.5	0.20	0.87	0.71
Conference	0.66	0.34	0.5	0.22	0.88	0.62
High school	0.89	0.40	0.5	0.05	0.94	0.59
Office building	0.71	0.12	0.5	0.05	0.90	0.41
Friends & Family	0.86	0.60	0.5	0.10	0.93	0.72

Table 6.4: AUROC and AUPR for geometric link prediction in real networks and their synthetic counterparts. The day of interest is day 3 in the hospital and day 2 in the rest of the networks. Geometric link prediction uses the effective distances among the nodes inferred from the time-aggregated network of the previous day. "AUPR chance" corresponds to link prediction based on pure chance in the real networks. It equals the ratio of the number of connected pairs to the total number of pairs in the time-aggregated network of the day of interest. AUPR chance values for the synthetic counterparts are similar as in the real networks and not shown for brevity.

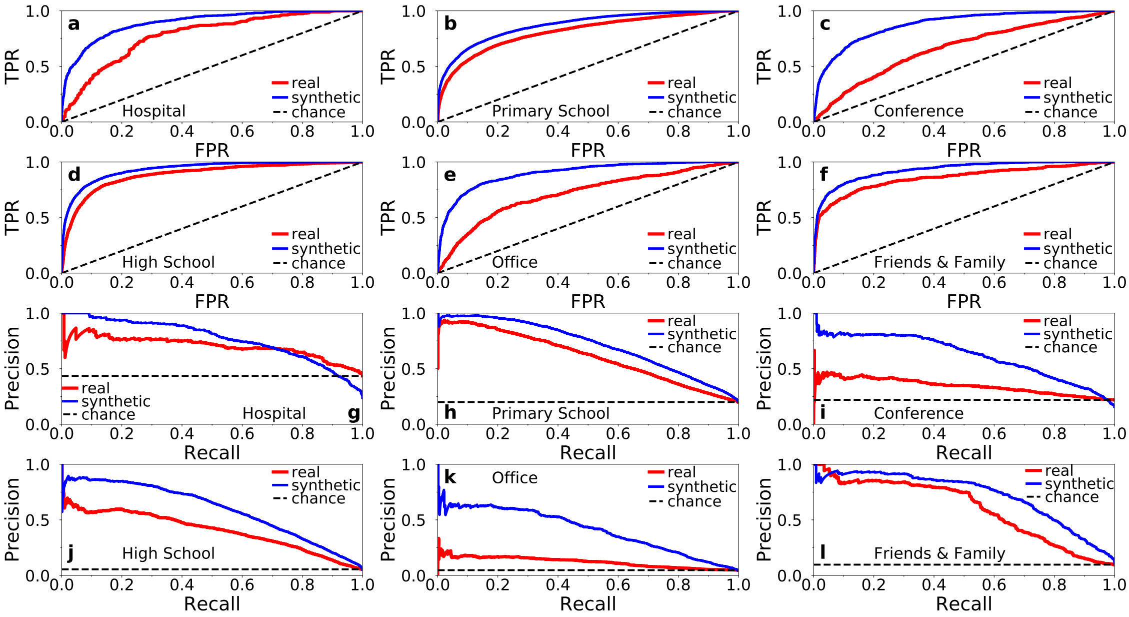

Figure 6.7: ROC and PR curves for geometric link prediction in real networks and their synthetic counterparts. (**a-f**) show the ROC curves, while (**g-l**) the PR curves, corresponding to the results in Table 6.4. The dashed black lines correspond to link prediction based on chance; these lines in (**g-l**) correspond to the AUPR chance values in Table 6.4.

Network	AUROC real	AUPR real	AUROC synthetic	AUPR synthetic
Hospital	0.75	0.79	0.85	0.69
Primary school	0.79	0.52	0.84	0.62
Conference	0.67	0.37	0.85	0.57
High school	0.88	0.44	0.89	0.52
Office building	0.73	0.10	0.86	0.35
Friends & Family	0.85	0.54	0.89	0.64

Table 6.5: Same as in Table 6.4 but for the CN approach.

6.1.6 Epidemic spreading

Finally, we consider epidemic spreading. Here, predicting the arrival time of an epidemic is crucial for developing better containment measures for infectious diseases [14, 34]. In the context of the global air transportation network, Brockmann and Helbing showed that the epidemic arrival time in a country can be well predicted by the effective distance between the country and the infection source country [14]. The effective distance between two countries is defined as the length of the shortest weighted path connecting the two countries in the air transportation network, where the weight of a link is a decreasing function of the air traffic between the endpoints of the link [14].

In a similar vein, here we show that in human proximity networks, the epidemic arrival time, i.e., the time slot at which a node becomes infected, is positively correlated with the hyperbolic distance between the node and the infected source node in the time-aggregated network. [We note that while in Ref. [14] the effective distances are directly defined by observable (weighted) path lengths, the effective distances in our case are defined by the nodes' latent coordinates that manifest themselves indirectly via the nodes' connections and disconnections in the (unweighted) time-aggregated network.] To this end, we consider the Susceptible-Infected (SI) epidemic spreading model [52]. In the SI, each node can be in one of two states, susceptible (S) or infected (I). At any time slot infected nodes infect susceptible nodes with whom they are within proximity range, with probability α. Thus, the transition of states is S→I. To simulate the SI process on temporal networks we use the dynamic SI implementation of the Network Diffusion library [90].

Figs. 6.8 and 6.9 show the results for the considered real networks and their synthetic counterparts, respectively. We see that the epidemic arrival times are significantly correlated with the hyperbolic distance from the infected source node. The correlation in each case is measured in terms of Spearman's rank correlation coefficient ρ (see Sec. 6.3.2).

Fig. 6.10 shows the spread of a single SI process during one day in each hyperbolic map of the real networks. In each network we divide the time slots in the day into 6 period of equal duration and color the nodes according to the period when they became infected. In the primary school and high school networks it is particularly clear that the infection first spreads within the community of the source infected node and then proceeds to infect adjacent communities.

These results indicate that hyperbolic embedding could provide a new perspective for understanding and predicting the behavior of epidemic spreading in human proximity systems. We leave further explorations for future work.

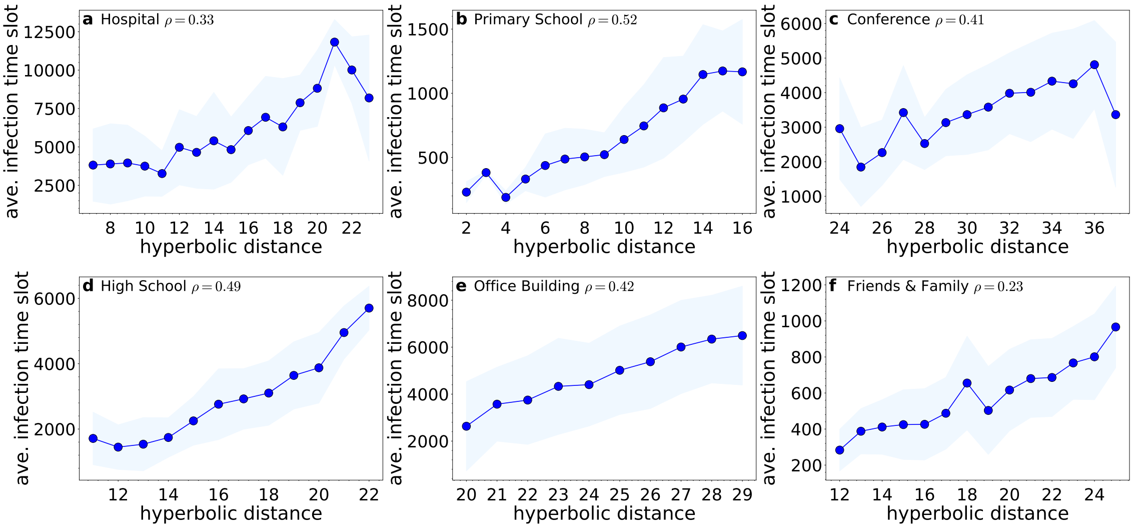

Figure 6.8: Average infection time slot as a function of the hyperbolic distance from the infected source node in real networks. In each case we consider the inferred hyperbolic distances in the time-aggregated network formed over the full observation duration. The hyperbolic distance is binned into bins of size $\delta = 1$ and the plots show the average infection time slot for nodes whose hyperbolic distance from the source node falls within each bin. The shaded area identifies the region corresponding to one standard deviation away from the average. Bins with less than 5 samples are ignored. The results are averaged over 10 simulated SI processes. Each process starts with a different infected source node selected at random, while the infection probability per time slot is $\alpha = 0.05$. Each plot indicates the average Spearman rank correlation coefficient ρ between the infection time slot and the hyperbolic distance across the 10 SI processes. In these plots we consider the hyperbolic distance instead of the equivalent effective distance $\tilde{\chi}$, as the former is more convenient for binning purposes.

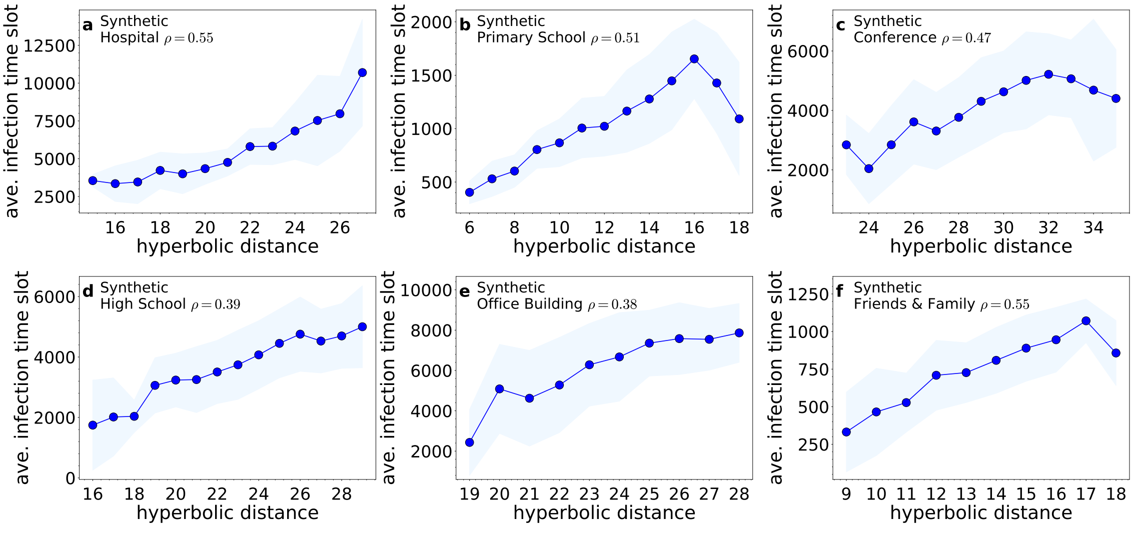

Figure 6.9: Same as in Fig. 6.8 but for the synthetic counterparts (using inferred hyperbolic distances).

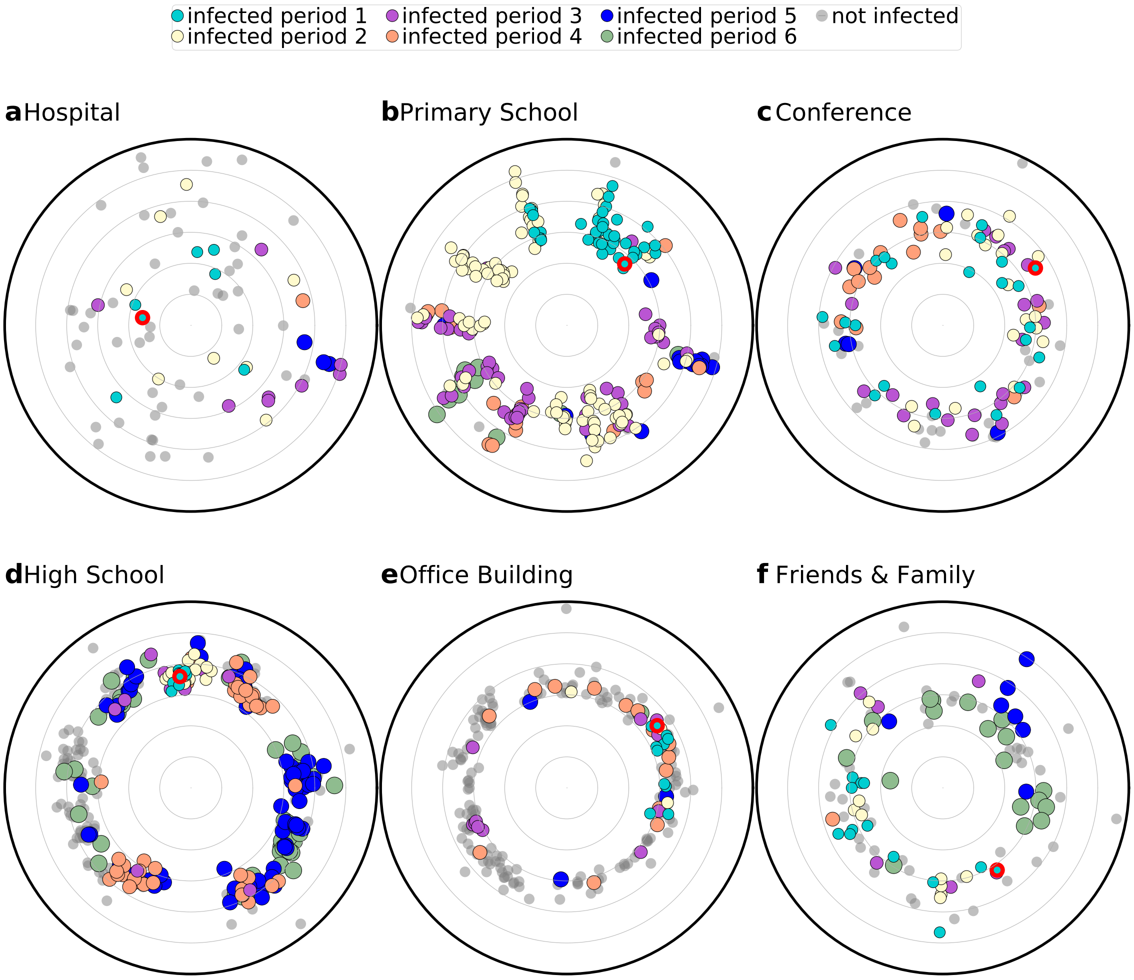

Figure 6.10: Evolution of the SI process in the hyperbolic embeddings of the real human proximity networks. (**a**) Hospital, (**b**) Primary School, (**c**) Conference, (**d**) High School, (**e**) Office Building, (**f**) Friends & Family. The nodes are positioned according to their inferred hyperbolic coordinates as in Fig. 6.4 and Appendix C.3. A single SI process is simulated on each network during the first day of observation with a single source infected node (marked with a red outline). The time slots considered in the SI process of each network are divided into six periods of equal duration and each period is assigned a different color. Nodes are colored according to the period when they became infected. The nodes were not infected are colored with a shaded gray color.

6.2 Discussion

Individual snapshots of human proximity networks are often very sparse, consisting of a small number of interacting nodes. Nevertheless, we have shown that meaningful hyperbolic embeddings of such systems are still possible. Our approach is based on embedding the time-aggregated network of such systems over an adequately large observation period, using mapping methods developed for traditional complex networks. We have justified this approach by showing that the connection probability in the time-aggregated network is compatible with the Fermi-Dirac connection probability in random hyperbolic graphs, on which existing embedding methods are based. From an applications' perspective, we have shown that the hyperbolic maps of real proximity systems can be used to identify communities, facilitate efficient greedy routing on the temporal network, and predict future links. Further, we have shown that epidemic arrival times in the temporal network are positively correlated with the distance from the infection sources in the maps. Overall, our work opens the door for a geometric description of human proximity systems.

6.3 Methods

6.3.1 Mercator

Mercator [32] combines the Laplacian Eigenmaps (LE) approach of Ref. [70] with maximum likelihood estimation (MLE) to produce fast and accurate embeddings. It can embed networks with arbitrary degree distributions. In a nutshell, Mercator takes as input the network's adjacency matrix. It infers the nodes' latent degrees ($\tilde{\kappa}$) using the nodes' observed degrees in the network and the connection probability in the $\mathbb{S}^1$ model. To infer the nodes' angular coordinates (θ), Mercator first utilizes the LE approach adjusted to the $\mathbb{S}^1$ model, in order to determine initial angular coordinates for the nodes. These initial angular coordinates are then refined using MLE, which adjusts the angular coordinates by maximizing the probability that the given network is produced by the $\mathbb{S}^1$ model. Mercator also estimates the value of the temperature parameter T. The code implementing Mercator is made publicly available by the authors of [32] at https://github.com/networkgeometry/mercator. We have used the code as is without any modifications.

We also considered a modified version of Mercator that replaces the connection probability of the $\mathbb{S}^1$ model in Eq. (5.1), Chapter 5 with the connection probability in Eq. (6.4). This modification requires several changes to the original Mercator implementation that we describe in Appendix C.7.

6.3.2 Epidemic arrival time and hyperbolic distance correlation

To quantify the correlation between the time slot at which a node becomes infected and its hyperbolic distance from the infected source node, we use Spearman's rank correlation coefficient ρ [98]. Formally, given n values X_i, Y_i, the values are converted to ranks rg_{X_i}, rg_{Y_i}, and Spearman's ρ is computed as

$$\rho = \frac{\operatorname{cov}(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}}, \tag{6.8}$$

where $\operatorname{cov}(rg_X, rg_Y)$ is the covariance of the rank variables, while $\sigma_{rg_X}, \sigma_{rg_Y}$ are the standard deviations of the rank variables. Spearman's ρ takes values between -1 and 1, and assesses monotonic relationships. $\rho = 1$ ($\rho = -1$) occurs when there is a perfect monotonic increasing (decreasing) relationship between variables X and Y, while $\rho = 0$ indicates that there is no tendency for Y to either increase or decrease when X increases.

Chapter 7
Conclusions

In this thesis we uncovered mechanisms responsible for the observed properties of real human proximity networks, including the puzzling recurrent formation of groups of the same people. In Chapter 4, we have shown that these groups can be formed by modeling human motion patterns with motion equations akin to molecular dynamics in a model of mobile agents. In the model, hidden similarities among the agents are the forces that direct their motion towards each other and determine the duration of their interactions. In this regard, similar techniques from physics could shed further insights about the dynamics of human proximity networks, for example what are the motion equations that govern the motion of the agents in the latent space? In other words, can social influence [57] (agents becoming more similar or dissimilar due to their interactions in the physical space) be modeled with already known motion equations? Another direction is extending the model with the addition of static nodes that exist both in the physical and in the latent space that represent locations rather than people. Can the same properties of these networks be reproduced if we assume that similar people are attracted to similar locations?

In Chapter 5, we have proposed a minimal latent space model, where a latent hyperbolic space abstracts the popularity of the nodes as well as the similarities among them [79]. The model generates network snapshots and in each snapshot nodes with smaller (larger) effective/hyperbolic distance are more (less) likely to be connected. This minimal model reproduces the main properties of human proximity networks as well as the complex formation of recurrent components. The simplicity of the model allows for mathematical analysis, we have proven three main properties of the model but other properties could be proven in future work. We have demonstrated that the model can be used to simulate realistic dynamics processes that run on generated synthetic human proximity networks and it is simple to use. Further work can be done in this area, a current hot topic is the spread of diseases transmitted through close contacts.

Finally, our results with the embedding of real human proximity networks further indicate that the node coordinates change over time in the hyperbolic space. Thus, a challenging yet promising task is to identify the stochastic differential equations that dictate this motion of nodes. Such equations would allow us to make predictions about the future positions of nodes in their hyperbolic spaces over different timescales. This, in turn, could allow us to improve the performance of tasks such as greedy routing and link prediction. This problem is relevant not only for human proximity systems, but for all complex networks where the hyperbolic node coordinates are expected to change over time, such as in social networks and the Internet [75]. Another problem is to extend existing hyperbolic embedding methods so that they can refine the nodes' coordinates on a snapshot-by-snapshot basis as new snapshots become available, without having to recompute each time a new embedding from scratch. Such methods could be based on the idea that a local change in the system (new connections or disconnections) should involve mostly the neighborhood (coordinates of the nodes) around the change. For this purpose, techniques based on quadtree structures as in Ref. [59] appear promising. Further, one might want to penalize large displacements based on the idea that the coordinates should be changing gradually from snapshot to snapshot. To this end, Gaussian transition models for the coordinates as in Ref. [53] seem appropriate. Methods for dynamic embedding in hyperbolic spaces should be useful not only for human proximity systems, but for temporal networks in general.

Taken altogether, our results pave the way towards more realistic modeling of human proximity networks with minimal models capable of reproducing even non-trivial social group dynamics, which are crucial for understanding and predicting the behavior of different fast-evolving processes on the networks. For example, determining the structural and dynamical properties of the networks that can affect the spread of diseases and information, is an important task to device more efficient containment and navigation strategies [1, 9, 16, 24, 41, 42, 44, 48].

Bibliography

[1] Aharony, N. et al. “Social fMRI: Investigating and shaping social mechanisms in the real world”. In: *Pervasive and Mobile Computing* vol. 7, no. 6 (2011), pp. 643–659.

[2] Alanis-Lobato, G., Mier, P., and Andrade-Navarro, M. A. “Manifold learning and maximum likelihood estimation for hyperbolic network embedding.” In: *Appl. Netw. Sci.* vol. 1, no. 10 (2016), p. 10.

[3] Alanis-Lobato, G., Mier, P., and Andrade-Navarro, M. A. “The latent geometry of the human protein interaction network”. In: *Bioinformatics* vol. 34, no. 16 (Apr. 2018), pp. 2826–2834.

[4] Allard, A. and Serrano, M. Á. “Navigable maps of structural brain networks across species”. In: *PLOS Computational Biology* vol. 16, no. 2 (Feb. 2020), pp. 1–20.

[5] Andrianou, X., Makris, K., and Konstantinou, C. “[Survery of daily contacts in Cyprus during the COVID-19 pandemic]”. Unpublished dataset. 2020.

[6] Arnold, T. A. and Emerson, J. W. “Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions”. In: *The R Journal* vol. 3, no. 2 (2011), pp. 34–39.

[7] Backer, J., Klinkenberg, D., and Wallinga, J. “Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020”. In: *Eurosurveillance* vol. 25 (Feb. 2020).

[8] Barbosa, H. et al. “Human mobility: Models and applications”. In: *Physics Reports* vol. 734 (2018). Human mobility: Models and applications, pp. 1–74.

[9] Barrat, A. and Cattuto, C. “Face-to-face interactions”. In: *Social Phenomena: From Data Analysis to Models*. Springer, Cham, 2015, pp. 37–57.

[10] Barrat, A. and Cattuto, C. “Face-to-face interactions”. In: *Social Phenomena: From Data Analysis to Models*. Springer, New York, 2015, pp. 37–57.

[11] Bläsius, T. et al. “Efficient Embedding of Scale-Free Graphs in the Hyperbolic Plane”. In: *IEEE/ACM Transactions on Networking* vol. 26, no. 2 (2018), pp. 920–933.

[12] Boguñá, M., Papadopoulos, F., and Krioukov, D. “Sustaining the Internet with hyperbolic mapping”. In: *Nature Communications* vol. 1 (Sept. 2010), p. 62.

[13] Boguñá, M. and Pastor-Satorras, R. “Class of correlated random networks with hidden variables”. In: *Phys Rev E* vol. 68, no. 3 (Sept. 2003), p. 036112.

[14] Brockmann, D. and Helbing, D. “The Hidden Geometry of Complex, Network-Driven Contagion Phenomena”. In: *Science* vol. 342, no. 6164 (2013), pp. 1337–1342.

[15] Carcione, J. M. et al. “A Simulation of a COVID-19 Epidemic Based on a Deterministic SEIR Model”. In: *Frontiers in Public Health* vol. 8 (2020), p. 230.

[16] Chaintreau, A. et al. “Impact of Human Mobility on Opportunistic Forwarding Algorithms”. In: *IEEE Transactions on Mobile Computing* vol. 6, no. 6 (June 2007), pp. 606–620.

[17] Chung, F. and Lu, L. “The average distances in random graphs with given expected degrees”. In: *Proceedings of the National Academy of Sciences* vol. 99, no. 25 (2002), pp. 15879–15882.

[18] Conan, V., Leguay, J., and Friedman, T. “Characterizing Pairwise Inter-contact Patterns in Delay Tolerant Networks”. In: *Proceedings of the International Conference on Autonomic Computing and Communication Systems*. Autonomics ’07. Rome, Italy: ICTS, 2007, 19:1–19:9.

[19] *Contact duration.* http://www.sociopatterns.org/2008/10/contact-duration/. Accessed: 2019-11-8.

[20] Conti, M. and Giordano, S. “Mobile ad hoc networking: milestones, challenges, and new research directions”. In: *IEEE Communications Magazine* vol. 52, no. 1 (Jan. 2014), pp. 85–96.

[21] Cui, P. et al. “A Survey on Network Embedding”. In: *IEEE Transactions on Knowledge and Data Engineering* vol. 31, no. 5 (2019), pp. 833–852.

[22] Daley, D. J. and Kendall, D. G. “Stochastic Rumours”. In: *IMA Journal of Applied Mathematics* vol. 1, no. 1 (Mar. 1965), pp. 42–55.

[23] Diseases, E. C. for and Protection. *COVID-19 clusters and outbreaks in occupational settings in the EU/EEA and the UK.* Technical documents. 2020, 17 p.

[24] Dong, W., Lepri, B., and Pentland, A. “Modeling the Co-evolution of Behaviors and Social Relationships Using Mobile Phone Data”. In: *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia.* MUM ’11. New York, USA: ACM, 2011, pp. 134–143.

[25] Dorogovtsev, S. N. *Lectures on Complex Networks.* Oxford: Oxford University Press, 2010.

[26] Eagle, N. and (Sandy) Pentland, A. “Reality Mining: Sensing Complex Social Systems”. In: *Personal Ubiquitous Comput.* vol. 10, no. 4 (Mar. 2006), pp. 255–268.

[27] Fang, Y., Nie, Y., and Penny, M. “Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: A data-driven analysis”. In: *Journal of Medical Virology* vol. 92, no. 6 (2020), pp. 645–659. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmv.25750.

[28] Firth, J. et al. “Using a real-world network to model localized COVID-19 control strategies”. In: *Nature Medicine* vol. 26 (Oct. 2020).

[29] Fournet, J. and Barrat, A. “Contact Patterns among High School Students”. In: *PLOS ONE* vol. 9, no. 9 (Sept. 2014), pp. 1–17.

[30] Galler, B. A. and Fisher, M. J. “An Improved Equivalence Algorithm”. In: *Commun. ACM* vol. 7, no. 5 (May 1964), pp. 301–303.

[31] Gao, W. et al. “Multicasting in Delay Tolerant Networks: A Social Network Perspective”. In: *Proceedings of the ACM International Symposium on Mobile Ad Hoc Networking and Computing*. MobiHoc ’09. New Orleans, LA, USA: ACM, 2009, pp. 299–308.

[32] García-Pérez, G. et al. “Mercator: uncovering faithful hyperbolic embeddings of complex networks”. In: *New Journal of Physics* vol. 21, no. 12 (Dec. 2019), p. 123033.

[33] Gatto, M. et al. “Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures”. In: *Proceedings of the National Academy of Sciences* vol. 117, no. 19 (2020), pp. 10484–10491. eprint: https://www.pnas.org/content/117/19/10484.full.pdf.

[34] Gauvin, L. et al. “Activity clocks: spreading dynamics on temporal networks of human contact”. In: *Scientific Reports* vol. 3 (Oct. 2013), p. 3099.

[35] Génois, M. and Barrat, A. “Can co-location be used as a proxy for face-to-face contacts?” In: *EPJ Data Science* vol. 7, no. 1 (May 2018), p. 11.

[36] Génois, M. et al. “Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers”. In: *Network Science* vol. 3 (03 Sept. 2015), pp. 326–347.

[37] Greene, D., Doyle, D., and Cunningham, P. “Tracking the Evolution of Communities in Dynamic Social Networks”. In: *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining*. ASONAM ’10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 176–183.

[38] Guan, W.-j. et al. “Clinical characteristics of 2019 novel coronavirus infection in China”. In: *medRxiv* (2020). eprint: https://www.medrxiv.org/content/early/2020/02/09/2020.02.06.20020974.full.pdf.

[39] Henderson, T., Kotz, D., and Abyzov, I. “The Changing Usage of a Mature Campus-wide Wireless Network”. In: *Computer Networks* vol. 52 (Oct. 2008), pp. 2690–2712.

[40] Holme, P. “Modern temporal network theory: A colloquium”. In: *The European Physical Journal B* vol. 88 (Aug. 2015).

[41] Holme, P. “Temporal network structures controlling disease spreading”. In: *Phys. Rev. E* vol. 94 (2 Aug. 2016), p. 022305.

[42] Holme, P. and Litvak, N. “Cost-efficient vaccination protocols for network epidemiology”. In: *PLOS Computational Biology* vol. 13, no. 9 (Sept. 2017), pp. 1–18.

[43] Holme, P. and Saramäki, J., eds. *Temporal Network Theory*. 1st ed. Springer International Publishing, 2019.

[44] Hui, P. et al. “Pocket Switched Networks and Human Mobility in Conference Environments”. In: *Proceedings of the ACM SIGCOMM Workshop on Delay-tolerant Networking*. WDTN 05. Philadelphia, Pennsylvania, USA: ACM, 2005, pp. 244–251.

[45] Isella, L. et al. “What’s in a crowd? Analysis of face-to-face behavioral networks”. In: *Journal of Theoretical Biology* vol. 271, no. 1 (2011), pp. 166–180.

[46] J.S. Allen, L. “Some discrete-time SI, SIR, and SIS epidemic models”. In: *Mathematical biosciences* vol. 124 (Dec. 1994), pp. 83–105.

[47] Jarvis, C. et al. “Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK”. In: *BMC Medicine* vol. 18 (May 2020), p. 124.

[48] Karagiannis, T., Le Boudec, J.-Y., and Vojnovic, M. “Power Law and Exponential Decay of Intercontact Times Between Mobile Devices”. In: *IEEE Transactions on Mobile Computing* vol. 9, no. 10 (Oct. 2010), pp. 1377–1390.

[49] Karsai, M. and Perra, N. “Control Strategies of Contagion Processes in Time-Varying Networks”. In: *Temporal Network Epidemiology.* Springer Singapore, 2017, pp. 179–197.

[50] Karsai, M., Perra, N., and Vespignani, A. “Time varying networks and the weakness of strong ties”. In: *Scientific Reports* vol. 4 (2014), p. 4001.

[51] Keeling, M. J. and Rohani, P. *Modeling Infectious Diseases in Humans and Animals.* Princeton University Press, 2008.

[52] Kermack, W. O., McKendrick, A. G., and Walker, G. T. “A contribution to the mathematical theory of epidemics”. In: *Proceedings of the Royal Society of London A* vol. 115, no. 772 (1927), pp. 700–721.

[53] Kim, B. et al. “A review of dynamic network models with latent variables”. In: *Statist. Surv.* vol. 12 (2018), pp. 105–135.

[54] Kleineberg, K.-K. et al. “Hidden geometric correlations in real multiplex networks”. In: *Nature Physics* vol. 12, no. 11 (July 2016), pp. 1076–1081.

[55] Krioukov, D. et al. “Hyperbolic geometry of complex networks”. In: *Phys. Rev. E* vol. 82 (3 Sept. 2010), p. 036106.

[56] Krioukov, D. et al. “Curvature and temperature of complex networks”. In: *Phys. Rev. E* vol. 80, no. 3 (Sept. 2009), p. 035101.

[57] Leenders, R. *Longitudinal behavior of network structure and actor attributes: modelling interdependence of contagion and selection.* Gordon and Breach Publishers, 1997, pp. 165–184.

[58] Li, Q. et al. “Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia”. In: *New England Journal of Medicine* vol. 382, no. 13 (2020). PMID: 31995857, pp. 1199–1207. eprint: https://doi.org/10.1056/NEJMoa2001316.

[59] Looz, M. V. and Meyerhenke, H. “Updating Dynamic Random Hyperbolic Graphs in Sublinear Time”. In: *ACM J. Exp. Algorithmics* vol. 23 (Nov. 2018), 1.6:1–1.6:30.

[60] Machens, A. et al. “An infectious disease model on empirical networks of human contact: bridging the gap between dynamic network data and contact matrices”. In: *BMC Infectious Diseases* vol. 13, no. 1 (2013), p. 185.

[61] Machens, A. et al. “An infectious disease model on empirical networks of human contact: bridging the gap between dynamic network data and contact matrices”. In: *BMC infectious diseases* vol. 13 (Apr. 2013), p. 185.

[62] Madan, A. et al. “Sensing the "Health State" of a Community”. In: *IEEE Pervasive Computing* vol. 11, no. 4 (2012), pp. 36–45.

[63] Massey, F. J. “The Distribution of the Maximum Deviation Between two Sample Cumulative Step Functions”. In: *Ann. Math. Statist.* vol. 22, no. 1 (Mar. 1951), pp. 125–128.

[64] Mastrandrea, R., Fournet, J., and Barrat, A. “Contact Patterns in a High School: A Comparison between Data Collected Using Wearable Sensors, Contact Diaries and Friendship Surveys”. In: *PLoS ONE* vol. 10, no. 9 (Sept. 2015), e0136497.

[65] Masuda, N. and Lambiotte, R. *A Guide to Temporal Networks.* World Scientific, Oct. 2016.

[66] Mazzarisi, P. et al. “A dynamic network model with persistent links and node-specific latent variables, with an application to the interbank market”. In: *arXiv:1801.00185* (2018).

[67] McLachlan, G. J. and Peel, D. *Finite mixture models.* New York: Wiley Series in Probability and Statistics, 2000.

[68] Mossong, J. et al. “Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases”. In: *PLOS Medicine* vol. 5, no. 3 (Mar. 2008), pp. 1–1.

[69] Muscoloni, A. and Cannistraci, C.-V. “A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities”. In: *New Journal of Physics* vol. 20, no. 5 (2018).

[70] Muscoloni, A. et al. “Machine learning meets complex networks via coalescent embedding in the hyperbolic space”. In: *Nature Communications* vol. 8, no. 1 (2017), p. 1615.

[71] Ogura, M. and M. Preciado, V. “Optimal Containment of Epidemics in Temporal and Adaptive Networks”. In: *Temporal Network Epidemiology.* Springer Singapore, 2017, pp. 241–266.

[72] Olver, F. W. et al. *NIST Handbook of Mathematical Functions.* 1st. New York, USA: Cambridge University Press, 2010.

[73] Organization, W. H. *Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19).* Technical documents. 2020, 40 p.

[74] Ortiz, E., Starnini, M., and Serrano, M. Á. “Navigability of temporal networks in hyperbolic space”. In: *Scientific Reports* vol. 7, no. 1 (2017), p. 15054.

[75] Papadopoulos, F., Aldecoa, R., and Krioukov, D. “Network geometry inference using common neighbors”. In: *Phys. Rev. E* vol. 92 (2 2015), p. 022807.

[76] Papadopoulos, F. and Kleineberg, K.-K. “Link persistence and conditional distances in multiplex networks”. In: *Phys. Rev. E* vol. 99 (1 Jan. 2019), p. 012322.

[77] Papadopoulos, F., Psomas, C., and Krioukov, D. “Network Mapping by Replaying Hyperbolic Growth”. In: *IEEE/ACM Transactions on Networking* vol. 23, no. 1 (2015), pp. 198–211.

[78] Papadopoulos, F. and Rodríguez-Flores, M. A. “Latent geometry and dynamics of proximity networks”. In: *Phys. Rev. E* vol. 100 (5 Nov. 2019), p. 052313.

[79] Papadopoulos, F. et al. “Popularity versus similarity in growing networks”. In: *Nature* vol. 489 (Sept. 2012), pp. 537–540.

[80] Park, J. and Newman, M. E. J. “Statistical mechanics of networks”. In: *Phys. Rev. E* vol. 70 (6 Dec. 2004), p. 066117.

[81] Passarella, A. and Conti, M. “Analysis of Individual Pair and Aggregate Intercontact Times in Heterogeneous Opportunistic Networks”. In: *IEEE Transactions on Mobile Computing* vol. 12, no. 12 (Dec. 2013), pp. 2483–2495.

[82] Perra, N. et al. “Activity driven modeling of time varying networks”. In: *Scientific Reports* vol. 2 (June 2012), p. 469.

[83] Prem, K. et al. “The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study”. In: *The Lancet Public Health* vol. 5 (Mar. 2020).

[84] Press and Cyprus, I. O. of. *Coronavirus Disease 2019 (COVID-19) in Cyprus: Surveillance Report as of 05/05/2020.* Technical documents. 2020, 16 p.

[85] *Reality Commons by the MIT Human Dynamics Lab.* http://realitycommons.media.mit.edu/.

[86] Ridenhour, B., Kowalik, J., and Shay, D. “Unraveling R_0: Considerations for Public Health Applications”. In: *American journal of public health* vol. 104 (Dec. 2013).

[87] Rodríguez-Flores, M. A. and Papadopoulos, F. “Similarity Forces and Recurrent Components in Human Face-to-Face Interaction Networks”. In: *Phys. Rev. Lett.* vol. 121 (25 Dec. 2018), p. 258301.

[88] Rodríguez-Flores, M. A. and Papadopoulos, F. “Hyperbolic Mapping of Human Proximity Networks”. In: *Scientific Reports* vol. 10 (Nov. 2020), p. 20244.

[89] Rohlf, F. J. and Slice, D. “Extensions of the Procrustes Method for the Optimal Superimposition of Landmarks”. In: *Systematic Biology* vol. 39, no. 1 (Mar. 1990), pp. 40–59.

[90] Rossetti, G. et al. “NDlib: a python library to model and analyze diffusion processes over complex networks”. In: *International Journal of Data Science and Analytics* vol. 5, no. 1 (2018), pp. 61–79.

[91] Saito, T. and Rehmsmeier, M. “Precrec: fast and accurate precision–recall and ROC curve calculations in R”. In: *Bioinformatics* vol. 33, no. 1 (Sept. 2016), pp. 145–147.

[92] Scherrer, A. et al. “Description and simulation of dynamic mobility networks”. In: *Computer Networks* vol. 52, no. 15 (2008), pp. 2842–2858.

[93] Schlick, T. *Molecular Modeling and Simulation: An Interdisciplinary Guide.* Interdisciplinary Applied Mathematics. Springer, New York, 2010.

[94] Sekara, V., Stopczynski, A., and Lehmann, S. “Fundamental structures of dynamic social networks”. In: *Proceedings of the National Academy of Sciences* vol. 113, no. 36 (2016), pp. 9977–9982.

[95] Serrano, M. Á., Krioukov, D., and Boguñá, M. “Self-Similarity of Complex Networks and Hidden Metric Spaces”. In: *Phys. Rev. Lett.* vol. 100 (7 Feb. 2008), p. 078701.

[96] Smieszek, T. “A mechanistic model of infection: why duration and intensity of contacts should be included in models of disease spread”. In: *Theoretical Biology and Medical Modelling* vol. 6, no. 1 (2009), p. 25.

[97] *SocioPatterns*. http://www.sociopatterns.org/.

[98] Spearman, C. “The Proof and Measurement of Association between Two Things”. In: *The American Journal of Psychology* vol. 15, no. 1 (1904), pp. 72–101.

[99] Starnini, M., Baronchelli, A., and Pastor-Satorras, R. “Model reproduces individual, group and collective dynamics of human contact networks”. In: *Social Networks* vol. 47 (2016), pp. 130–137.

[100] Starnini, M., Baronchelli, A., and Pastor-Satorras, R. “Modeling Human Dynamics of Face-to-Face Interaction Networks”. In: *Phys. Rev. Lett.* vol. 110 (16 Apr. 2013), p. 168701.

[101] Starnini, M. et al. “Robust Modeling of Human Contact Networks Across Different Scales and Proximity-Sensing Techniques”. In: *Social Informatics*. Cham: Springer, 2017, pp. 536–551.

[102] Stehlé, J. et al. “High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School”. In: *PLoS ONE* vol. 6, no. 8 (Aug. 2011), e23176.

[103] Stehlé, J., Barrat, A., and Bianconi, G. “Dynamical and bursty interactions in social networks”. In: *Phys. Rev. E* vol. 81 (3 Mar. 2010), p. 035101.

[104] Takaguchi, T. et al. “Predictability of Conversation Partners”. In: *Phys. Rev. X* vol. 1 (1 Sept. 2011), p. 011008.

[105] Thurner, S., Klimek, P., and Hanel, R. “A network-based explanation of why most COVID-19 infection curves are linear”. In: *Proceedings of the National Academy of Sciences* vol. 117, no. 37 (2020), pp. 22684–22689. eprint: https://www.pnas.org/content/117/37/22684.full.pdf.

[106] Torricelli, M., Karsai, M., and Gauvin, L. “weg2vec: Event embedding for temporal networks”. In: *Scientific Reports* vol. 10, no. 1 (2020), p. 7164.

[107] Vanhems, P. et al. “Estimating Potential Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors”. In: *PLoS ONE* vol. 8, no. 9 (Sept. 2013), e73970.

[108] Vazquez, A. et al. “Impact of Non-Poissonian Activity Patterns on Spreading Processes”. In: *Phys. Rev. Lett.* vol. 98 (15 Apr. 2007), p. 158702.

[109] Vestergaard, C. L., Génois, M., and Barrat, A. “How memory generates heterogeneous dynamics in temporal networks”. In: *Phys. Rev. E* vol. 90 (4 Oct. 2014), p. 042805.

[110] *Worldometer: Coronavirus in Cyprus*. https://www.worldometers.info/coronavirus/country/cyprus/. Accessed: 2020-12-15.

[111] Zhang, J. et al. “Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China”. In: *Science* vol. 368, no. 6498 (2020), pp. 1481–1486. eprint: https://science.sciencemag.org/content/368/6498/1481.full.pdf.

[112] Zhao, K. et al. “Social network dynamics of face-to-face interactions”. In: *Phys. Rev. E* vol. 83 (5 May 2011), p. 056109.

Appendices

Appendix A

More results with the force-directed motion model

A.1 Recurrent components

A.1.1 Unique and recurrent components in real and modeled networks

Figs. A.1-A.4 show the unique and recurrent components in the real datasets and in corresponding simulated networks with the attractiveness and FDM models. For the Primary School and High School the results are shown for each activity cycle. For the Conference the results are shown for the whole duration (all activity cycles), as there were relatively few recurrent components in each individual activity cycle. Furthermore, for the Primary School we also show the results if we exclude the lunch break period in each activity cycle (12pm-2pm) where children of different classes have lunch in a common place and some children go home to have lunch [102]. Removing this period results in a more uniform pattern of recurrent components formation (Figs. A.1a,b vs. Figs. A.1c,d).

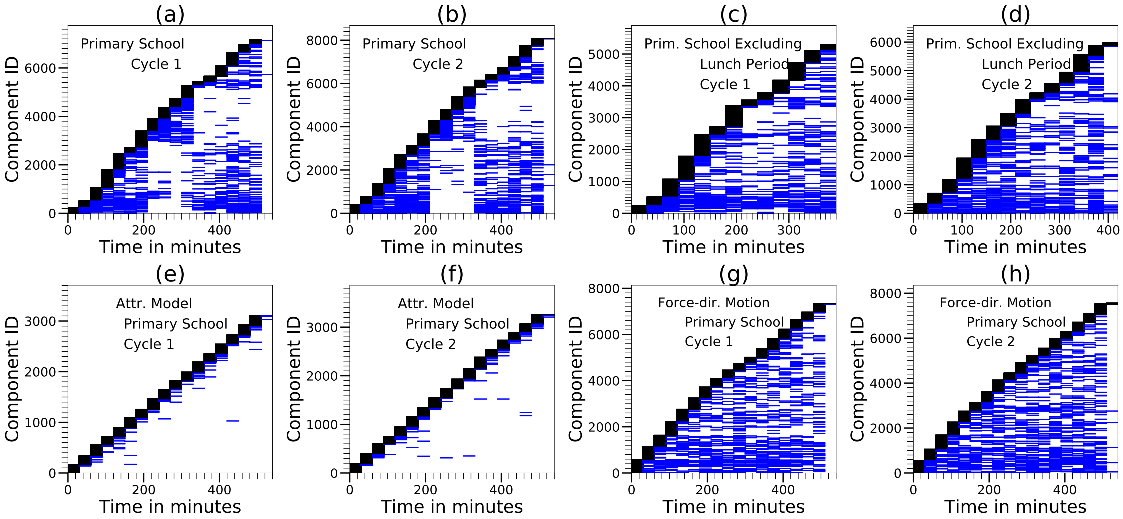

Figure A.1: **(a, b)** Unique and recurrent components found in each cycle of activity in the Primary School. **(c, d)** Same as (a, b) but excluding the lunch break period. **(e, f)** Components found in a simulation run of the attractiveness model assuming activity cycles of the same durations as in (a, b). **(g, h)** Same as (e, f) but for the FDM (Force-dir. Motion) model. All simulations use the Primary School parameters (Table 4.2 in Sec. 4.3).

Finally, in Fig. A.5 below, we plot the recurrent and unique components formed in a synthetic network of the Hospital's first cycle of activity, generated with the model of Ref. [103] that we described in Section 3.3. To generate the network we use the following parameters: number of nodes $N = 75$, number of time slots $\tau = 1100$

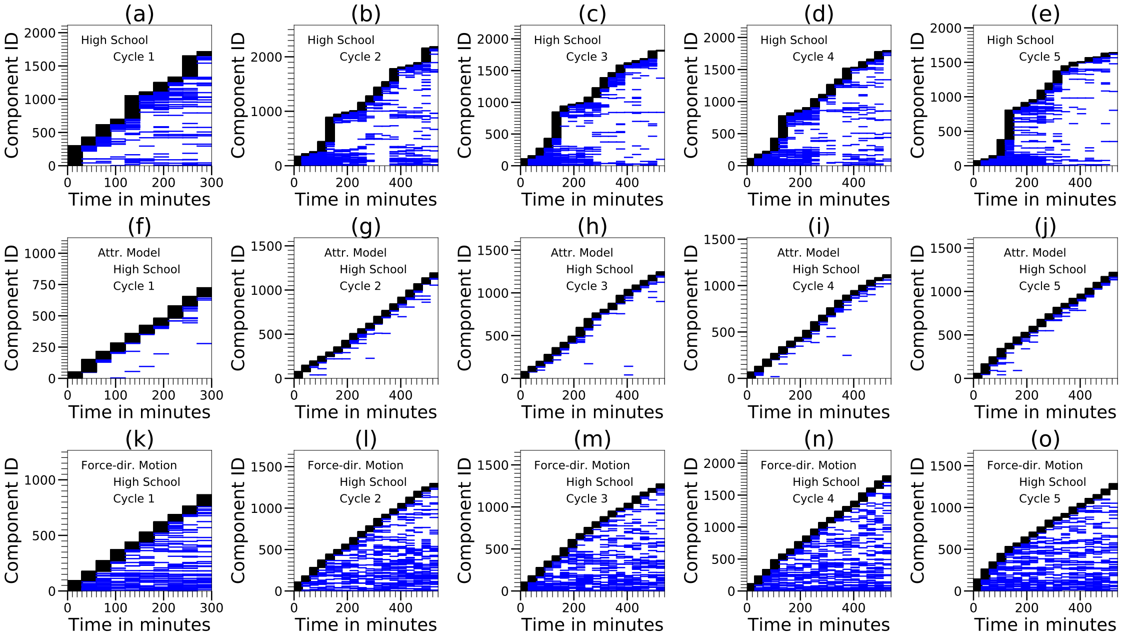

Figure A.2: **(a-e)** Unique and recurrent components found in each cycle of activity in the High School. **(f-j)** Components found in a simulation run of the attractiveness model assuming activity cycles of the same durations as in (a-e). **(k-o)** Same as (f-j) but for the FDM (Force-dir. Motion) model. All simulations use the High School parameters (Table 4.2 in Sec. 4.3).

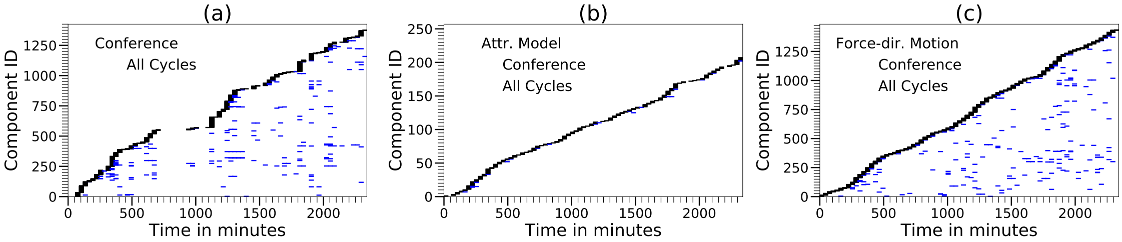

Figure A.3: **(a)** Unique and recurrent components found over the whole duration of the Conference dataset. **(b)** Components found in a simulation run of the attractiveness model with the same duration as in (a). **(c)** Same as (b) but for the FDM (Force-dir. Motion) model. All simulations use the Conference parameters (Table 4.2 in Sec. 4.3).

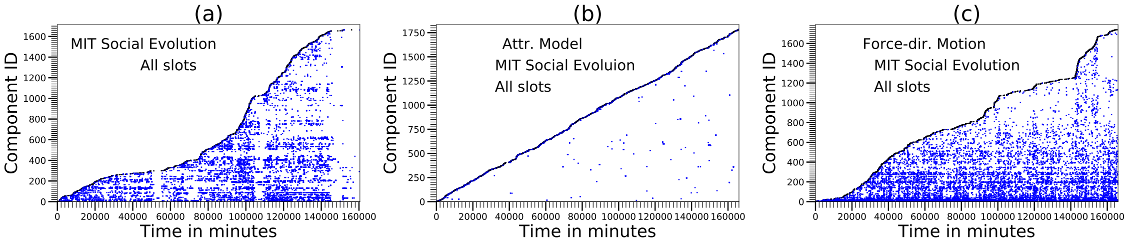

Figure A.4: **(a)** Unique and recurrent components found over the whole duration of the MIT Social Evolution dataset. **(b)** Components found in a simulation run of the attractiveness model with the same duration as in (a). **(c)** Same as (b) but for the FDM (Force-dir. Motion) model. All simulations use the MIT Social Evolution parameters (Table 4.2 in Sec. 4.3).

and parameter $\lambda = 0.64$, which we tuned to match the average number of interacting agents per time slot.

The mechanism used to form groups in this model also fails to form recurrent components in abundance, because groups are formed or grown randomly: in each snapshot the model selects a random isolated node to connect to another random isolated node or to a randomly selected group.

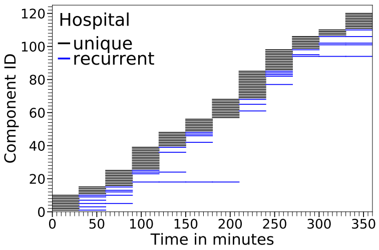

Figure A.5: Unique and recurrent components found in a synthetic network corresponding to the first cycle of activity in the Hospital. The network was generated with the model of Ref. [103], described in Section 3.3.

A.2 Other properties of real versus modeled networks

In Figs. A.6- A.9 we the properties described in Sec. 3.2 between the real networks and the corresponding simulated networks with the FDM model. We observe a good agreement between the model and reality, except for the deviation observed in the shortest time-respecting paths between the Conference and the model (Fig. A.8h). This is not due to the attraction forces in the FDM, since as we see in Fig. A.8h the attractiveness model also yields similar results to the FDM. In fact, we observe that this distribution is also very different between the Conference and the other datasets—as can be seen, in the Conference there are significantly longer paths. This difference might be due to the fact that interactions are less structured in this dataset, in the sense that participants move at will between different areas such as conference rooms, coffee break areas, etc. [45], which also justifies the fewer recurrent components in this dataset compared to the rest (see Appendix A.1).

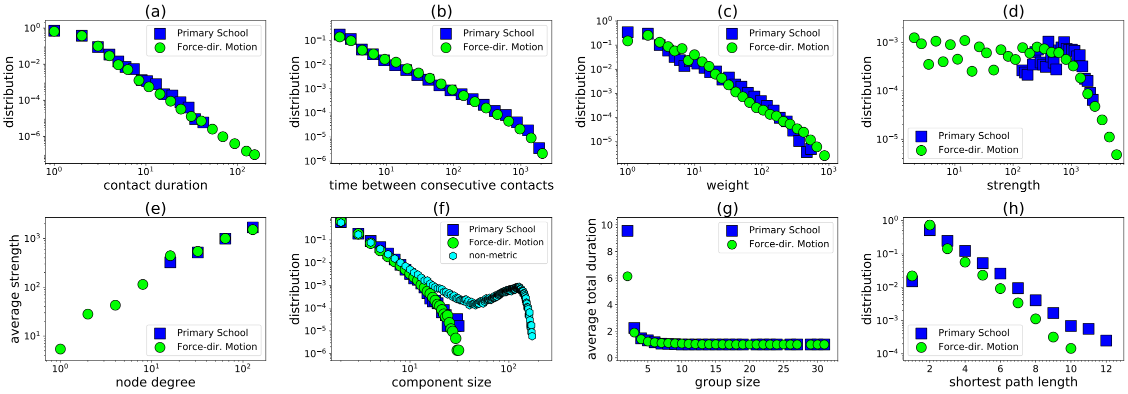

Figure A.6: Same as Fig. 4.7 but for the Primary School.

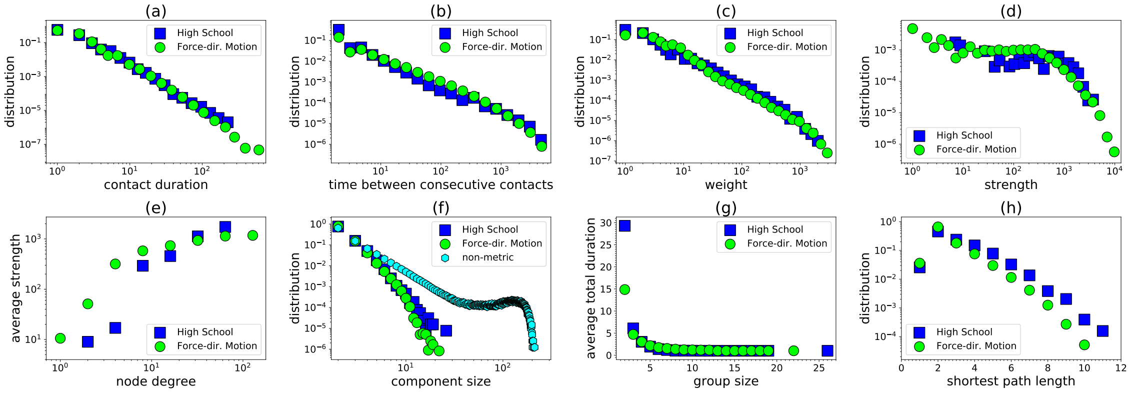

Figure A.7: Same as Fig. 4.7 but for the High School.

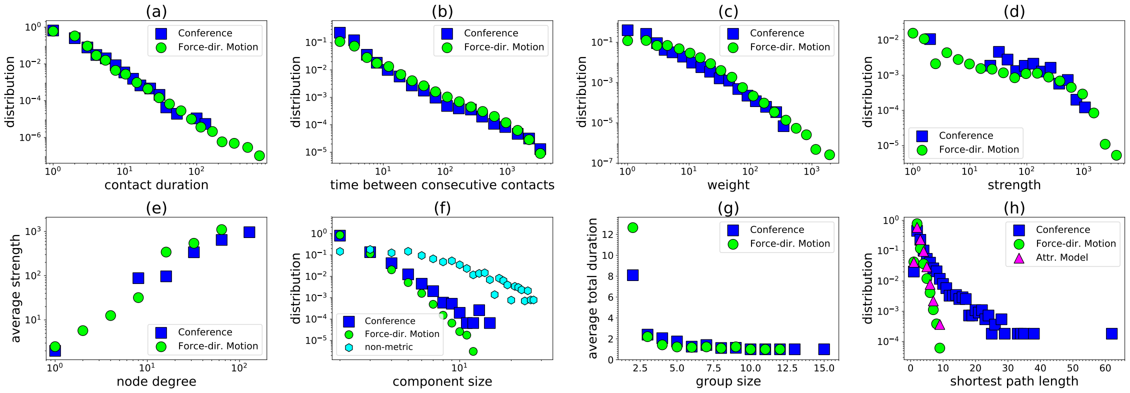

Figure A.8: Same as Fig. 4.7 but for the Conference. Plot (h) also shows the corresponding simulation results with the attractiveness model.

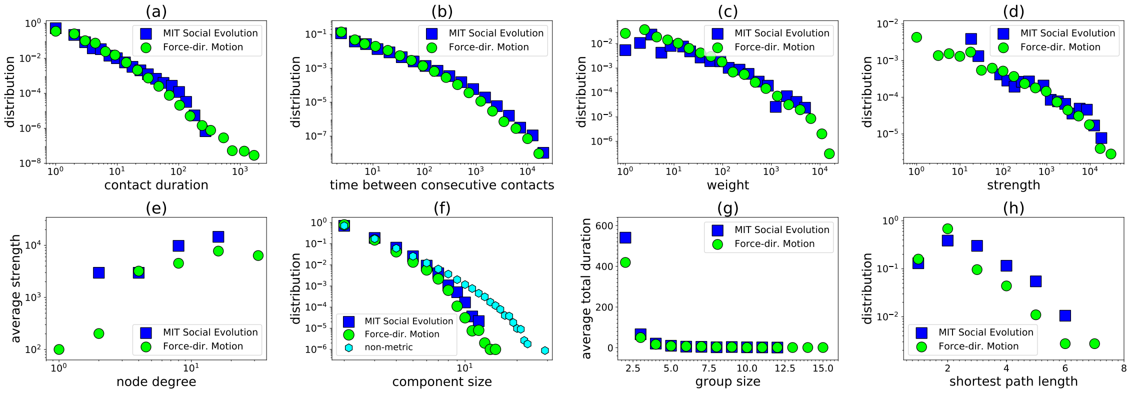

Figure A.9: Same as Fig. 4.7 but for the MIT Social Evolution. In all plots the simulation results are averages over 10 runs except from (h), which shows the results from one run as computing this metric in this large dataset is computationally expensive.

A.3 SIS spreading in real and modeled networks with the FDM

Here we provide more details on the susceptible-infected-susceptible (SIS) epidemic spreading model [51] considered in Chapter 4. In the SIS model each agent can be in one of two states at any time slot t, susceptible (S) or infected (I). At any time slot an infected agent recovers with probability β and becomes susceptible again, whereas infected agents infect the susceptible agents with whom they interact, with probability α. Therefore, the only transition of states is S → I → S.

To obtain the results in Fig. 4.3 we have used the dSIS (dynamic SIS) model for temporal networks from the Network Diffusion Library [90]. For each simulation of the process we compute the percentage of infected agents per slot, and then take the average of this percentage over the considered slots (prevalence). We consider the first activity cycle of the Hospital and Primary school and the second cycle of the High School—we consider the second cycle of the High School as its first cycle has fewer recorded slots than the rest of its cycles (Sec. 3.1). In all cases the results are similar in all activity cycles of similar durations. The corresponding simulated networks with the FDM (Table 4.2) are run (after τ_{warmup}) for the same duration as the corresponding cycles in the real networks and the prevalence is measured excluding the τ_{warmup} period. In the real networks the results are averages over 20 simulated SIS processes. The results with the FDM are averages across 10 simulated counterparts of each real network; in each counterpart the prevalence is averaged over 5 SIS processes. In all cases each SIS process has a different initial set of infected agents that consists of 10% of all agents selected at random.

In Fig. A.10 we also report prevalence results for the Conference (first activity cycle). As can be seen, in this case the SIS process performs differently than in the corresponding FDM networks. This was expected since as explained in Sec. A.2 this dataset has some different properties from the rest of the datasets we consider (cf. Fig. A.8h and the related discussion in Sec. A.2).

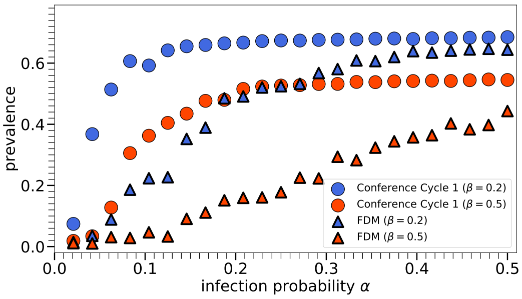

Figure A.10: Same as Fig. 4.3 but for the Conference dataset.

A.4 Network properties and recurrent components with effective distances

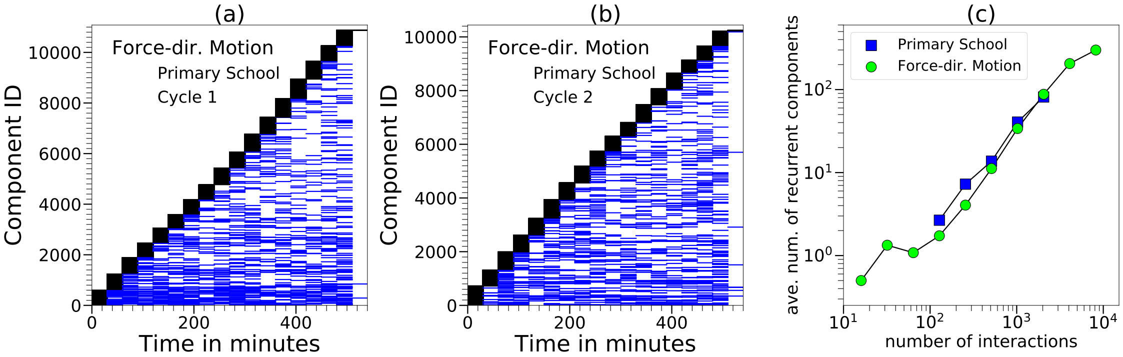

Figure A.11: Same as Fig. 4.17 but for the Primary School.

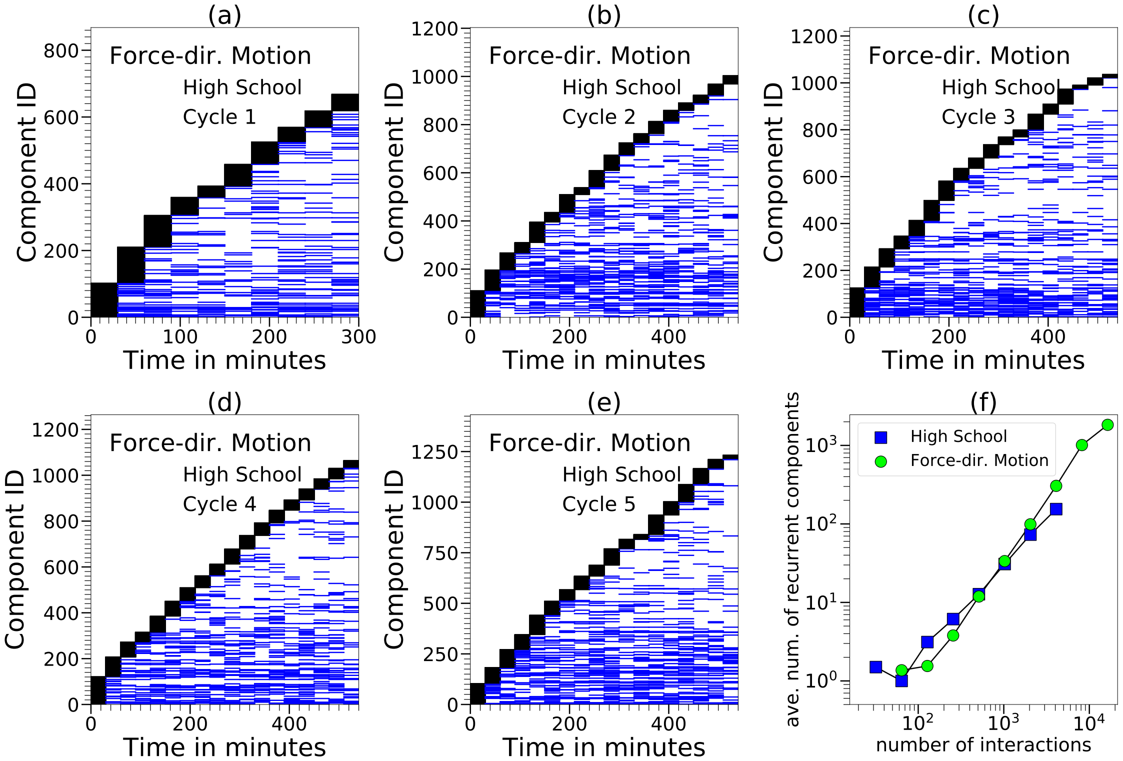

Figure A.12: Same as Fig. 4.17 but for the High School.

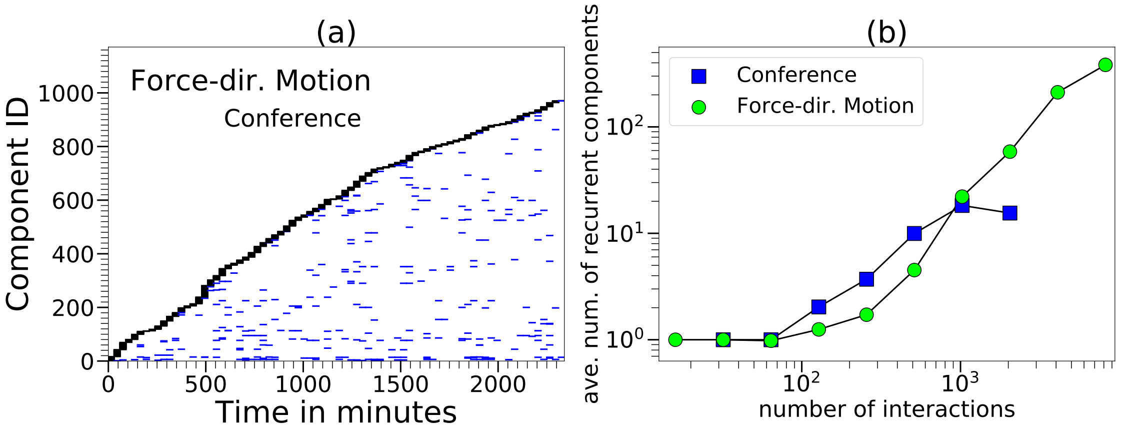

Figure A.13: Same as Fig. 4.17 but for the Conference.

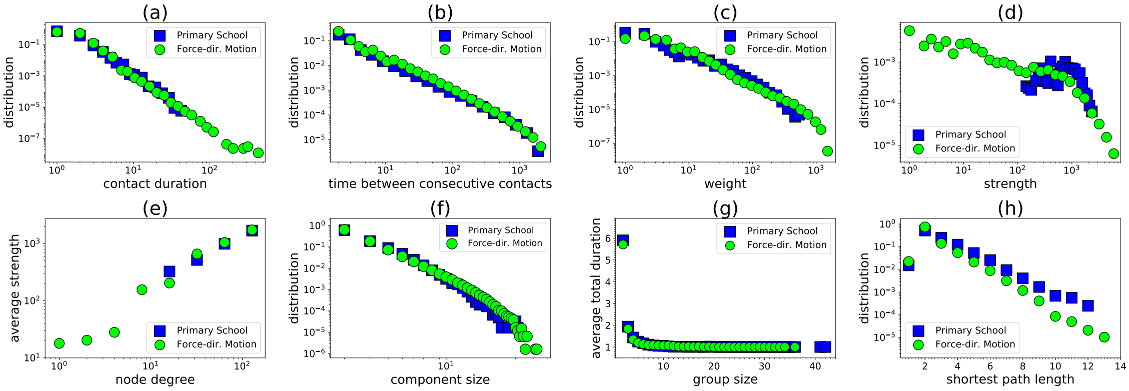

Figure A.14: Same as Fig. 4.18 but for the Primary School.

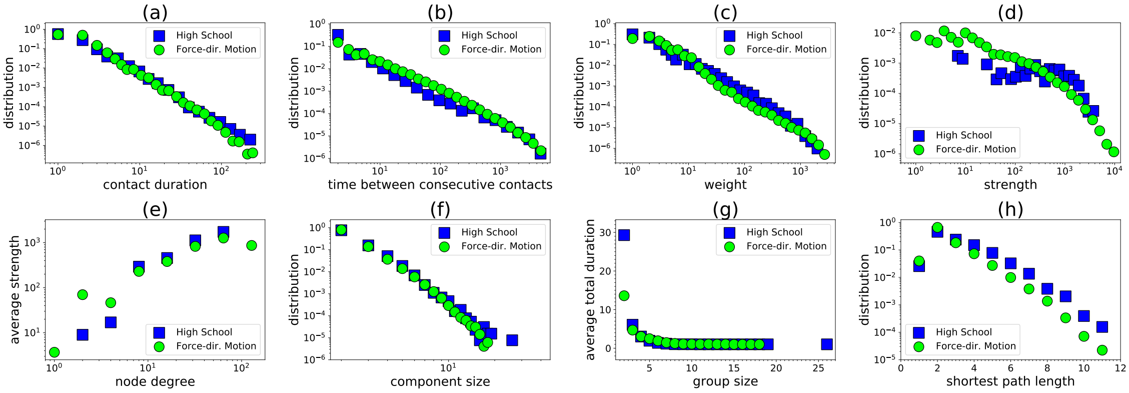

Figure A.15: Same as Fig. 4.18 but for the High School.

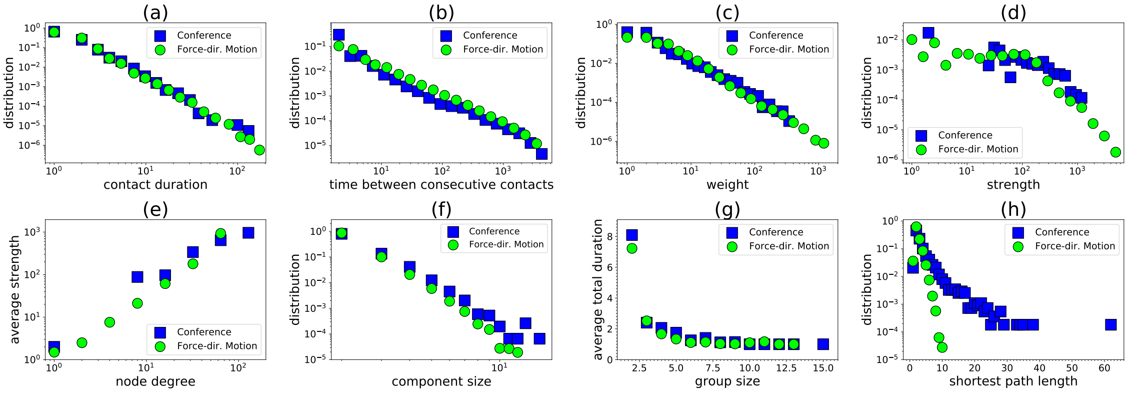

Figure A.16: Same as Fig. 4.18 but for the Conference.

Appendix B

More results with the dynamic-$\mathbb{S}^1$ model

B.1 COVID-19 SEIR simulations with non-normalized data.

In Fig. B.1 we present the properties of the synthetic networks generated with the dynamic-$\mathbb{S}^1$ from the non-normalized survey of daily contacts in Cyprus [5] (unpublished dataset). Whereas in Fig. B.2 we show the plots of total cases, daily new cases and active cases per day produced by the SEIR simulations with the COVID-19 parameters as described in Section 5.7, but using the synthetic networks generated from the non-normalized data. Fig. B.3 corresponds to the daily new cases originated in each setting: work, home or elsewhere.

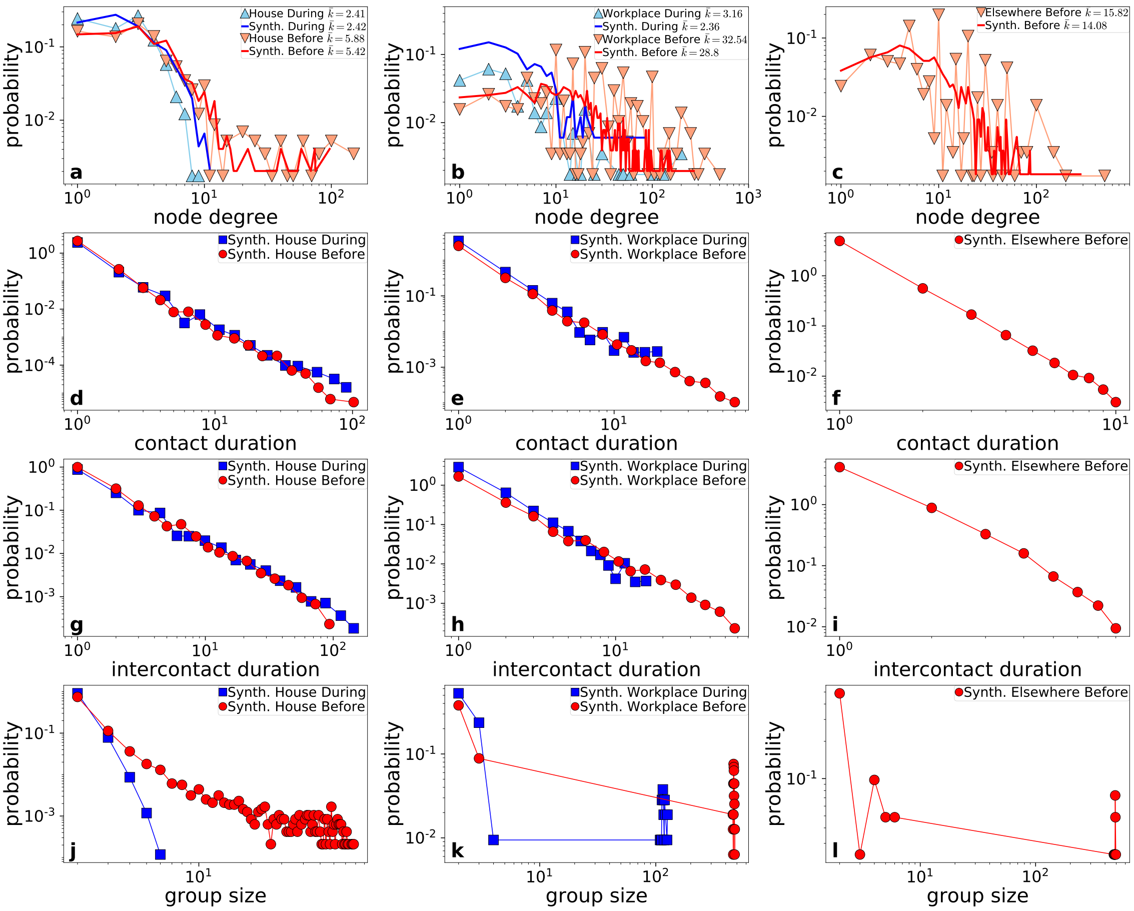

Figure B.1: Same as Fig. 5.17 but for the synthetic networks generated from the non-normalized data.

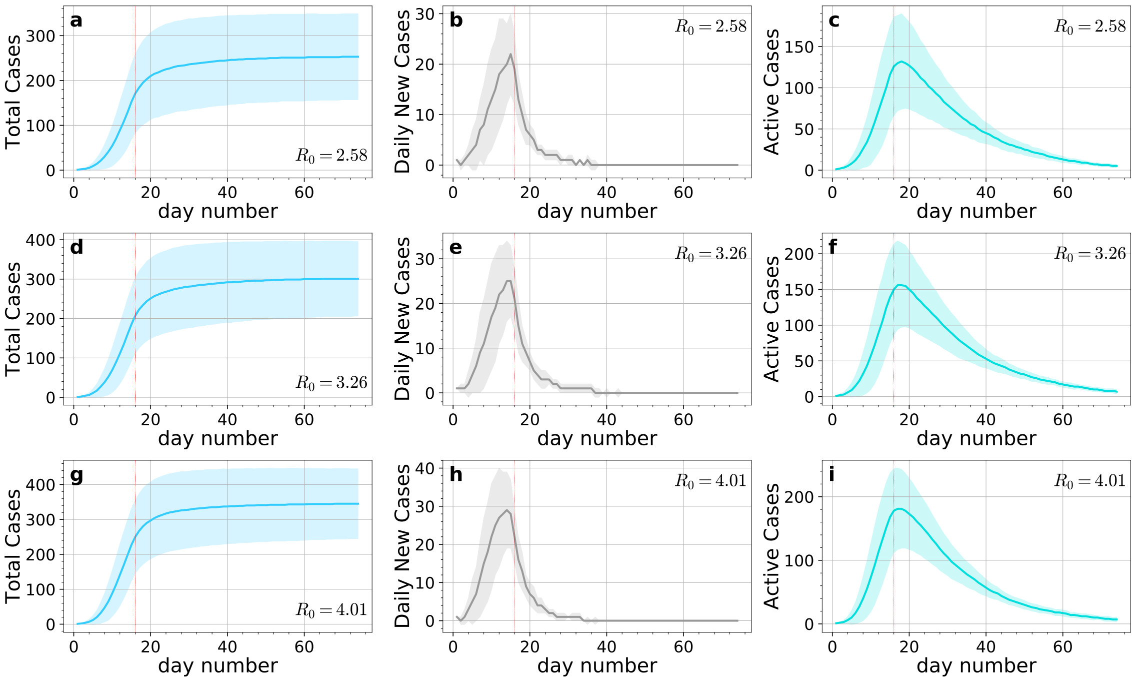

Figure B.2: Same as Fig. 5.18 but for the synthetic network generated from the non-normalize data.

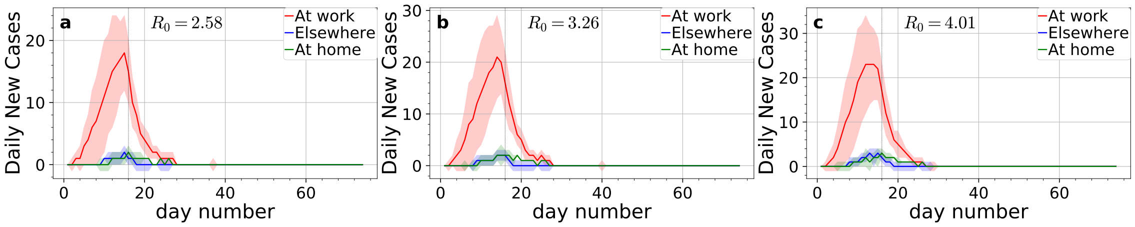

Figure B.3: Same as Fig. 5.19 but for the synthetic network generated from the non-normalized data.

B.2 dynamic-$\mathbb{S}^1$ vs. configuration model

The dynamic-$\mathbb{S}^1$ utilizes the $\mathbb{S}^1$ model at the cold regime where the temperature is $T < 1$ (Sec. 5.1). The $\mathbb{S}^1$ can be also defined at the hot regime, $T > 1$ [55].

Like traditional complex networks [55], proximity networks appear to belong to the cold regime. Indeed, as seen in Table 5.2, all counterparts have $T < 1$. Further, Fig. 5.16 shows that the number of recurrent components quickly decreases with $T \in (0, 1)$, becoming small at $T \to 1$, while real networks have large numbers of recurrent components (cf. Figs. 5.5(h), 5.6(h) and [87]).

Analyzing the dynamic-$\mathbb{S}^1$ at the hot regime is beyond the scope of this paper. However, we consider here a limiting case at this regime, where the $\mathbb{S}^1$ model degenerates to the configuration model, i.e., to the ensemble of graphs with given expected degrees [17, 80]. This case corresponds to letting $T \to \infty$, while completely ignoring the angular distances among the nodes, see [55] for details. The connection probability between two nodes i, j becomes

$$p_{\mathrm{cm}}(\kappa_i, \kappa_j) = \frac{1}{1 + N\bar{\kappa}^2/(\bar{k}\kappa_i\kappa_j)}. \tag{B.1}$$

For sparse networks ($\bar{k} \ll N$) and distributions of κ_i that are not too broad (conditions that hold in the considered networks, Fig. 5.3), we can write $p_{\mathrm{cm}}(\kappa_i, \kappa_j) \approx \bar{k}\kappa_i\kappa_j/(N\bar{\kappa}^2)$. Using this approximation, it is easy to see that the expected degree of a node with latent variable κ is given by (5.4), while the average degree in the resulting network is $\bar{k}$.

We now build synthetic counterparts for the real networks of Sec. 5.3.1 using the dynamic-$\mathbb{S}^1$ as described in Secs. 5.2 and 5.3.2, except that we connect the nodes in each snapshot with the connection probability in (B.1) [instead of (5.1)]. Since there is no temperature T in (B.1), we can no longer control the average time-aggregated degree, which is significantly larger in the counterparts, $\bar{k}_{\mathrm{aggr}} = 58, 214, 242, 76, 125$, for the hospital, primary school, high school, conference and Friends & Family, respectively (vs. the ones in Table 5.1). As expected, we see in Fig. B.4 that the configuration model cannot reproduce the abundance of recurrent components observed in the real networks. Further, it cannot capture their broad contact, inter-contact and weight distributions (Fig. B.4).

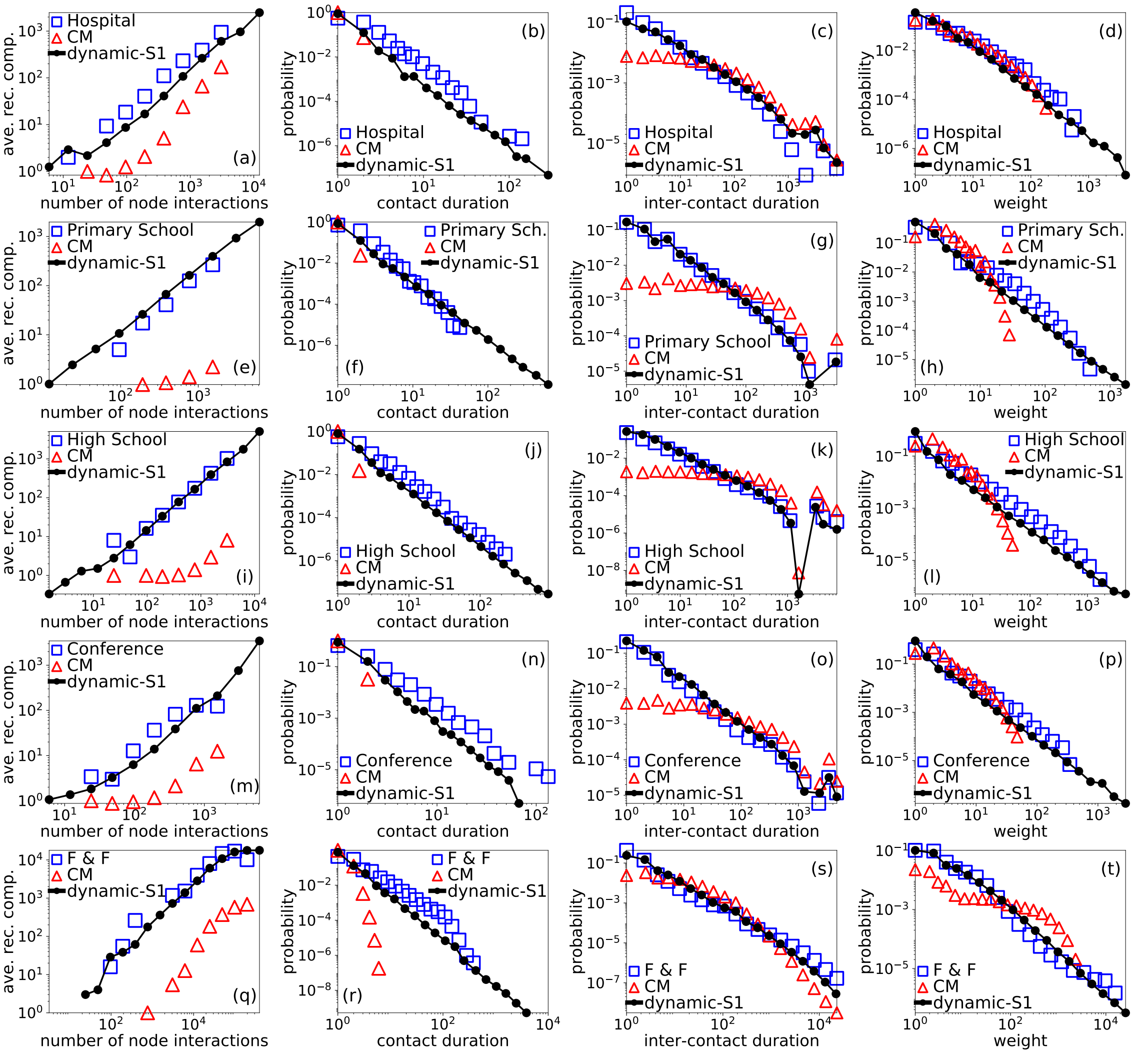

Figure B.4: Real face-to-face interaction networks vs. simulated networks with the configuration model (CM). **(a,e,i,m,q)** Average number of recurrent components where an agent participates as a function of the total number of interactions of the agent. **(b,f,j,n,r)** Contact distribution. **(c,g,k,o,s)** Inter-contact distribution. **(d,h,l,p,t)** Weight distribution. For comparison the results with the dynamic-$\mathbb{S}^1$ considered in Chapter 5 are also shown. The results with the models are averages over 20 simulation runs except from the Friends & Family where the averages are over 5 runs.

Appendix C

More results with hyperbolic embeddings of human proximity networks

C.1 Connection probability in the time-aggregated network

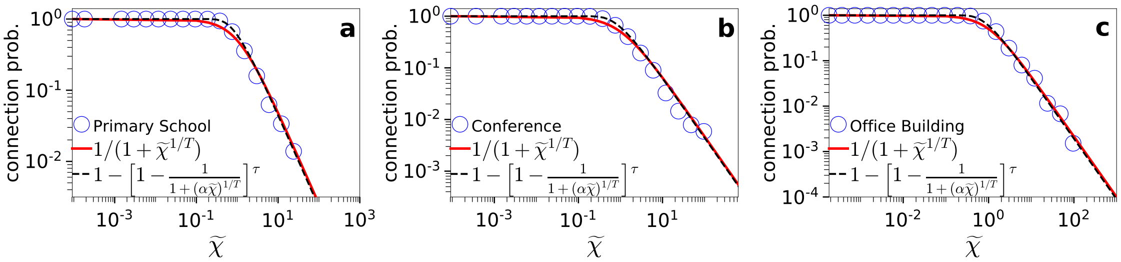

Figure C.1: **Connection probability in the time-aggregated network versus Fermi-Dirac connection probability.** Same as in Fig. 6.1 but for the synthetic counterparts of the primary school, conference and office building.

C.2 Inference of latent coordinates with the original and modified Mercator

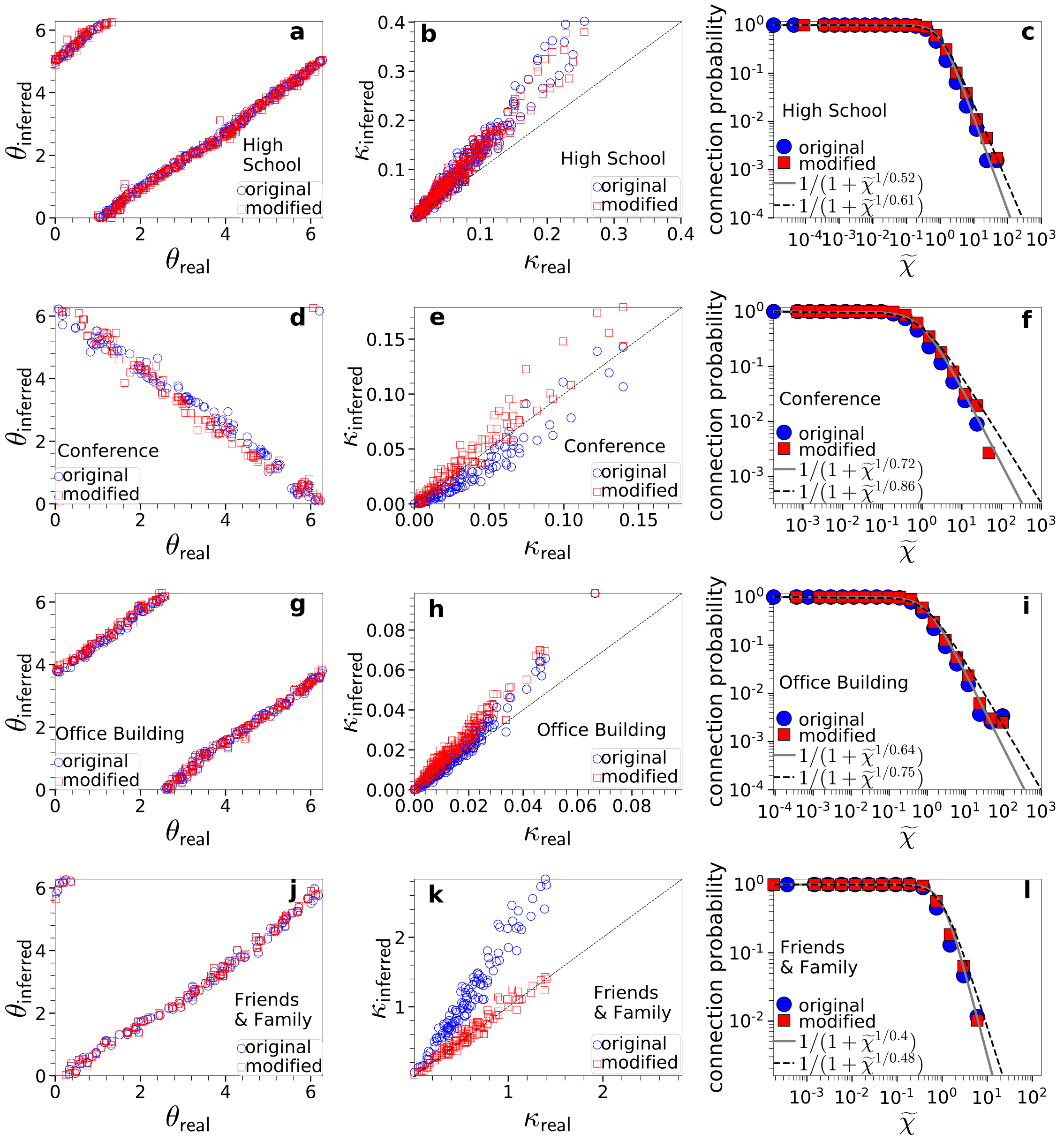

Figure C.2: **Inference of latent coordinates (κ, θ) with the original and modified versions of Mercator.** Same as in Fig. 6.2 but for the synthetic counterparts of the high school, conference, office building and Friends & Family. For the four networks, the original version estimates $T = 0.52, 0.72, 0.64$ and 0.4, the modified version estimates $T = 0.61, 0.86, 0.75$ and 0.48, while the actual values are $T = 0.61, 0.85, 0.74$ and 0.48, respectively.

C.3 Hyperbolic maps of the conference and Friends & Family

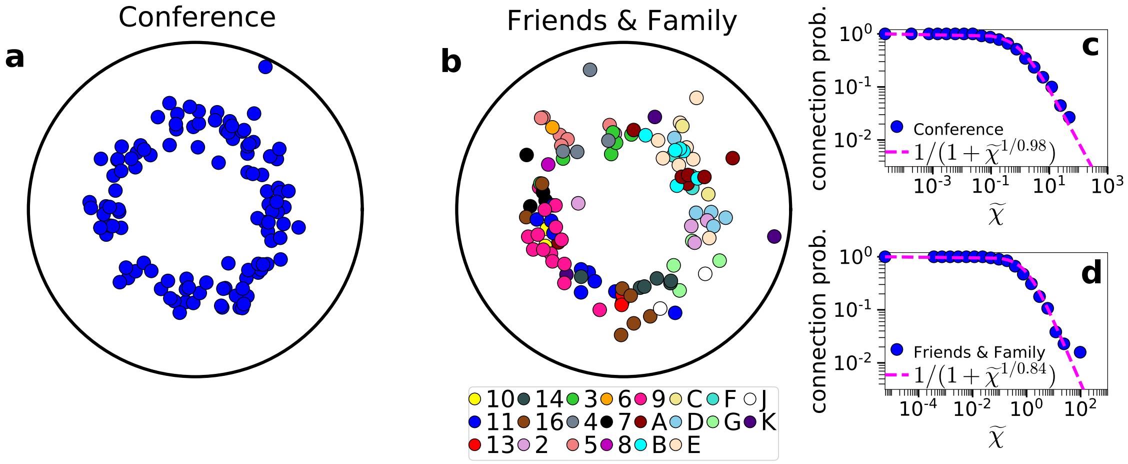

Figure C.3: **Hyperbolic embeddings of the conference and Friends & Family.** Same as in Fig. 6.4 but for the conference and Friends & Family. In (**a**) all nodes have the same color as there is no group membership information available for the conference. In (**b**) the nodes are colored according to the partial apartment number or letter where they live, as given in the Friends & Family metadata. The pink dashed lines in (**c**) and (**d**) are Fermi-Dirac connection probabilities with temperatures T as inferred by Mercator, $T = 0.98$ and 0.84, respectively.

C.4 Human-to-human greedy routing

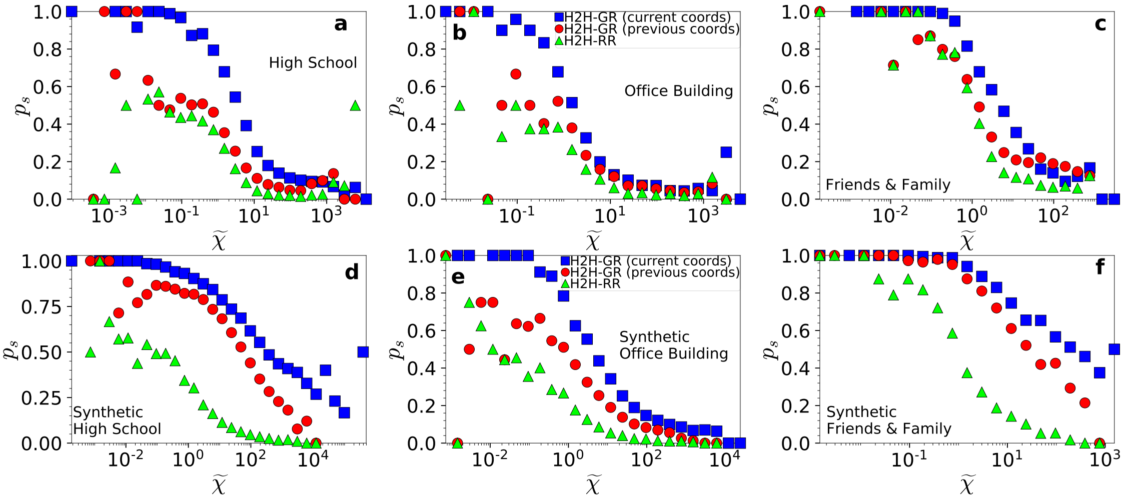

Figure C.4: **Success ratio p_s of H2H-GR and H2H-RR as a function of the effective distance $\tilde{\chi}$ between source-destination pairs.** Same as in Fig. 6.5 but for the high school, office building and Friends & Family. The top row shows the results for the real networks, while the bottom row shows the results for the synthetic counterparts.

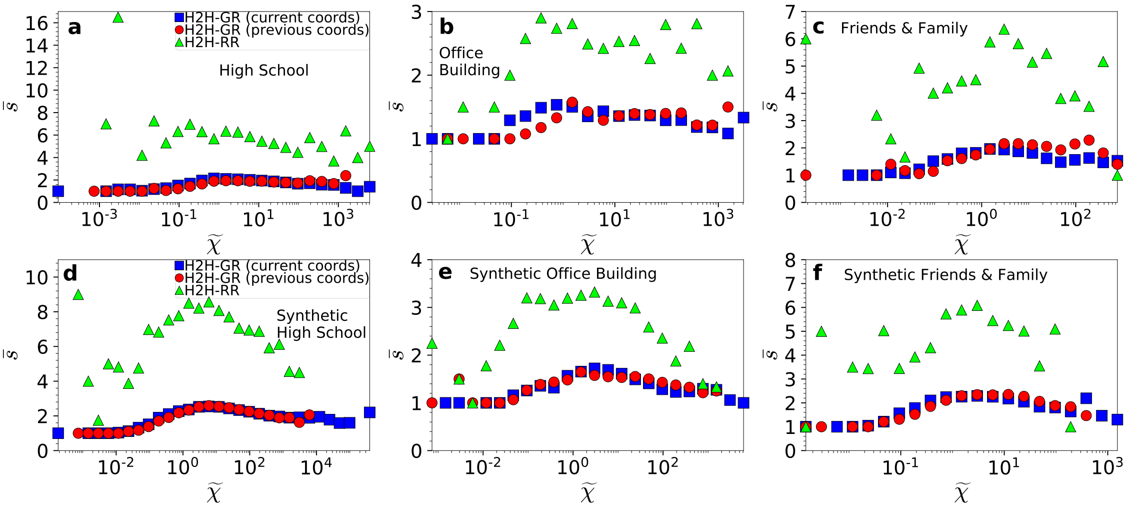

Figure C.5: **Average stretch $\bar{s}$ of H2H-GR and H2H-RR as a function of the effective distance $\tilde{\chi}$ between source-destination pairs.** The results correspond to the networks of Fig. C.4.

Real Network	H2H-GR (current angular coordinates)	H2H-GR (previous angular coordinates)
Hospital	$p_s = 0.69,\ \bar{s} = 2.16$	$p_s = 0.39,\ \bar{s} = 1.98$
Primary School	$p_s = 0.69,\ \bar{s} = 4.40$	$p_s = 0.65,\ \bar{s} = 3.88$
Conference	$p_s = 0.55,\ \bar{s} = 2.25$	$p_s = 0.35,\ \bar{s} = 2.11$
High School	$p_s = 0.20,\ \bar{s} = 2.12$	$p_s = 0.10,\ \bar{s} = 1.84$
Office Building	$p_s = 0.11,\ \bar{s} = 1.42$	$p_s = 0.09,\ \bar{s} = 1.39$
Friends & Family	$p_s = 0.42,\ \bar{s} = 2.20$	$p_s = 0.28,\ \bar{s} = 2.03$

Table C.1: **Success ratio p_s and average stretch $\bar{s}$ of H2H-GR that uses only the angular (similarity) distances in real networks.** Same as in Table 6.2 but when using only the inferred angular coordinates (current and previous) in H2H-GR.

Synthetic Network	H2H-GR (current angular coordinates)	H2H-GR (previous angular coordinates)
Hospital	$p_s = 0.71,\ \bar{s} = 2.46$	$p_s = 0.59,\ \bar{s} = 2.44$
Primary School	$p_s = 0.97,\ \bar{s} = 4.77$	$p_s = 0.91,\ \bar{s} = 5.29$
Conference	$p_s = 0.70,\ \bar{s} = 2.83$	$p_s = 0.51,\ \bar{s} = 2.77$
High School	$p_s = 0.35,\ \bar{s} = 2.93$	$p_s = 0.25,\ \bar{s} = 2.90$
Office Building	$p_s = 0.13,\ \bar{s} = 1.66$	$p_s = 0.10,\ \bar{s} = 1.61$
Friends & Family	$p_s = 0.58,\ \bar{s} = 2.58$	$p_s = 0.49,\ \bar{s} = 2.47$

Table C.2: Same as in Table C.1 but for the synthetic counterparts of the real systems.

C.5 Stability of the inferred node coordinates in different days

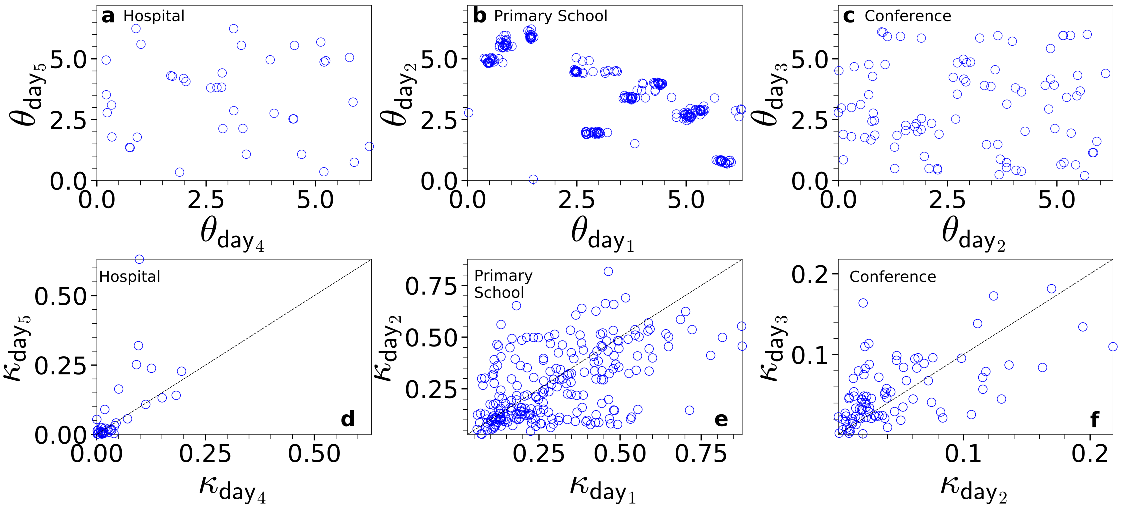

Figure C.6: **Inferred node coordinates** (κ, θ) **from the time-aggregated network of different observation days.** The results correspond to real networks, and the considered days are as in Table 6.2. (**a**) Inferred angles in day 4 versus inferred angles in day 5 in the hospital. The numbers of time slots for days 4 and 5 are $\tau = 3889$ and 2177, respectively, while Mercator's inferred temperature for the time-aggregated network is $T = 0.99$ for both days. (**b**) Inferred angles in day 1 versus inferred angles in day 2 in the primary school. For days $1, 2$, $\tau = 1555, 1545$ and $T = 0.43, 0.36$. (**c**) Inferred angles in day 2 versus inferred angles in day 3 in the conference. For days $2, 3$, $\tau = 3216, 1946$ and $T = 0.99, 0.98$. (**d-f**) Same as in (**a-c**) but for the inferred latent degrees κ. For each node, κ is estimated as $\kappa = \tilde{\kappa}/\alpha$, where $\tilde{\kappa}$ is the node's inferred latent degree in the time-aggregated network of the corresponding day, while $\alpha = \tau^T/\Gamma(1+T)$. Due to rotational symmetry of the model, the inferred angles in a day can be globally shifted compared to the inferred angles in another day by any value in $[0, 2\pi]$.

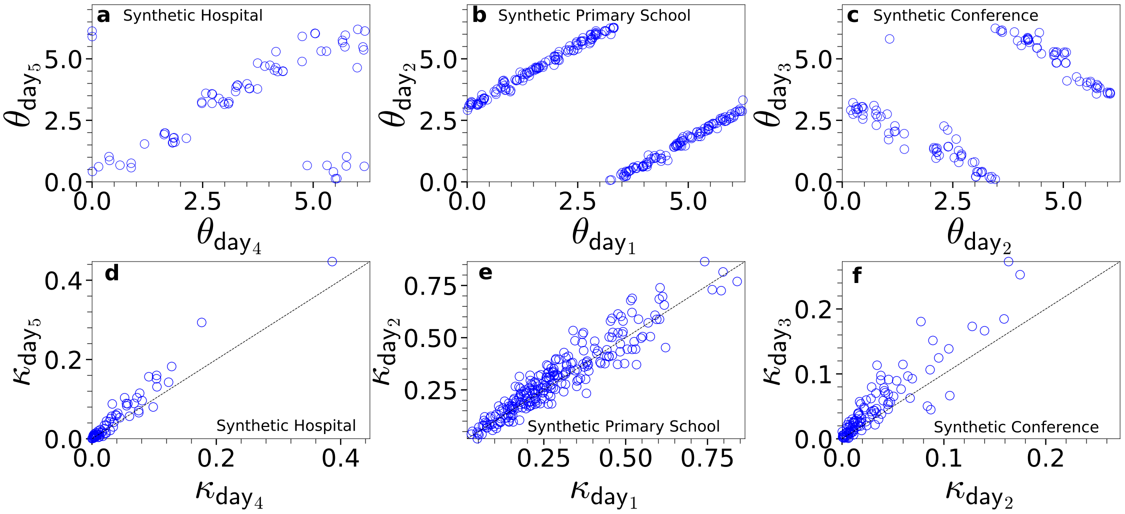

Figure C.7: **Inferred node coordinates** (κ, θ) **from the time-aggregated network of different days.** Same as in Fig. C.6 but for the synthetic counterparts of the real systems. The days in each counterpart have the same duration τ as in the corresponding real system. The temperatures inferred by Mercator are $T = 0.57$ for both days of the hospital, $T = 0.60, 0.64$ for days $1, 2$ of the primary school, and $T = 0.64$ for both days of the conference.

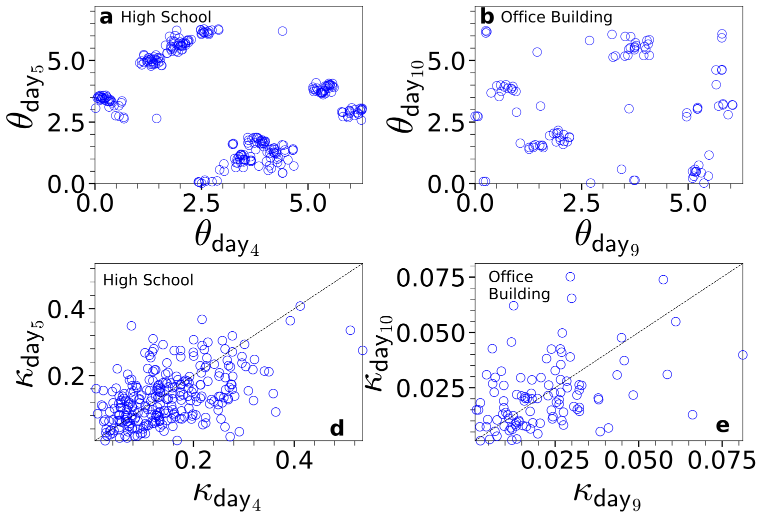

Figure C.8: **Inferred node coordinates** (κ, θ) **from the time-aggregated network of different observation days.** Same as in Fig. C.6 but for the high school, office building and Friends & Family. (**a**) Inferred angles in day 4 versus inferred angles in day 5 in the high school. For days $4, 5$, $\tau = 1619$ and $T = 0.54, 0.49$. (**b**) Inferred angles in day 9 versus inferred angles in day 10 in the office building. For days $9, 10$, $\tau = 2153, 2148$ and $T = 0.57, 0.47$. (**c**) Inferred angles in the 30th of March, 2011 versus inferred angles in the 31st of March, 2011 in the Friends & Family. For the two days, $\tau = 289$ and $T = 0.52, 0.49$. (**d-f**) Same as in (**a-c**) but for the inferred latent degrees κ.

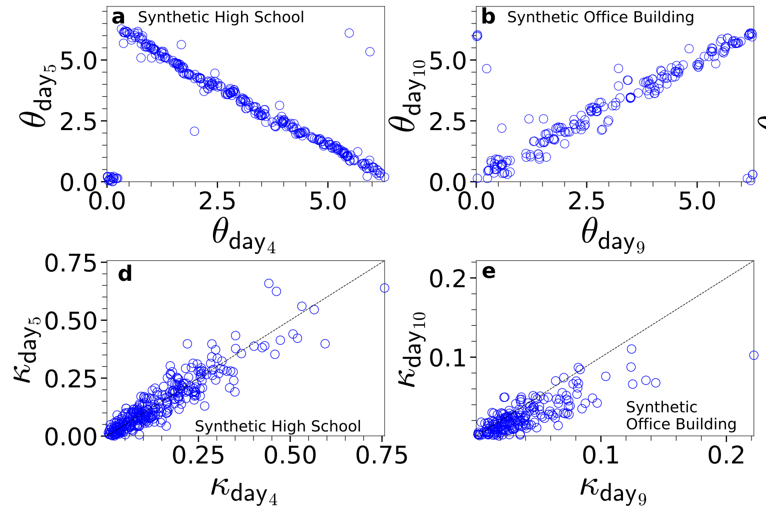

Figure C.9: **Inferred node coordinates** (κ, θ) **from the time-aggregated network of different days.** Same as in Fig. C.8 but for the synthetic counterparts of the real systems. The days in each counterpart have the same duration τ as in the corresponding real system. The temperatures inferred by Mercator are $T = 0.54$ for both days of the high school, $T = 0.68$ for both days of the office building, and $T = 0.45, 0.43$ for the two days of the Friends & Family.

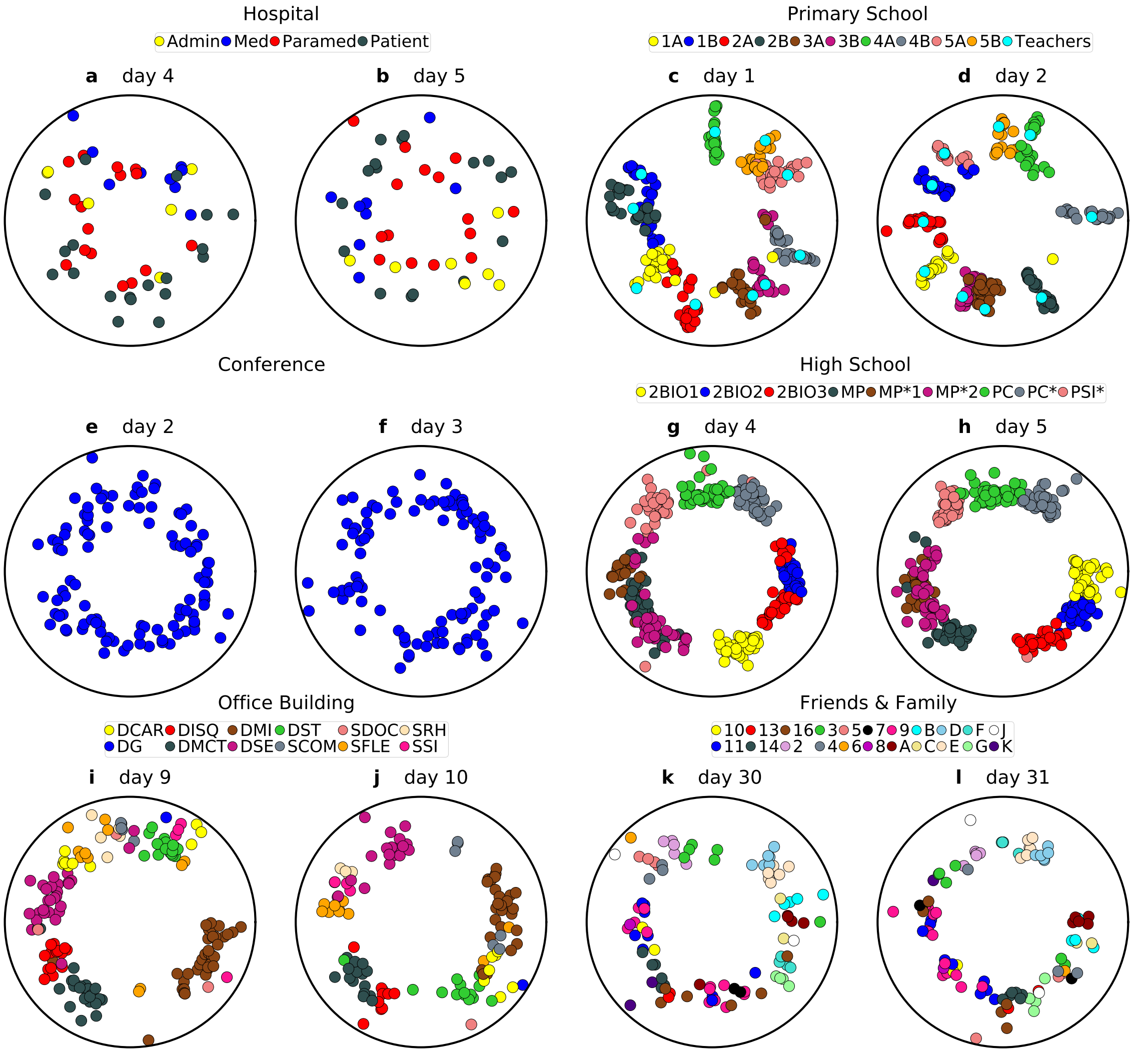

Figure C.10: **Daily hyperbolic maps of the considered real systems.** The maps correspond to the days considered in Figs. C.6 and C.8. The nodes are positioned according to their inferred hyperbolic coordinates (r, θ) in the time-aggregated network of the corresponding day. For an easier inspection of how node coordinates change between days, the map of each day is rotated such that it minimizes the sum of the squared distances between the inferred angles in the day and the angles inferred by considering the full duration of the corresponding network (Fig. 6.4 and Fig. C.3). The nodes are colored according to group membership information available in the metadata of each network as described in Fig. 6.4 and in Fig. C.3.

C.6 Link prediction: common neighbors benchmark ROC and PR curves

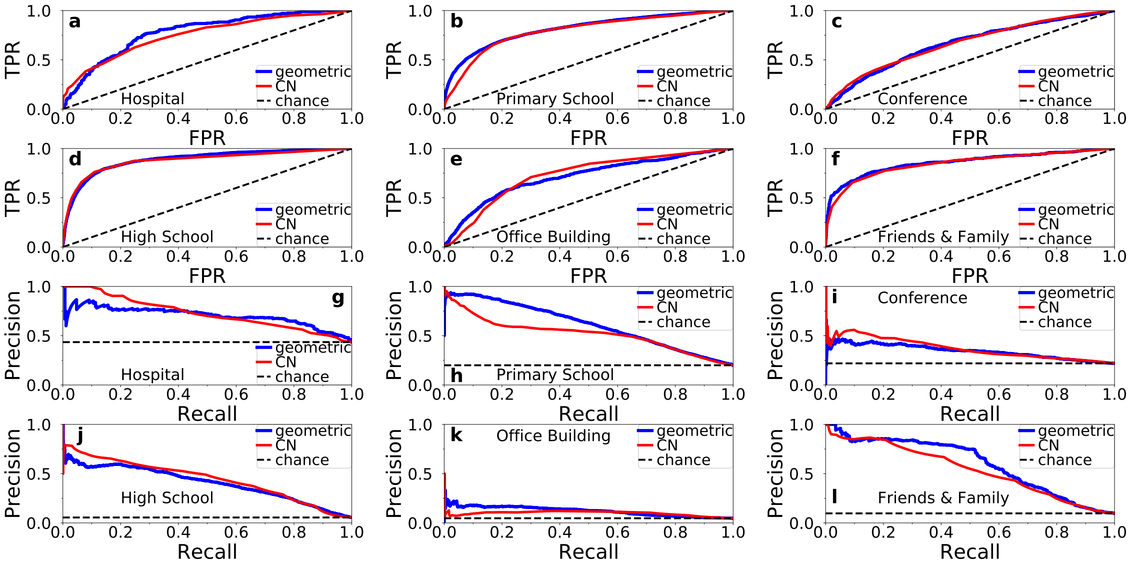

Figure C.11: ROC and PR curves for geometric link prediction and common neighbors in real networks. (a-f) show the ROC curves, while (g-l) the PR curves. The dashed black lines correspond to link prediction based on chance.

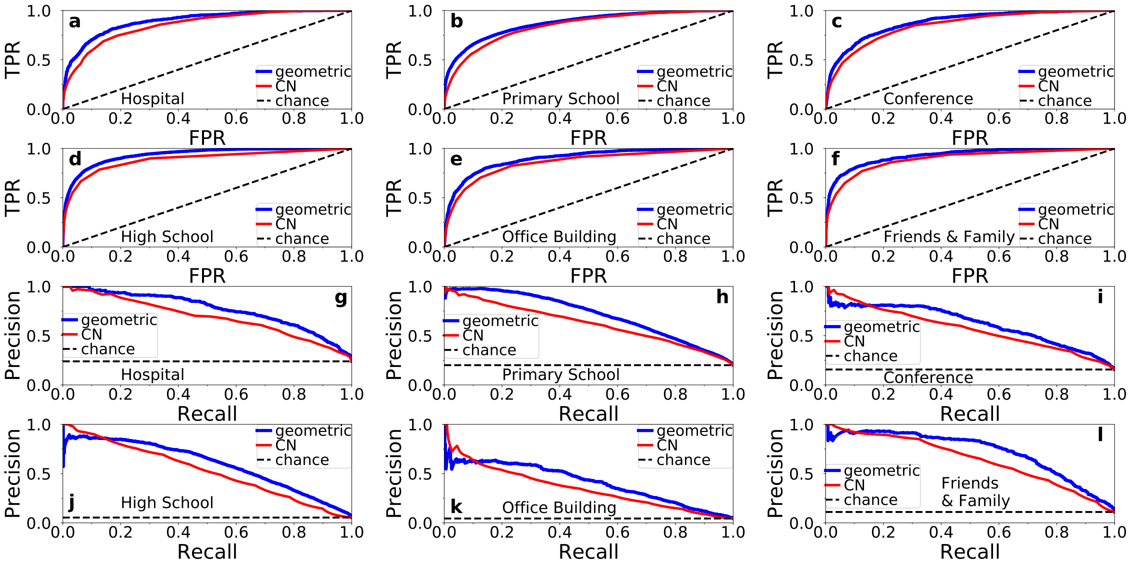

Figure C.12: Same as Fig. C.11 but for the synthetic networks.

C.7 Modified Mercator

In the modified version of Mercator we replace the connection probability of the $\mathbb{S}^1$ model [Eq. (5.1) in Chapter 5] with the connection probability in the time-aggregated network of the dynamic-$\mathbb{S}^1$ model [Eq. (6.4) in Chapter 6]. This modification requires replacing all relations in the original Mercator implementation, which are derived using the original connection probability, with the corresponding relations derived using the new connection probability. Below we list all modifications made in each step of the original Mercator implementation [32]. For convenience, we express all new relations in terms of the nodes' latent degrees per snapshot κ. We recall that $\kappa = \tilde{\kappa}/\alpha$, where $\tilde{\kappa}$ denotes the node's latent degree in the time-aggregated network, while $\alpha = \tau^T/\Gamma(1+T)$. The modified Mercator takes also as input the value of the duration τ in which the time-aggregated network is computed, while like the original version it infers the value of the temperature parameter T along with the nodes' coordinates $(\tilde{\kappa}, \theta)$.

In the first step, Mercator uses an iterative procedure that adjusts the nodes' latent degrees so that the expected degree of each node as prescribed by the $\mathbb{S}^1$ model matches the node's observed degree in the given network. We adapt this step by replacing the relation for the probability that two nodes with latent degrees κ and κ' are connected [Eq. (A1) in Ref. [32]], with

$$p(a_{\kappa\kappa'} = 1) =$$

$$1 - \tfrac{2\mu\kappa\kappa'}{N}\left[\tfrac{T\Gamma(\tau+T)\Gamma(-T)}{\Gamma(\tau)} + \left(\tfrac{1}{1+\left(\frac{N}{2\mu\kappa\kappa'}\right)^{1/T}}\right)^{-T} {}_2F_1\left(-T, 1-\tau-T; 1-T; \tfrac{1}{1+\left(\frac{N}{2\mu\kappa\kappa'}\right)^{1/T}}\right)\right]. \tag{C.1}$$

The above relation is derived in Sec. 5.5.4 in Chapter 5. $a_{\kappa\kappa'}$ is an indicator function, $a_{\kappa\kappa'} = 1$ if two nodes with latent degrees κ and κ' are connected in the time-aggregated network, and $a_{\kappa\kappa'} = 0$ otherwise, $\mu = \sin(T\pi)/(2\bar{\kappa}T\pi)$, and ${}_2F_1(a, b; c; z)$ is the Gauss hypergeometric function.

In the second step, Mercator uses an iterative procedure that adjusts the temperature T so that the value of the average clustering coefficient as prescribed by the $\mathbb{S}^1$ model matches the value of the average clustering coefficient in the given network. We adapt this step by replacing the relation for the distribution of the angular distance $\Delta\theta$ between two connected nodes with latent degrees κ and κ'

[Eq. (A3) in Ref. [32]], with

$$
\begin{aligned}
\rho(\Delta\theta|a_{\kappa\kappa'}=1) &= \frac{p(a_{\kappa\kappa'}=1|\Delta\theta)\rho(\Delta\theta)}{p(a_{\kappa\kappa'}=1)} \\
&= \frac{\frac{1}{\pi}\left[1-\left(1-\frac{1}{1+\left(\frac{R\Delta\theta}{\mu\kappa\kappa'}\right)^{1/T}}\right)^{\tau}\right]}{1-\frac{2\mu\kappa\kappa'}{N}\left[\frac{T\Gamma(\tau+T)\Gamma(-T)}{\Gamma(\tau)}+\left(\frac{1}{1+\left(\frac{N}{2\mu\kappa\kappa'}\right)^{1/T}}\right)^{-T}{}_2F_1\left(-T,1-\tau-T;1-T;\frac{1}{1+\left(\frac{N}{2\mu\kappa\kappa'}\right)^{1/T}}\right)\right]}.
\end{aligned}
\tag{C.2}
$$

In the above relation, $p(a_{\kappa\kappa'}=1|\Delta\theta)$ is the probability that two nodes with latent degrees κ and κ' and angular distance $\Delta\theta$ are connected in the time-aggregated network, $\rho(\Delta\theta)=1/\pi$ is the uniform distribution of the angular distances in the model, and $p(a_{\kappa\kappa'}=1)$ is given by (C.1).

In the third step, Mercator adapts Laplacian Eigenmaps (LE) to the $\mathbb{S}^1$ model in order to determine initial angular coordinates for the nodes. We adapt this step by replacing the relation for the expected angular distance between two nodes with latent degrees κ_i and κ_j conditioned on the fact that they are connected [Eq. (A8) in Ref. [32]], with

$$
\begin{aligned}
\langle\Delta\theta_{ij}\rangle &= \int_0^{\pi}\Delta\theta_{ij}\rho(\Delta\theta_{ij}|a_{\kappa_i\kappa_j}=1)\mathrm{d}\Delta\theta_{ij} \\
&= \frac{N\pi\Gamma(\tau)\left[\tau+2T-2T\left(\frac{\pi R}{\mu\kappa_i\kappa_j}\right)^{\tau/T}{}_2F_1\left(\tau,\tau+2T;\tau+2T+1;-\left(\frac{\pi R}{\mu\kappa_i\kappa_j}\right)^{1/T}\right)\right]}{2(\tau+2T)\left[\left(N+2T\mu\kappa_i\kappa_j\mathrm{B}(x;-T,\tau+T)\right)\Gamma(\tau)+2\mu\kappa_i\kappa_j\Gamma(1-T)\Gamma(\tau+T)\right]},
\end{aligned}
\tag{C.3}
$$

where $\mathrm{B}(x;a,b)$ is the incomplete beta function and $x=1/\left[1+\left(\frac{N}{2\mu\kappa_i\kappa_j}\right)^{1/T}\right]$. In this step, we also replace the probability in the $\mathbb{S}^1$ model of having the observed connection (or disconnection) among each pair of consecutive nodes i and $i+1$ on the similarity circle, conditioned on their angular separation gap g_i and their latent degrees κ_i and κ_j [Eq. (A13) in Ref. [32]], with

$$p(a_{i+1,i}|g_i) = \left[1-\left(1-\frac{1}{1+\left(\frac{Rg_i}{\mu\kappa_i\kappa_j}\right)^{1/T}}\right)^{\tau}\right]^{a_{i+1,i}} \times \left[\left(1-\frac{1}{1+\left(\frac{Rg_i}{\mu\kappa_i\kappa_j}\right)^{1/T}}\right)^{\tau}\right]^{1-a_{i+1,i}}, \qquad \text{(C.4)}$$

where $a_{i,i+1} = 1$ if the two nodes are connected in the time-aggregated network, and $a_{i,i+1} = 0$ otherwise.

In the fourth step, Mercator refines the initial angular coordinates by (approximately) maximizing the likelihood that the given network is produced by the $\mathbb{S}^1$ model. We adapt this step by replacing the local log-likelihood for each node i [Eq. (A20) in Ref. [32]], with

$$\ln \mathcal{L}_i = \sum_{j\neq i} a_{ij} \ln \left[1-\left(1-\frac{1}{1+\left(\frac{R\Delta\theta_{ij}}{\mu\kappa_i\kappa_j}\right)^{1/T}}\right)^{\tau}\right] + (1-a_{ij}) \ln \left[\left(1-\frac{1}{1+\left(\frac{R\Delta\theta_{ij}}{\mu\kappa_i\kappa_j}\right)^{1/T}}\right)^{\tau}\right], \qquad \text{(C.5)}$$

where $a_{ij} = 1$ if nodes i and j are connected in the time-aggregated network, and $a_{ij} = 0$ otherwise.

In the final (optional) step, Mercator re-adjusts the latent degrees of the nodes according to the inferred angular coordinates so that the expected degree of each node indeed matches its observed degree in the given network. We adapt this step by replacing the connection probability of the $\mathbb{S}^1$ model in Eq. (A21) in Ref. [32], with the connection probability in the time-aggregated network of the dynamic-$\mathbb{S}^1$ model [Eq. 6.4 in Chapter 6].

C.8 Aggregation interval and rotation of angular coordinates

In Fig. 6.3 in Chapter 6 we quantify the difference between inferred and real coordinates as a function of the aggregation interval τ in a synthetic counterpart of the primary school. In this section, Figs. C.13-C.17 correspond to the same results but for synthetic counterparts of the hospital, conference, high school, office building and Friends & Family. Specifically, each of Figs. C.13-C.17 shows the metrics $D_\kappa(\tau)$, $D_\theta(\tau)$ and $d(\tau)$ defined in the caption of Fig. 6.3 in Chapter 6. Further, Figs. C.18-C.29 juxtapose the inferred against the real coordinates in each synthetic counterpart, as a function of the aggregation interval τ.

Before computing $D_\theta(\tau)$, we globally shift (rotate) the inferred angles such that the sum of the squared distances (SSD) between real and rotated inferred angles is minimized. To this end, we apply a Procrustean rotation (Ref. [89]), as follows:

1. We transform the real and inferred angles $\{\theta^i_{\text{real}}\}$ and $\{\theta^i_{\text{inferred}}\}$ to Cartesian coordinates $\{x_i, y_i\} = \{\cos\theta^i_{\text{real}}, \sin\theta^i_{\text{real}}\}$ and $\{w_i, z_i\} = \{\cos\theta^i_{\text{inferred}}, \sin\theta^i_{\text{inferred}}\}$ for all nodes $i = 1, \ldots, N$.[1]

2. A rotation of the points $\{w_i, z_i\}$ by an angle ϕ is given by $\{u_i, v_i\} = \{w_i \cos\phi - z_i \sin\phi, w_i \sin\phi + z_i \cos\phi\}$, where u_i, v_i are the coordinates of the rotated point w_i, z_i. The SSD between $\{u_i, v_i\}$ and $\{x_i, y_i\}$ is SSD $= \sum_{i=1}^N (u_i - x_i)^2 + (v_i - y_i)^2$. The optimal rotation angle ϕ^* is computed by taking the derivative of the SSD with respect to ϕ and solving for ϕ when the derivative is zero,

   $$\phi^* = \tan^{-1}\left(\frac{\sum_{i=1}^N (w_i y_i - z_i x_i)}{\sum_{i=1}^N (w_i x_i + z_i y_i)}\right). \tag{C.6}$$

   We compute the optimally rotated inferred angles as $\{\theta^i_{\text{rotated}}\} = \{\tan^{-1}(v_i^*/u_i^*)\}$, where $\{u_i^*, v_i^*\} = \{w_i \cos\phi^* - z_i \sin\phi^*, w_i \sin\phi^* + z_i \cos\phi^*\}$.[2]

3. We repeat the above procedure after replacing $\{\theta^i_{\text{inferred}}\}$ with $\{2\pi - \theta^i_{\text{inferred}}\}$, which is the reflection of the former across the x-axis, and compute the optimally rotated inferred angles in this case as well, $\{\tilde{\theta}^i_{\text{rotated}}\}$.

4. We compute $D_\theta(\tau) = \sum_{i=1}^N |\theta^i_{\text{rotated}} - \theta^i_{\text{real}}|/N$ and $\widetilde{D}_\theta(\tau) = \sum_{i=1}^N |\tilde{\theta}^i_{\text{rotated}} - \theta^i_{\text{real}}|/N$. The optimally rotated inferred angles are $\{\theta^i_{\text{rotated}}\}$ if $D_\theta(\tau) < \widetilde{D}_\theta(\tau)$, and $\{\tilde{\theta}^i_{\text{rotated}}\}$ otherwise.

We follow a similar procedure for the rotations in Fig. C.10.

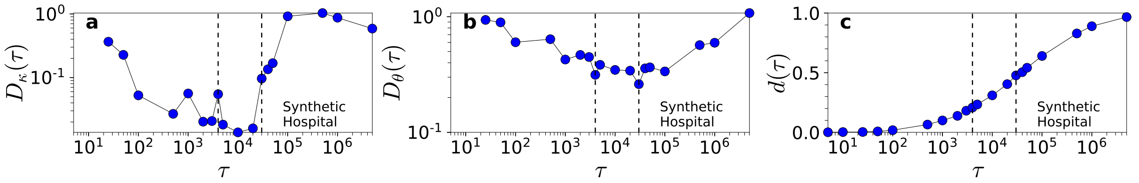

Figure C.13: **Inference accuracy vs. aggregation interval.** Same as in Fig. 6.3 but for a synthetic counterpart of the hospital. The vertical dashed lines indicate the interval $4000 \leq \tau \leq 30000$. In this interval, $D_\kappa(\tau) < 0.1$, $D_\theta(\tau) < 0.4$, and $0.20 < d(\tau) < 0.48$.

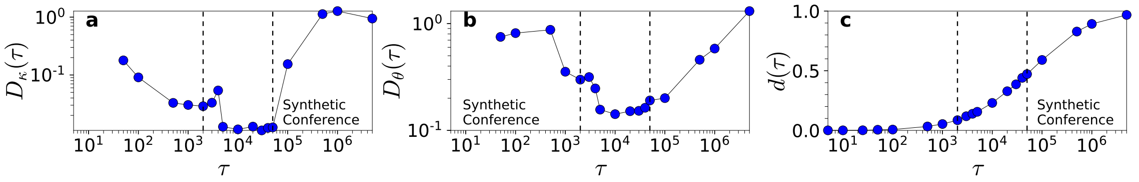

Figure C.14: **Inference accuracy vs. aggregation interval.** Same as in Fig. C.13 but for a synthetic counterpart of the conference. The vertical dashed lines indicate the interval $2000 \leq \tau \leq 50000$. In this interval, $D_\kappa(\tau) < 0.1$, $D_\theta(\tau) < 0.4$, and $0.09 < d(\tau) < 0.47$.

[1] Notation“{ }" denotes a set. For example, $\{x_i, y_i\} = \{x_1, y_1, x_2, y_2, \ldots, x_N, y_N\}$.

[2] If $\theta^i_{\text{rotated}} < 0$, then $\theta^i_{\text{rotated}} := 2\pi + \theta^i_{\text{rotated}}$.

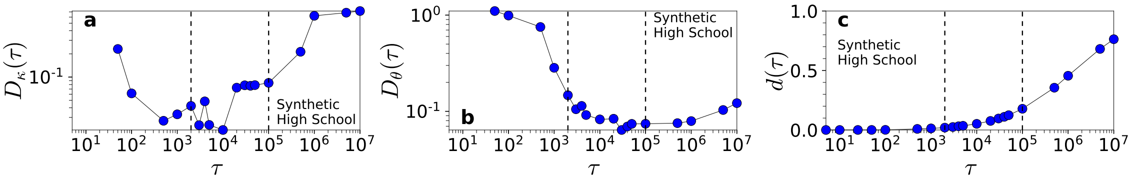

Figure C.15: **Inference accuracy vs. aggregation interval.** Same as in Fig. C.13 but for a synthetic counterpart of the high school. The vertical dashed lines indicate the interval $2000 \leq \tau \leq 100000$. In this interval, $D_\kappa(\tau) < 0.1$, $D_\theta(\tau) < 0.2$, and $0.01 < d(\tau) < 0.18$.

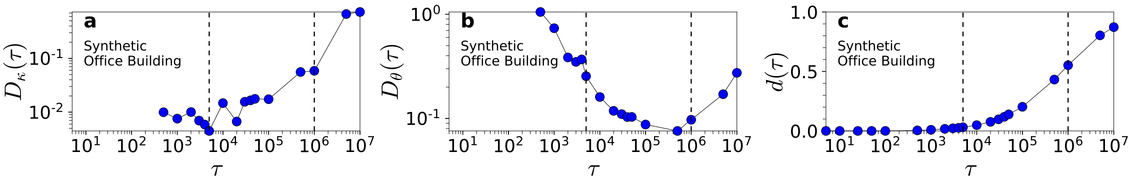

Figure C.16: **Inference accuracy vs. aggregation interval.** Same as in Fig. C.13 but for a synthetic counterpart of the office building. The vertical dashed lines indicate the interval $5000 \leq \tau \leq 1000000$. In this interval, $D_\kappa(\tau) < 0.1$, $D_\theta(\tau) < 0.3$, and $0.03 < d(\tau) < 0.55$.

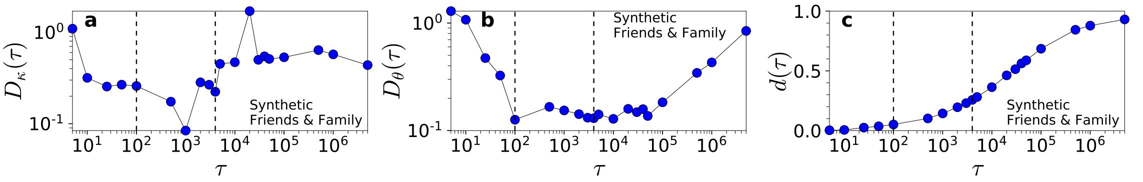

Figure C.17: **Inference accuracy vs. aggregation interval.** Same as in Fig. C.13 but for a synthetic counterpart of the friends & family. The vertical dashed lines indicate the interval $100 \leq \tau \leq 4000$. In this interval, $D_\kappa(\tau) < 0.3$, $D_\theta(\tau) < 0.2$, and $0.05 < d(\tau) < 0.26$.

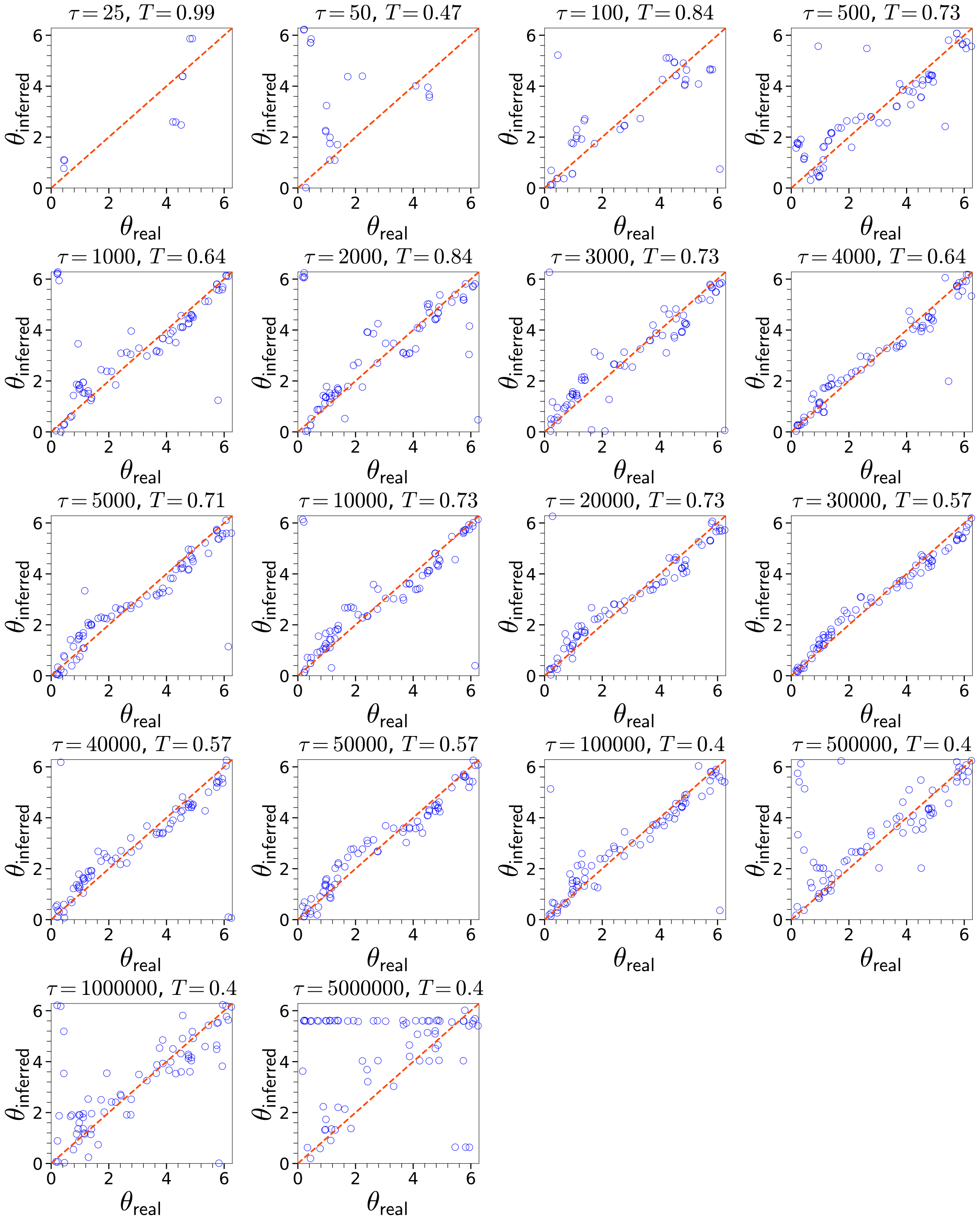

Figure C.18: **Inferred vs. real θ for different aggregation intervals τ.** The results correspond to the synthetic counterpart of the hospital. For each τ we also indicate the temperature T inferred by Mercator. The diagonal dashed line indicates $x = y$.

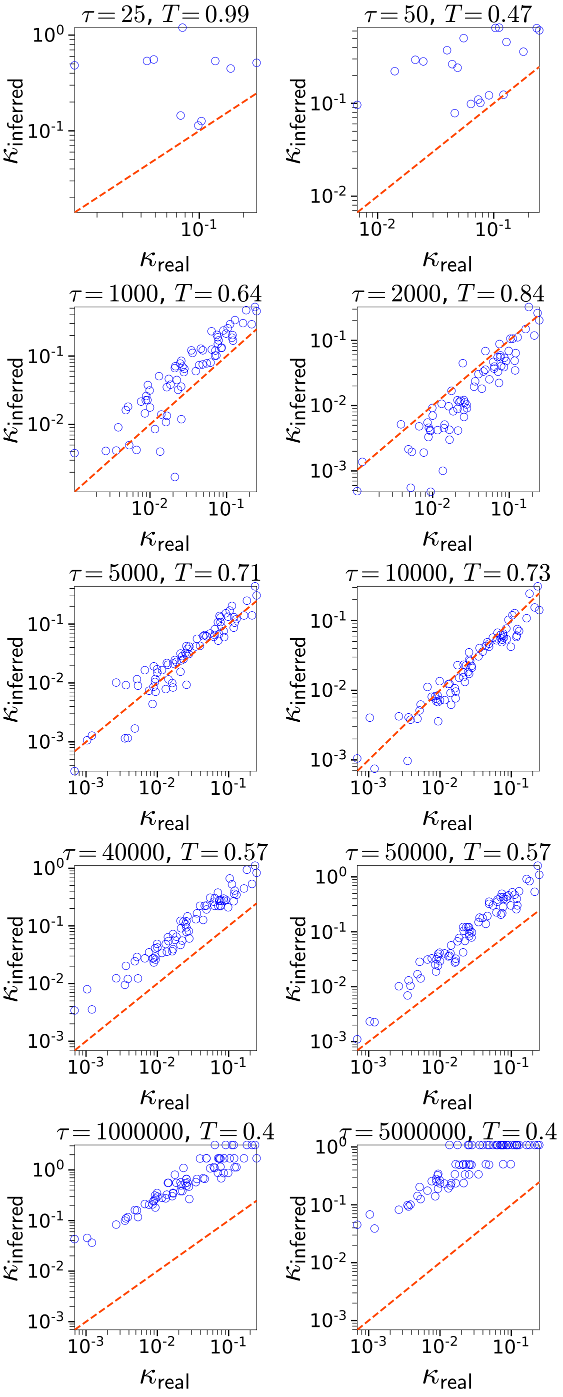

Figure C.19: **Inferred vs. real κ for different aggregation intervals τ.** The results correspond to the synthetic counterpart of the hospital. The κ_{inferred} are estimated as described in the caption of Fig. 6.2. For each τ we indicate the temperature T inferred by Mercator. The diagonal dashed line indicates $x = y$.

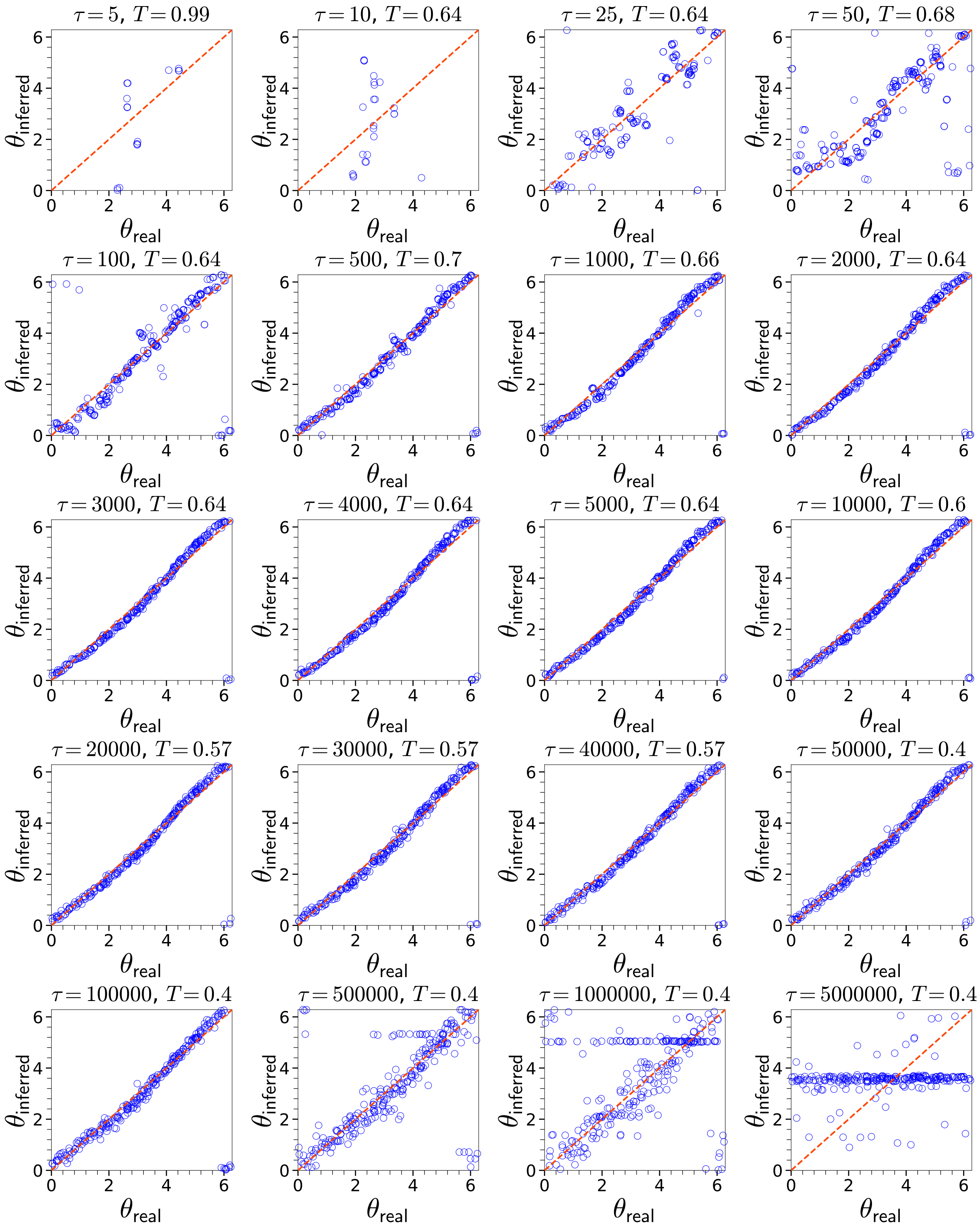

Figure C.20: **Inferred vs. real θ for different aggregation intervals τ.** Same as in Fig. C.18 but for the synthetic counterpart of the primary school.

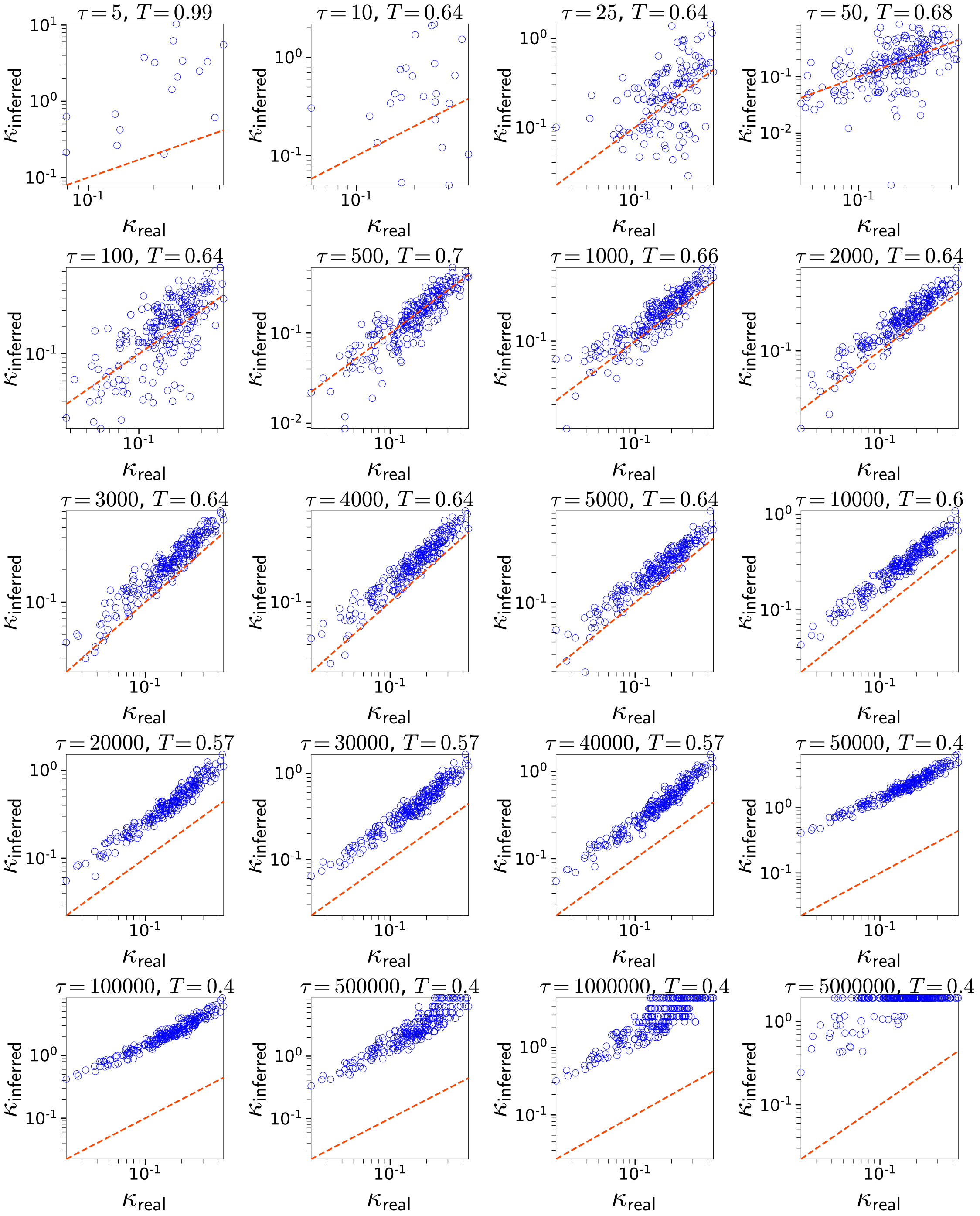

Figure C.21: **Inferred vs. real κ for different aggregation intervals τ.** Same as in Fig. C.19 but for the synthetic counterpart of the primary school.

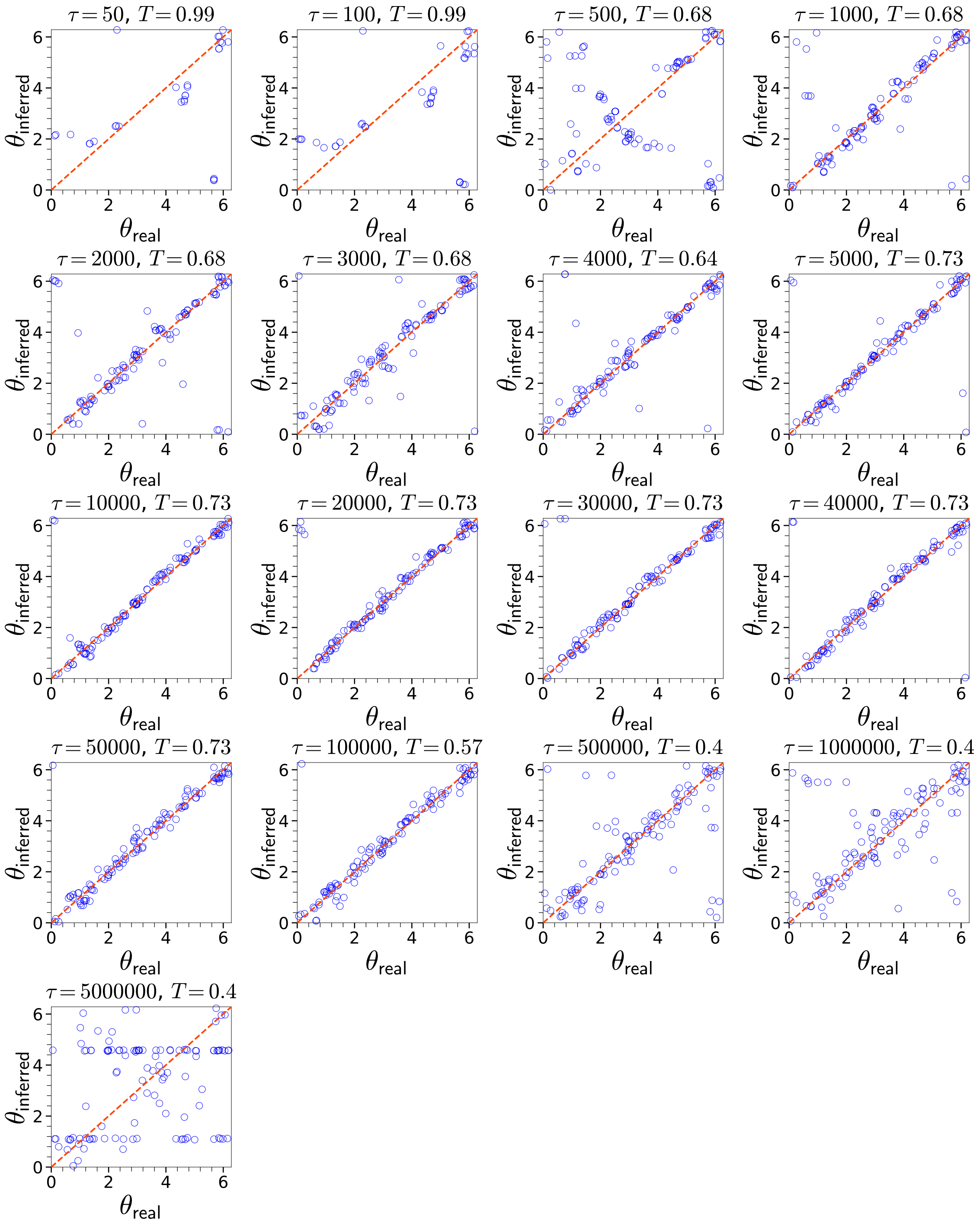

Figure C.22: **Inferred vs. real θ for different aggregation intervals τ.** Same as in Fig. C.18 but for the synthetic counterpart of the conference.

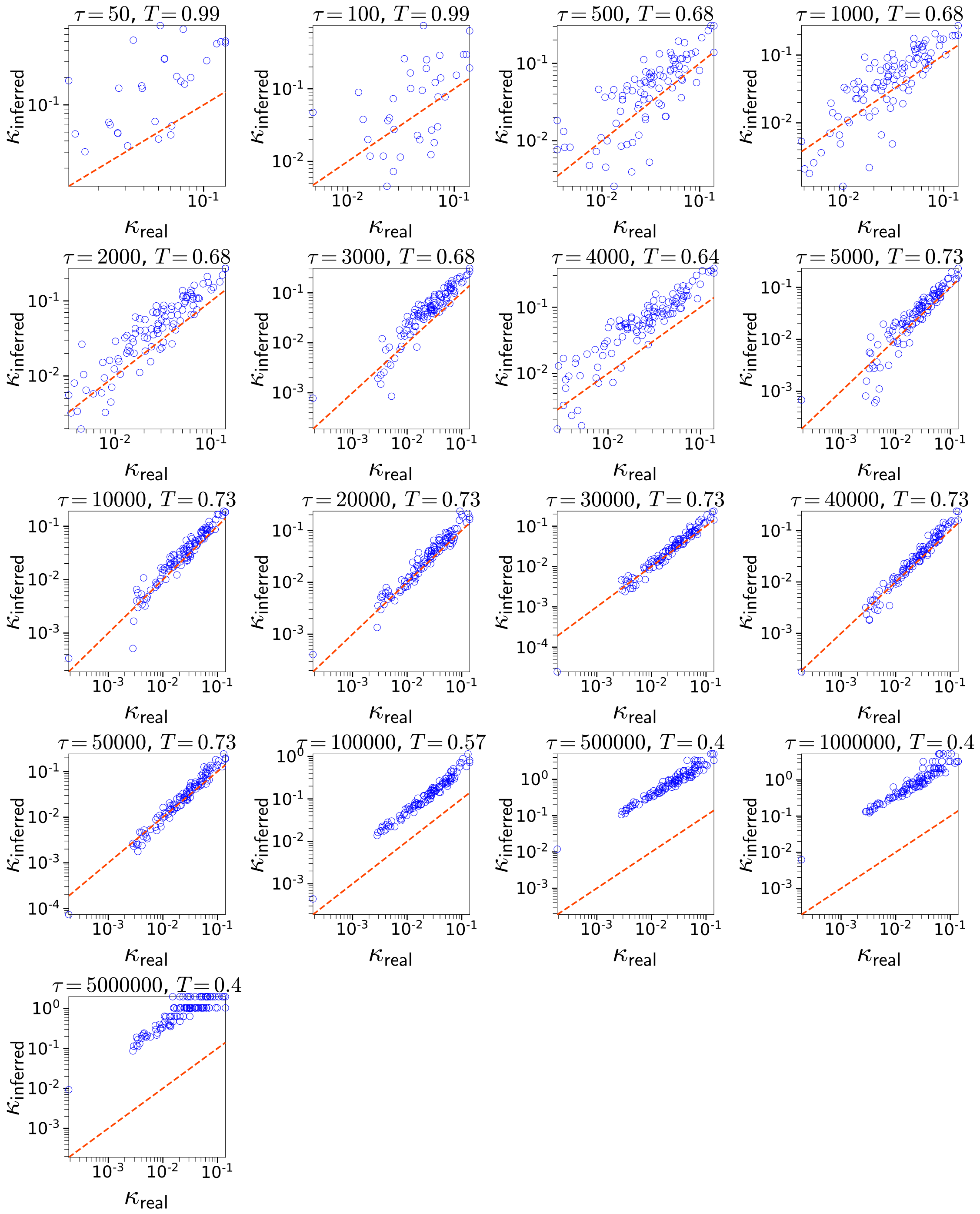

Figure C.23: **Inferred vs. real κ for different aggregation intervals τ.** Same as Fig. C.19 but for the synthetic counterpart of the conference.

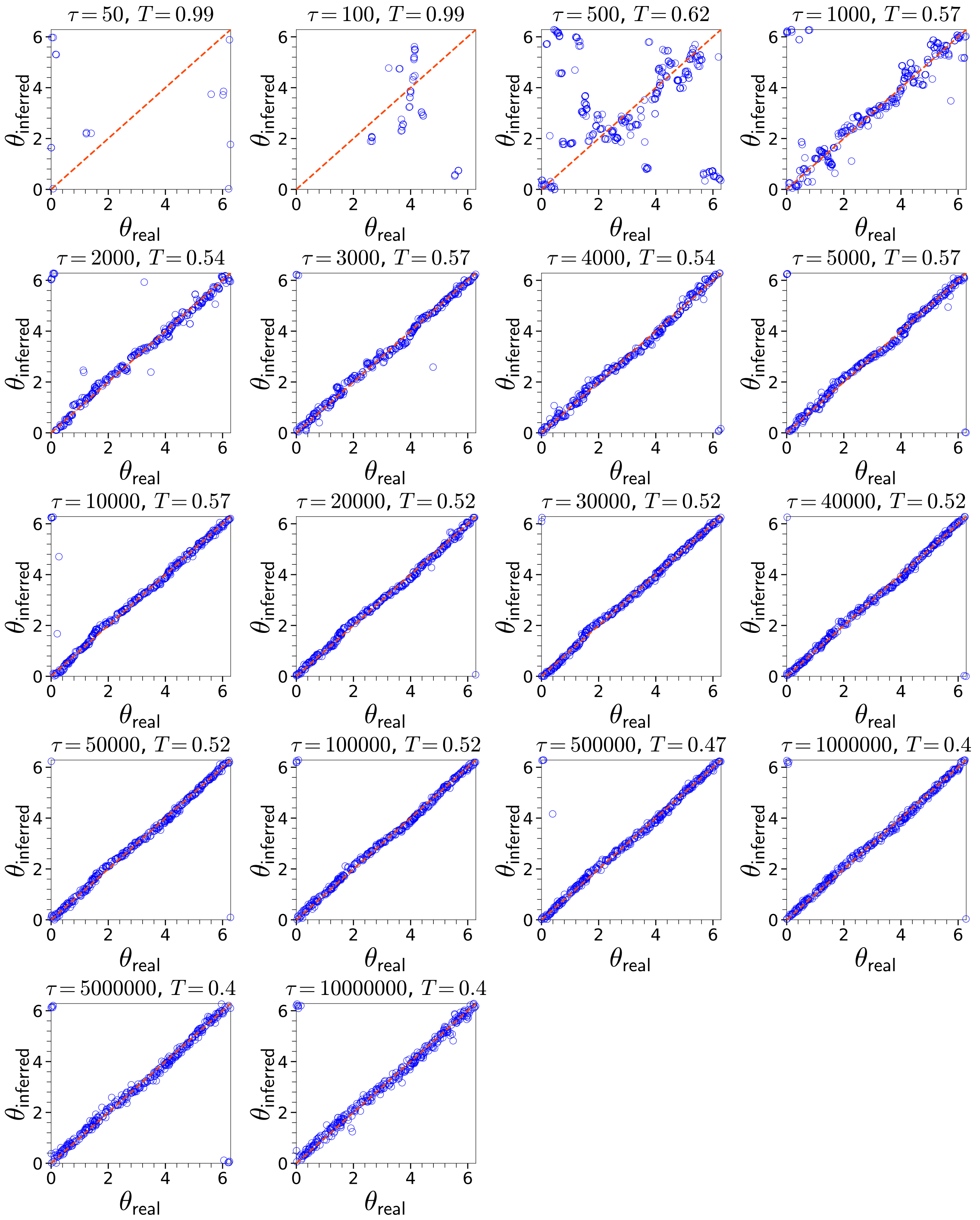

Figure C.24: **Inferred vs. real θ for different aggregation intervals τ.** Same as in Fig. C.18 but for the synthetic counterpart of the high school.

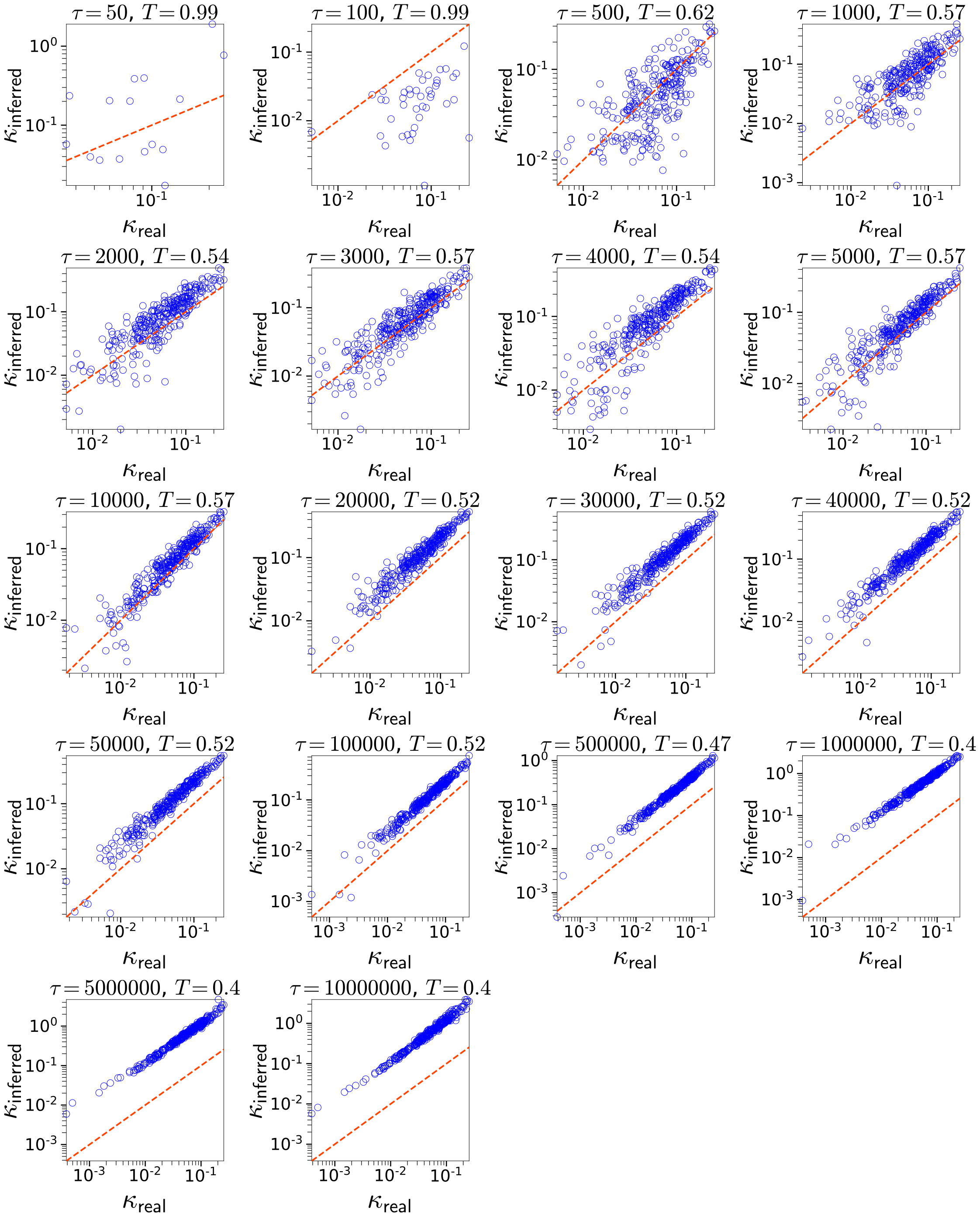

Figure C.25: **Inferred vs. real κ for different aggregation intervals τ.** Same as in Fig. C.19 but for the synthetic counterpart of the high school.

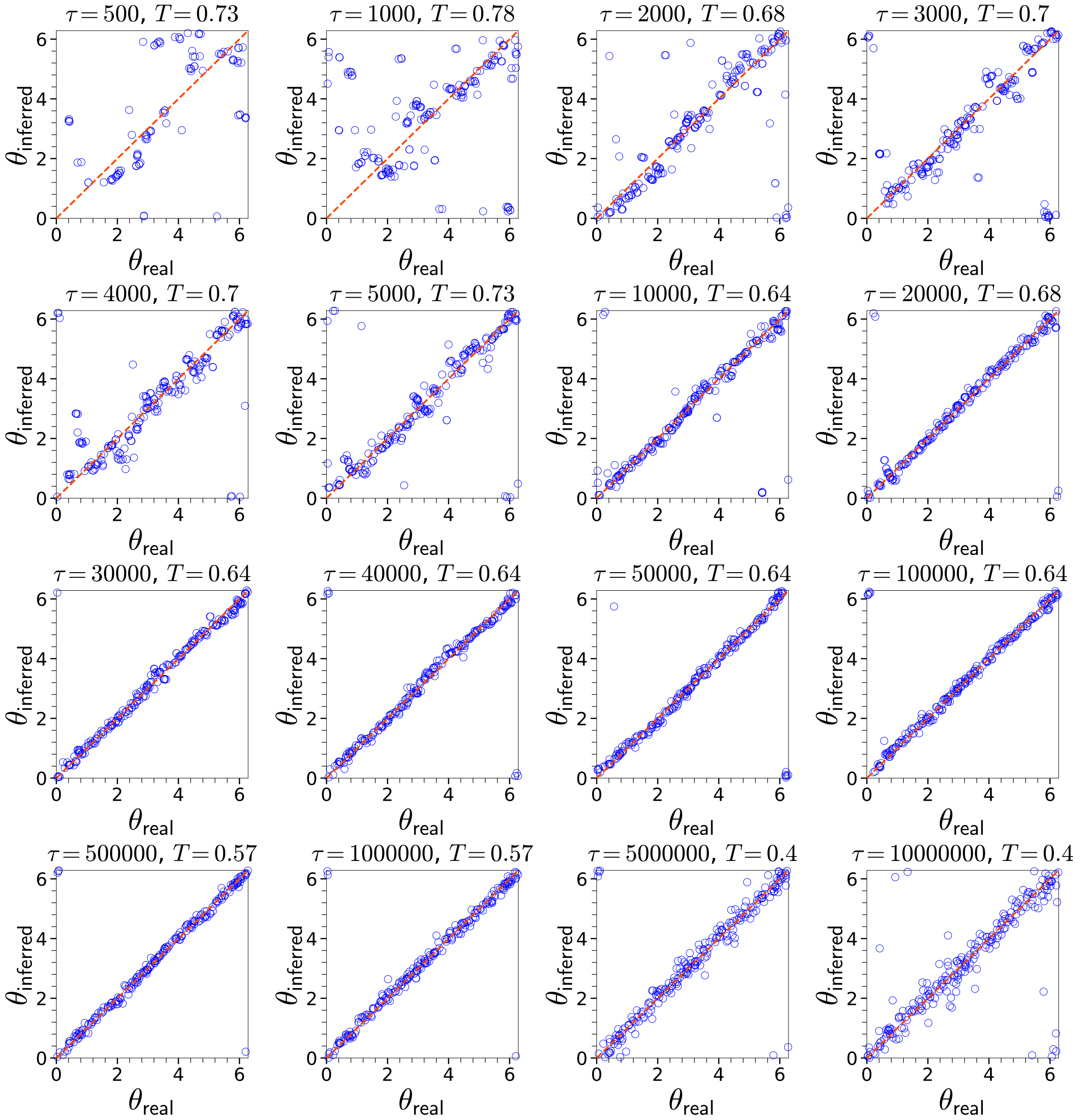

Figure C.26: **Inferred vs. real θ for different aggregation intervals τ.** Same as in Fig. C.18 but for the synthetic counterpart of the office building.

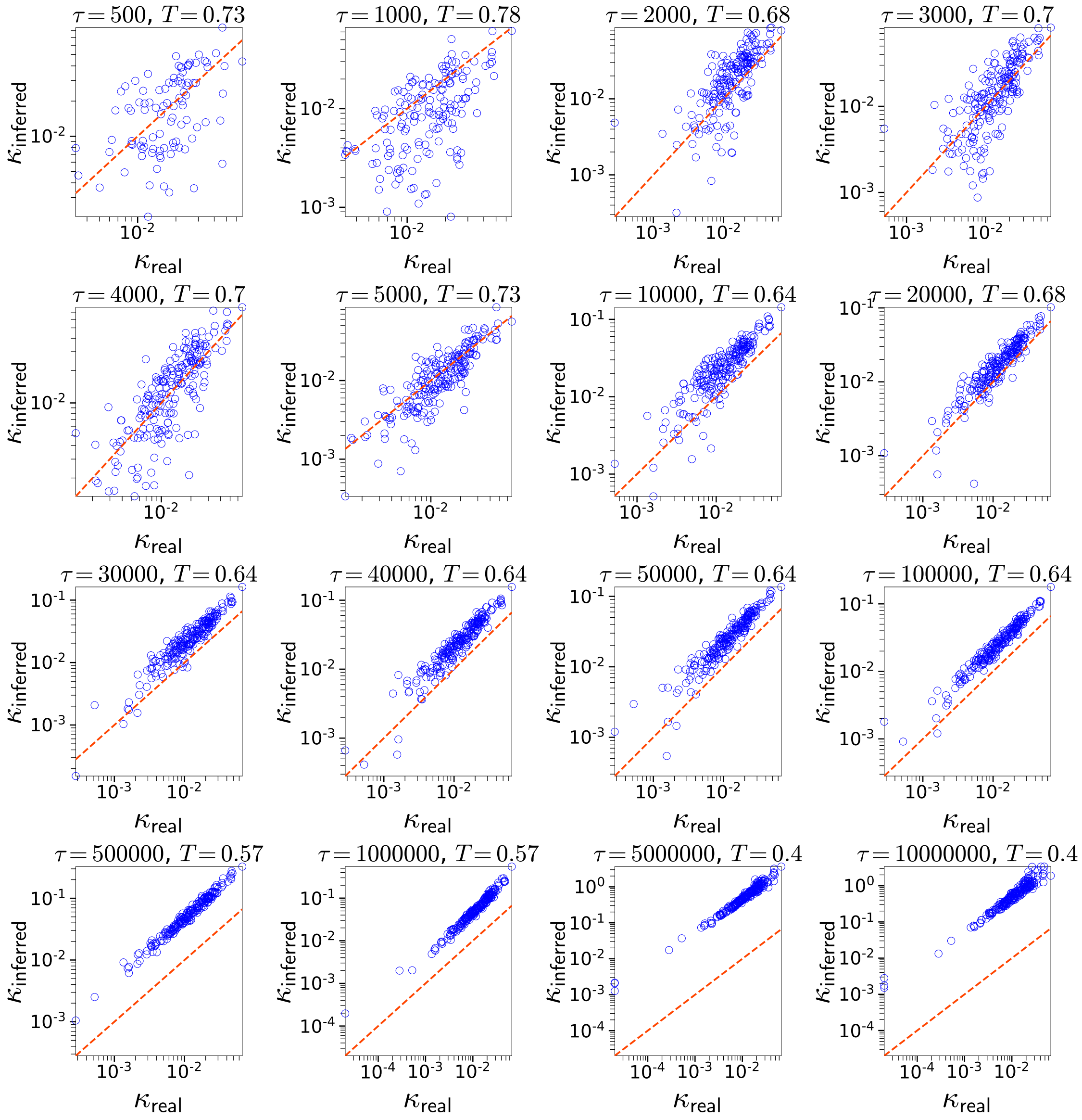

Figure C.27: **Inferred vs. real κ for different aggregation intervals τ.** Same as in Fig. C.19 but for the synthetic counterpart of the office building.

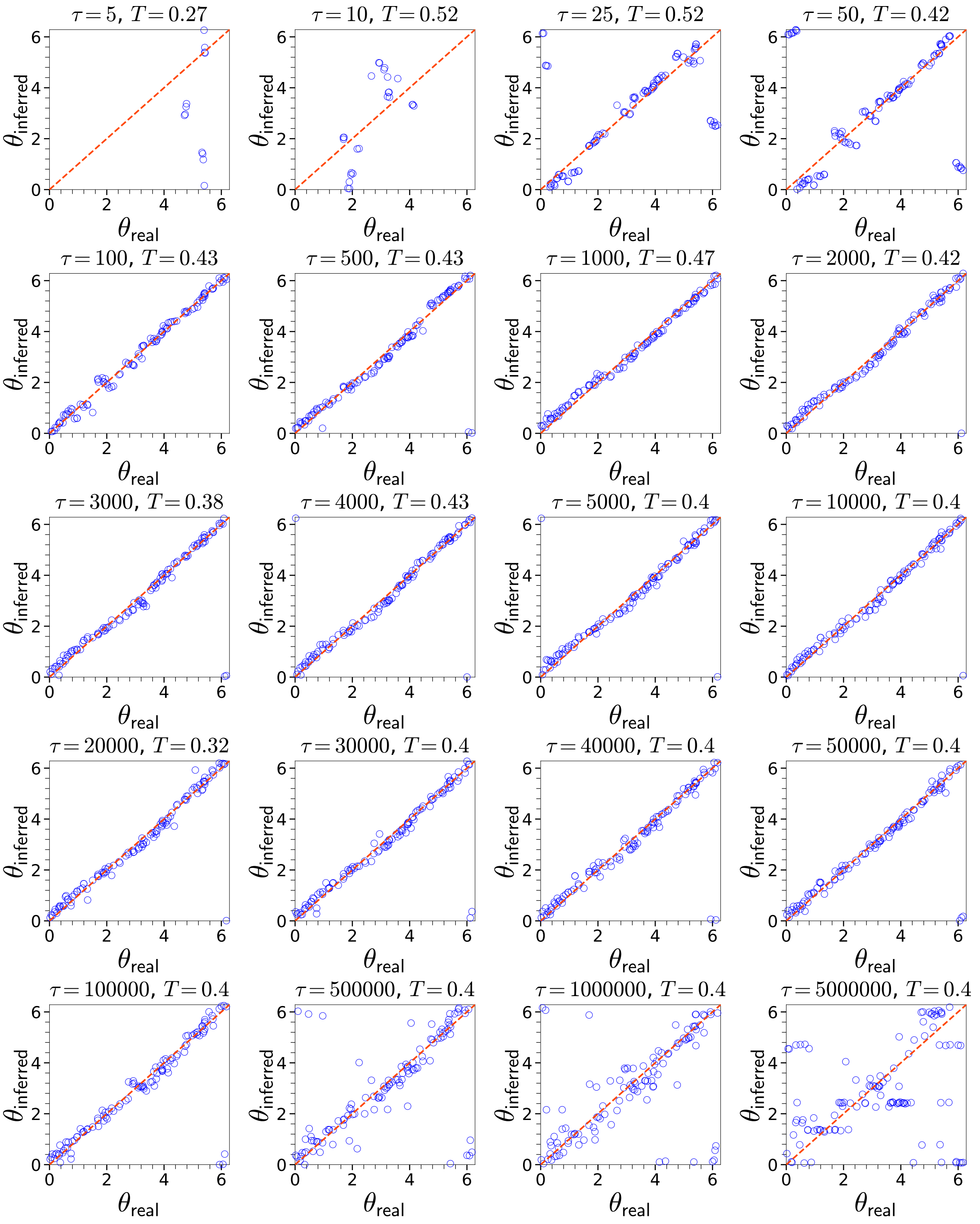

Figure C.28: **Inferred vs. real θ for different aggregation intervals τ.** Same as in Fig. C.18 but for the synthetic counterpart of the Friends & Family.

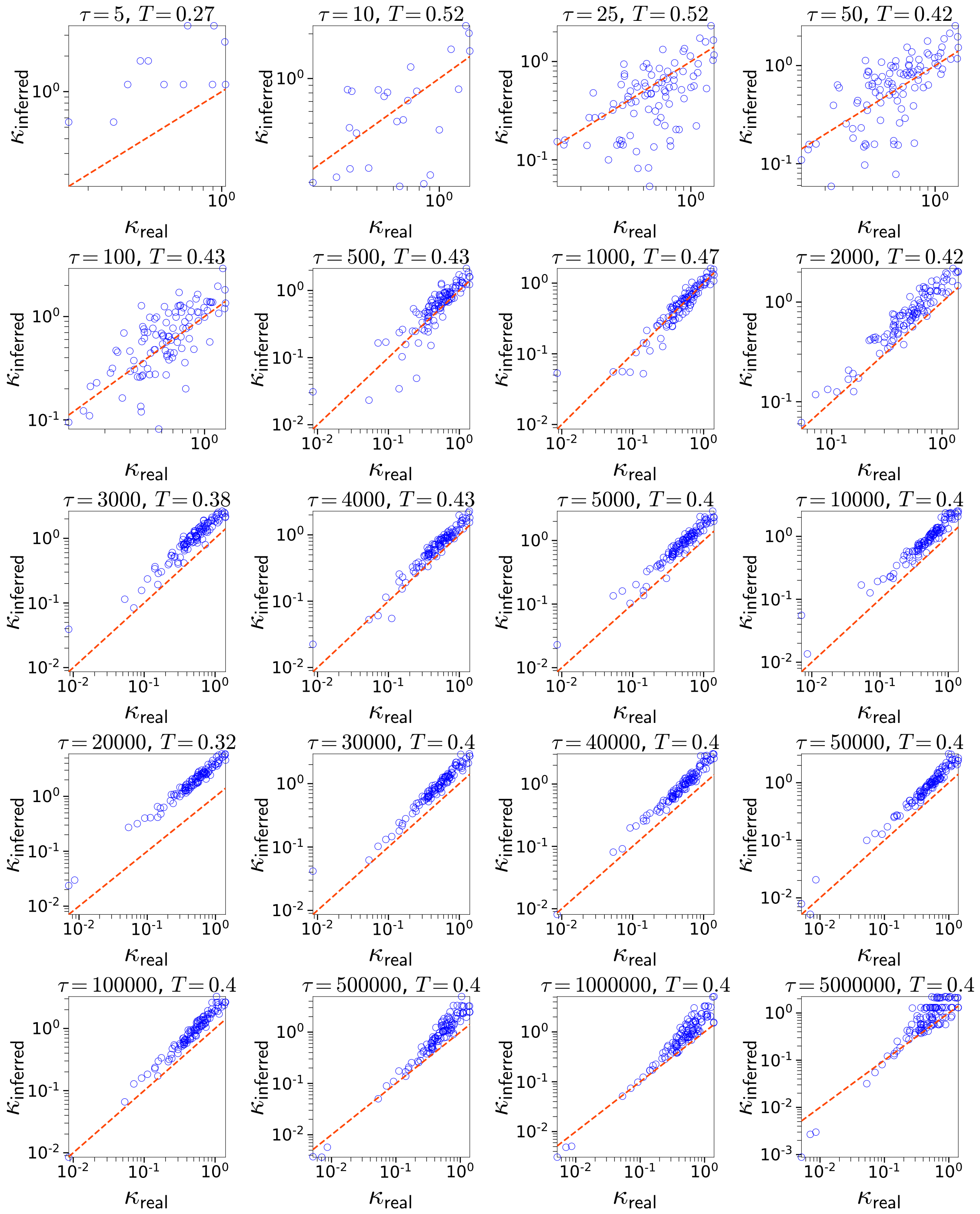

Figure C.29: **Inferred vs. real κ for different aggregation intervals τ.** Same as in Fig. C.19 but for the synthetic counterpart of the Friends & Family.