Study of dynamical social networks of pre-school children using wearable wireless sensors

Sicheng Dai



Numéro National de Thèse : 2022LYSEN015

THESE DE DOCTORAT DE L'UNIVERSITE DE LYON

opérée par

l'Ecole Normale Supérieure de Lyon

en cotutelle avec

East China Normal University

Ecole Doctorale N° 512

École Doctorale en Informatique et Mathématiques de Lyon

Discipline : Informatique

Soutenue publiquement le 23/05/2022, par :

Sicheng DAI

Study of dynamical social networks of pre-school children using wearable wireless sensors

Étude des réseaux sociaux dynamiques d'enfants d'âge préscolaire à l'aide de capteurs portables

---

Devant le jury composé de :

Alain, BARRAT DR, Centre de Physique Théorique, Marseille, Rapporteur

Nathalie, MITTON DR, INRIA Lille Nord-Europe, Rapporteure

Sara, ALOUF CR-HDR, INRIA Sophia Antipolis, Examinatrice

Jean-Philippe, MAGUE MCF, ENS de Lyon, Examinateur

Xiaoling, WANG Professeur, East China Normal University, Examinatrice

Changbo, WANG Professeur, East China Normal University, Co-tuteur de these

Márton, KARSAI MCF–HDR, Central European University, Vienne, Directeur de thèse

Abstract

Wearable wireless radio-frequency identification (RFID) devices provide novel ways to track face-to-face human interactions in various settings. Their application in pre-schools to record daily contacts and verbal interactions of children can reveal how the social networks and linguistic skills of children co-develop over time during early time of schooling.

This thesis addresses this challenge through the design, performance, and analysis of a large-scale social experiment carried out in a French pre-school. In this project we used wireless RFID sensors to collect proximity and voice data from over 200 participants (children and staff) over an observation period of three years, for one week in each month. In parallel, we collected extensive ground-truth data and using periodic questionnaires in which we determined the socio-demographic background and linguistic development of children.

The first goal of the thesis focused on the collection and processing of raw RFID sensor data using conventional data cleaning and signal processing techniques. The second goal was to develop techniques to precisely reconstruct the interactions of participants as temporal networks, using advanced machine learning methods applied on sequential data. Using the reconstructed social networks and recorded linguistic and socio-demographic attributes of children, I conducted a multivariable statistical analysis to study the effects of homophily, inducing over-represented social interactions between linguistically and demographically similar individuals. Finally, a visualisation technique of lineage graphs is presented.

Résumé

Les dispositifs portables RFID offrent de nouveaux moyens de suivre les interactions en face à face des personnes dans divers environnements. Leur application dans les écoles maternelles pour objectif d'enregistrer les contacts quotidiens et les interactions orales des enfants permet de révéler comment les réseaux sociaux et les compétences linguistiques des enfants se développent conjointement au fil du temps pendant les premières années de scolarité.

Ma thèse aborde ce défi à travers la conception, la réalisation et l'analyse d'une expérience sociale à grande échelle menée dans une école maternelle française. Dans ce projet, nous avons utilisé des capteurs RFID pour collecter des données de proximité et vocales auprès de plus de 200 participants (enfants et personnels) pendant trois ans de période d'observation, à raison d'une semaine par mois. En parallèle, nous avons collecté de nombreuses données de vérité terrain et, à l'aide des questionnaires périodiques, nous avons suivi le contexte sociodémographique et le développement linguistique des enfants.

Le premier objectif de la thèse portait sur la collecte et le traitement des données brutes des capteurs RFID à l'aide de techniques classiques de nettoyage des données et de traitement du signal, avec l'alimentation de méthodes personnalisées. Par rapport au deuxième objectif, j'ai développé des techniques pour reconstruire précisément les interactions des participants sous forme de réseaux temporels, en utilisant des méthodes avancées d'apprentissage automatique appliquées aux données séquentielles. En utilisant les réseaux sociaux reconstruits et les attributs linguistiques et socio-démographiques enregistrés des enfants, j'ai mené une analyse statistique multivariable afin d' étudier les effets de l'homophilie, induisant des interactions sociales surreprésent- ées entre des individus linguistiquement et démographiquement similaires. Enfin, une technique de visualisation des graphiques de lignage est présentée.

List of publications

- Liu, Y., **Dai, S.**, Wang, C., Zhou, Z., Qu, H. (2017). GenealogyVis: A System for Visual Analysis of Multidimensional Genealogical Data. IEEE Transactions on Human-Machine Systems, 47(6), 873-885.
- **Dai, S.**, Bouchet, H., Nardy, A., Fleury, E., Chevrot, J. P., Karsai, M. (2020). Tem- poral social network reconstruction using wireless proximity sensors: model selection and consequences. EPJ Data Science, 9(1), 19.
- **Dai, S.**, Bouchet, H., Karsai, M., Chevrot, J. P., Fleury, E., Nardy, A., (2022) Lon- gitudinal data collection to follow social network and language development dynamics at preschool. (Submitted).
- **Dai, S.**, Chevrot, Nardy, A., Karsai, M. Multivariate network homophily patterns between pre-school children. (working paper)

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Márton Karsai. His dedication, erudition and enthusiasm left an indelible influence on me. My four years working with him has been an enlightening and highly rewarding journey, and all the efforts he devoted to my guidance and growth have been indispensable in every aspect. I would also like to express my deep appreciation for my Chinese supervisor Changbo Wang, whom I hold in the highest regard, and who led me to the academia. He has provided unrelenting support since I entered my masters program seven years ago.

I'm grateful for Aurélie Nardy, Hélène Bouchet and Jean-Pierre Chevrot, for their inspiring ideas and efforts in the outstanding collaboration on the DyLNet project. In particular, I would like to thank Eric Fleury, without whom I may never have had the chance to start this journey. I would also like to thank Yuhua Liu, from whom I benefited a lot, especially on the first research paper.

Great appreciation goes out to my jury for spending time and effort examining my work, namely Alain Barrat, Nathalie Mitton, Sara Alouf, Jean-Philippe Mague and Xiaoling Wang.

Most importantly, I owe an inexpressible gratitude towards my parents, who have been the greatest support and encouragement in my life. I am also very grateful for the accompaniment, support and care of my girlfriend Yaqi over all these years.

Contents

Chapter 1

Introduction

1.1 Modelling Human Dynamics

To understand, explain, and predict the behaviour of the natural world, scientists have developed diverse modelling approaches, from the descriptive to the mechanistic. This is true not only in natural sciences such as physics and biology, but also in the social sciences, such as economics, political science, and linguistics [Loayza (1996), Alvarez and Nagler (2000), Kornai (2007)].

Complex, emergent phenomena has been observed in seemingly unrelated natural systems. It is often repeated that in recent decades, researchers across a variety of domains have encountered systems where collective, large-scale behaviour cannot be explained by studying individual constituents of the system in isolation. Systems demonstrating this feature are called *complex systems*. Said differently, a complex system is a system composed of many interacting parts, or agents, which displays collective behaviour that does not follow trivially from the behaviours of the individual parts [Newman (2011)]. The challenge of modelling complex systems is due to the intrinsic heterogeneity of different system components, and the dynamics and interactions among them. The challenges addressed in this work involves not the system itself, but the dynamical processes through which it evolves. Its complexity arises through several distinct properties, such as non-linearity, feedback, emergence and self-organisation [Ladyman et al. (2013)]. Real-world complex systems are as ubiquitous as they are diverse, from networks of biological nerves, financial markets, transport networks and the Internet [Bullmore and Sporns (2009), Mantegna and Stanley (1999), Helbing (2001)]. Building on methods from statistical physics, computational science and mathematics, the study of complex systems has come to define itself as a discipline in its own right, devoted to the study of collective behaviour [Bar-Yam (2002)].

Human dynamics, as a field in complex systems, aims to explain how human behavioural patterns emerge and shape observed collective social phenomena. Human dynamics have long been an intriguing subject of modern science, and of potential use in explaining large scale social, technological and economic phenomena [Vázquez et al. (2006)]. One of the most significant advances that complex systems have made is in understanding human behaviour, such as in decision making, scheduling, collaboration and community formation. For example, in the study of epidemiology, compartmental models such as the SIR (Susceptible, Infectious, Recovered) model, initially assumed populations to be homogeneously mixed. In order to more simulate more realistic scenarios, [Edmunds et al. (1997)] studied the mixing contact patterns among people using questionnaires. The study used multilevel modelling to identify and quantify the contact patterns of potential epidemiological significance. [Wallinga et al. (1999)] proposed using structural characteristics of contact patterns such as the number of contacts per person, transitivity and characteristic path lengths to increase our understanding of the impact of human contact networks on the spread of infectious disease. [Colizza et al. (2006)] forecast global epidemics by considering detailed worldwide air travel infrastructure, complemented with census population data. [Musse and Thalmann (1997)] developed an approach based on autonomous virtual crowds to describe the emergent behaviour observed in large groups of people in the real world. When studying leadership in organisations, [Marion and Uhl-Bien (2001)] used complexity theory to understand the emergence of structure, fitness, and innovation in organisation, and how emergence is influenced by leadership behaviour. An often-cited example is the *Zachary's karate club*, first studied in [Zachary (1977)], where the author studied a karate club and defined a network that represents its 34 members as nodes and interactions among them as links. After the club split into two parts, Zachary correctly classified all but one member into the groups they actually joined using the Ford–Fulkerson algorithm [Ford and Fulkerson (1956)]. This dataset became an illustrative example of social networks, including in the pioneering study of community structure in [Girvan and Newman (2002)]. Not being limited to using digital records to portray communication network's dynamics and phenomena, human dynamics also include research topics in the traditional scope of social science. This was recognised in [Cameron and Larsen-Freeman (2007)] that introduced complexity theory as a metaphor for systems in applied linguistics owing to their shared characteristics of temporal evolution and heterogeneity, as well as in

[Horn (2008)], which identified complexity as a new scientific paradigm, could offer the potential to develop a new kind of social science. Sustained research efforts in this direction have lead to the emergence of a new field, that of *computational social science* [Lazer et al. (2009a)], which aims to observe and understand social phenomena with an emphasis on data.

1.2 Networks

The most often used abstraction of a complex system is that of a *network*. Borrowing the language of graph theory, a system can be represented by a network where nodes (vertices) represent the system's components and links (edges) represent their interactions. Such representation provided a universal model to study various systems even with great difference. For example, protein–protein interaction network was initiated in model organisms' biological processes [Walhout et al. (2000)], where proteins are nodes and their interactions are edges. In social network the nodes are social entities, which could be people, groups, web sites, publications, and links are connection between the entities. There are some fundamental definitions and attributes that can be computed to analyse networks and distinguish how they contrast to each other. *Complex network* is an inevitable concept when studying complex systems. Behind each complex system there is an intricate network that encodes the interactions between the system's components [Barabási and Pósfai (2016)]. Again borrowing from graph theory, complex networks can be characterised by their topological features, which are irregular and heterogeneous, in contrast to chains, grids, lattices or fully-connected graphs. Network representation has been applied in a wide range of real complex systems in nature and society [Bhalla and Iyengar (1999), Donges et al. (2009), Cohen et al. (2012), Bassett and Sporns (2017)] resulting in network science being one of the fastest growing disciplines in recent decades.

To understand the structure, formation and dynamics of empirical systems, various network models have been proposed, and we discuss them below.

1.2.1 Models of complex networks

Erdős–Rényi model. The story of complex networks started with two closely related models proposed by Paul Erdős and Alfréd Rényi for generating random graphs [Erdos et al. (1960)]. For a more precise definition, here we use the representation $G(N, p)$, where N is the number of nodes and p is the probability that

an edge exists between a randomly selected pair of nodes. Each edge generated in an Erdős–Rényi (ER) model network is independent from each other, which straightforwardly determines the distribution of their degrees, i.e. the number of connection each node has in the graph. Degrees in an ER graph follow a binomial degree distribution. That is, for a randomly chosen node, the probability of its degree to be equal to k is

$$P_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}. \tag{1.1}$$

Most real networks are sparse, meaning that the average degree is much smaller than the system size, that is, $\langle k \rangle \ll N$. With this condition, when the ER graph is large ($N \to \infty$), its degree distribution (Eq. 1.1) tends to the Poisson distribution,

$$P_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}. \tag{1.2}$$

The Poisson distribution has the advantage of analytical simplicity since its properties are independent of the network size and depend only on a single parameter $\langle k \rangle$ [Barabási and Pósfai (2016)].

One important characteristic of complex networks is their connected component structure, which is a connected subgraph that is not part of any larger connected subgraph. Among the connected components of a network the *largest component* deserves special attention, as its size (denoted N_G) determines the behaviour of any large scale, or global phenomena (such as epidemics, opinion dynamics and synchronisation) observable on the structure. For random networks, one would expect that the N_G grows gradually with increasing of $\langle k \rangle$. However counter-intuitively, N_G/N stays close to 0 for small $\langle k \rangle$, even with increasing average degree. Once $\langle k \rangle$ exceeds a critical value, N_G/N starts to increase rapidly, resulting the emergence of a giant component. More precisely, for inreeasing $\langle k \rangle$, networks generated by the Erdős–Rényi model pass through four regimes,

- $\langle k \rangle < 1$ ($Np < 1$): sub-critical, the largest component in this regime is in $O(\log N)$ scale, only a small number of links exist in this regime, therefore only small components will be observed.

- $\langle k \rangle = 1$ ($Np = 1$): critical point, the largest component in this regime is in $O(N^{\frac{2}{3}})$ scale. Compared with sub-critical regime, the size of the largest component starts to show a significant raise. The size of connected components in this regime follows the power law distribution.

- $\langle k \rangle > 1$ $(Np > 1)$: super-critical, the size of largest component in this regime is $\sim yN$, where $y = p - \frac{1}{N}$. This regime has the most relevance to real systems, a giant component covers large fraction of nodes and no other component will have size larger than $O(\log N)$.
- $\langle k \rangle > \ln N$ $(Np > \ln N)$: connected, with sufficiently large p, the giant component will include almost all nodes, where $N_G \simeq N$, therefore the whole network becomes connected.

The critical point of random network is arguably the most noteworthy regime of all. The phase transition in many physical systems is resembled by the emergence of giant component in random networks near critical point, representing the system which goes through a continuous phase transition such as water-ice transition and magnetism [Barabási and Pósfai (2016)].

Average path length. Here we focus only on the connected regime for simplicity. Distance between two nodes in a connected network is defined as the number of links one need to pass through the shortest walk from one node to another. For most values of p, the diameter, i.e. the maximum of any distances in a given network, with the same N and p have the same values concentrated at $\log N / \log Np$ (or equivalently $\log N / \log\langle k \rangle$) [Chung and Lu (2001)]. Consequently, the average path length of a random graph is

$$\langle l \rangle = \frac{\ln(N/z_1)}{\ln(z_2/z_1)} + 1, \tag{1.3}$$

where z_m is the average number of m-th nearest neighbours. With the Erdős–Rényi model, $z_1 = \langle k \rangle$, and $z_2 = \langle k \rangle^2$, then $\langle l \rangle = \dfrac{\ln N}{\ln\langle k \rangle}$, which is the same as the diameter of the network.

Clustering coefficient (in a local sense) is defined as the number of connected neighbours of a node with degree k divided by the number of possible connected neighbours it can have, that is $(k(k-1)/2)$. If we consider a node in a random graph and its neighbours, the probability that two of its neighbours are connected is equal to the probability that two randomly selected nodes are connected. Consequently, the clustering coefficient of a random graph is

$$C_i = p = \frac{\langle k \rangle}{N}. \tag{1.4}$$

Watts–Strogatz model. While the Erdős–Rényi model is concise and enlightening, the assumptions that edges are independent, and that edges between all node pairs

are present with same probability, fails to model many phenomena observed in real networks. Many biological, technological and social networks exhibit high clustering, like regular lattices, yet have small average path lengths, like random graphs [Watts and Strogatz (1998)]. Many network models have been proposed to better explain such phenomena. The first, and most well-knwon, was the Watts–Strogatz model, which generates so called *small-world* networks, combining the small-world phenomenon (colloquially known as "six degrees of separation") with properties such as small average shortest path lengths and high clustering. The Watts–Strogatz model constructs networks by interpolating between regular and random networks. The model starts from a ring lattice with N vertices and k edges per vertex, then each edge will have probability p to be randomly rewired. This construction allows the network to interpolate between order ($p = 0$) and disorder ($p = 1$).

Average path length. We could analyse by first checking two extrema of p. As $p \to 0$, $\langle l \rangle \sim N/2k \gg 1$, while $p \to 1$, the network turns into a random network, where $\langle l \rangle \sim \ln(N)/\ln(k) \gg 1$. The rapid drop in $\langle l \rangle$ is caused by the introduction of a few long-range edges, which result in small-world networks. For small p, each short cut has an effect on $\langle l \rangle$ owing to not only the reduction of distance between the pair of vertices that rewiring connects, but also the pass-through effect between their neighbourhoods, and neighbourhoods of neighbourhoods. A large drop in average path length can be observed by the time $p = 10^{-2}$. A more in-depth analysis is performed in [Barthélémy and Amaral (1999)], where crossover size $n^* \sim p^{-\tau}$ with $\tau \approx 1$, and average distance $\langle l \rangle$ between any two vertices of the network is a scaling function of n/n^*. This implies that rewiring a finite number of links already has a strong influence on $\langle l \rangle$ of a network.

Clustering coefficient. In the study of [Barrat and Weigt (2000)], the authors define $C(p)$ as the average of cluster coefficient over all vertices with each edge has probability p to be randomly rewired. Each vertex is assumed to have $2k$ neighbours for simplicity. For $p = 0$, it is easy to see that the number of links between these neighbours is $N_0 = 3k(k-1)/2$. Then $C(0) = \dfrac{3(k-1)}{2(2k-1)}$. For $p > 0$, two neighbours of node i that were connected at $p = 0$ are still neighbours of i (their link to node i is not rewired) and linked together with probability $(1-p)^3$, up to terms of order $\frac{1}{N}$. Instead of considering $C(p)$ which computes over all nodes, the authors define $\tilde{C}$ as the ratio of the mean number of links between the neighbours of a vertex and the mean number of possible links between the neighbours of a

vertex, it could calculated as

$$\tilde{C}_p = \frac{3(k-1)}{2(2k-1)}(1-p)^3. \tag{1.5}$$

By comparing these two characteristics with their counterparts in the Erdős–Rényi model, this model captured two crucial features of many real world networks: the first is the small world property which is depicted by $\langle l \rangle \sim \ln N$; the second feature is high clustering coefficient that obviously exceeds its counterpart in Erdős–Rényi model, where $C_i \sim \langle k \rangle / N$.

Barabási-Albert model of scale-free networks. The Watts–Strogatz model generates networks with small-world features: small average path length and large clustering coefficient, which is typical in real-world networks. This model takes a significant step towards reproducing the characteristics of real-world complex system. However, it exhibits degree distributions that fail to match with those of real-world networks. For example, the Internet, online social networks and citation networks typically contain a few nodes, called hubs, with unusually high degree. In [Albert et al. (1999)], the authors found that the distributions of in-degree and out-degree of World Wide Web follow power-law degree distributions $P(k) \sim k^{-\gamma}$ with a tail ranging over several orders of magnitude with exponent $\gamma_{out} = 2.45$ and $\gamma_{in} = 2.1$. This is in strong contrast to the Poisson distribution predicted by the Erdős–Rényi random graph model and the bounded distribution predicted by the Watts–Strogatz model. Networks with power-law degree distribution are called *scale free network*, due to the fact that in power-law distributions, all moments larger than $\gamma - 1$ diverge. The moment of degree distribution is defined as

$$\langle k^n \rangle = \int_{k_{min}}^{\infty} k^n P(k) dk. \tag{1.6}$$

This scale-free property is essential to produce network structures such as hubs. Also, the Watts–Strogatz model sets a fixed number of nodes and thus cannot be used to model networks with growth. The two aforementioned properties of networks can be generated by a preferential attachment model proposed in [Barabási and Albert (1999)]. This model incorporates two important properties,

- Growth, which describe that networks expand over time with new vertices joining.
- Preferential attachment, which regulates that new vertices attach preferentially to vertices with higher degree.

The construction of Barabási-Albert model is as follows. Starting with m_0 vertices, at every time step a new vertex with $m \leq m_0$ edges will be introduced in the network and link to m different vertices already present in the network. Then, to introduce the preferential attachment mechanism, each new vertex connects to existing nodes with a probability that is proportional to the degree of the latter. Formally, the probability that the new vertex is connected to vertex i with degree k_i is

$$p_i = \frac{k_i}{\sum_j k_j}. \tag{1.7}$$

Evolving in this way, after t time steps the network turns to a scale-free network with $t + m_0$ vertices and mt edges. The rate at which a vertex i acquires edges is

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\frac{1}{2}}, \tag{1.8}$$

where t_i is the time at which vertex i was added to the system. The so-called rich-gets-richer phenomenon are able to be observed in this way. Based on Eq. 1.8, we could then calculate the exponent γ of the generated network. The probability that a vertex i has a connectivity smaller than k, $P(k_i(t) < k)$, can be written as $P(t_i > m^2t/k^2)$, which is equal to $1 - P(t_i \leq m^2t/k^2) = 1 - m^2t/k^2(t + m_0)$. Then the probability density $P(k)$ can be obtained from $P(k) = \partial P(k_i(t) < k)/\partial k$ which will leads to the stationary solution

$$P(k) \sim k^{-\gamma}, \tag{1.9}$$

where giving $\gamma = 3$, independent of m, reproduces the observed scale-free distribution. Furthermore, it is easy to modify the model to account for exponents different from $\gamma = 3$, for example, when introducing new vertices, we could assume that a fraction p of the links is directed, we obtain $\gamma(p) = 3 - p$. Also some networks evolve by adding links not only between new vertices and existed counterparts, but also between existed vertices, this modification could also alter the value of γ [Barabási and Albert (1999)].

Average path length. In the work Cohen and Havlin (2003), author used analytical arguments to show that scale-free networks with $2 < \gamma < 3$ have a much smaller diameter as $d \sim \ln\ln N$, featuring an ultra-small network. For $\gamma = 3$, the analysis yields $d \sim \ln N/\ln\ln N$, while for $\gamma > 3$, $d \sim \ln N$.

Clustering coefficient. Here we assume that one node is added for each time step, and that nodes are indexed by the time step they are added to the network. When

adding node j to the network, the probability for one link of node j to connect with node i is the ratio of the degree of the node i, k_i, and the sum of all node degrees in the network, $2mj$. The probability for the existence of a link from j to i is given by

$$\text{Prob}\{(ij)\} = \frac{m}{2}(ij)^{-\frac{1}{2}}. \tag{1.10}$$

The local clustering of node l in a network of size N is denoted by $C_l(N)$. Taking into account expectation values and treating the nodes as a continuum, it can be calculated to be

$$C_l(N) = \frac{\int_1^N di \int_1^N dj \text{Prob}\{(ij)\}\text{Prob}\{(il)\}\text{Prob}\{(jl)\}}{{k_l}^2(N)}, \tag{1.11}$$

using ${k_l}^2(N) = m^2 N/l$ (from Eq. 1.8) and approximating the total number of neighbors by $k_l^2/2$. Then, substituting Eq. 1.10 into Eq. 1.11, we derive

$$C_l(N) = \frac{8}{m}\frac{(\ln N)^2}{N}. \tag{1.12}$$

A more detailed deduction can be found in the work of [Klemm and Eguiluz (2002)].

As we could see from Eq. 1.12, Barabási–Albert model fails to produce the high levels of clustering that is frequently observed in real networks, something that the Watts-Strogatz model succeeded in doing. A more general preferential attachment is studied in [Krapivsky et al. (2000)], where attachment probability is replaced by a more general form that proportional to

$$p_i = \frac{k_i^\alpha}{\sum_j k_j^\alpha}. \tag{1.13}$$

where α is used to tune the preferential attachment. Moreover, in many real world networks, the oldest node may not always have the most links. That is, a newcomer may attract a large number of links within a short period of time and become the largest hub. A variant of the Barabási–Albert model called *Bianconi-Barabási model* [Bianconi and Barabási (2001)] is proposed in to simulate this observed phenomenon.

1.2.2 Temporal networks

While early network studies focused on static structures, more advanced and realistic models have over time been proposed in accordance with observations of real-world systems. One such observation was of their tendency to evolve in time, through mechanisms such as the addition or removal of nodes, and the forming and severing of links, probabilistically and sometimes as a function of local topological features. These models have proved successful in explaining observed phenomena in many empirical systems. However when the network and its dynamic process evolve at comparable timescales, classic network models are incapable of capturing the fine-grained time-varying nature of real systems connectivity patterns [Karsai et al. (2014)]. For such cases, the temporal structure can affect the dynamics of systems interacting through the network [Holme and Saramäki (2012)], from the spread of disease to the diffusion of information in online social settings. For example in information propagation, two peers connected by a link in static network indicates information could pass between between them at any time. In contrast, in case of temporal networks, information can be passed between connected nodes only at the time of their interactions, a property which could greatly affect properties such as reachability and transitivity of the network [Nicosia et al. (2013), Williams and Musolesi (2016), Badie-Modiri et al. (2020)]. Also in [Li et al. (2017)], the authors demonstrated that temporal networks, compared with their aggregated counterpart, could reach controllability faster, they demand orders of magnitude less control energy, and they have significantly more compact control trajectories to reach its final states. Therefore if applicable, it is essential to incorporate temporal dynamics between pairs of nodes into the modelling process. Thanks to the digital revolution and development of related technologies, large volumes of network data with time and other metadata details on nodes and links have been generated and collected.

Temporal networks are network structures with edges vary with time. The inclusion of temporal dimension has brought network modelling to another perspective that other types of network modelling are unable to hold. Therefore temporal network could provide better understanding about the mechanism behind the observed phenomena and in turn could greatly improve their predictions. In many studies the temporarily of interactions is considered. One of the most well-studied types of temporal networks are *human proximity networks*, which measure contact

between individuals in close range, as well as the duration of this contact. Various technologies have been developed to collect such data, e.g. radio-frequency identification (RFID) [Isella et al. (2011), Barrat and Cattuto (2013), Kibanov et al. (2014), Voirin et al. (2015)], autonomous low-power wearable devices (LPWD) [Duval et al. (2018)] and infrared sensors [Takaguchi et al. (2011)]. This kind of experiment are widely applied in multiple settings, like schools [Stehlé et al. (2011b), Fournet and Barrat (2014)], museums or conferences [Isella et al. (2011)]. *Brain networks* is another example of temporal network application. In the work [Bassett et al. (2015)], authors built temporal networks from functional magnetic resonance imaging (fMRI) data and extracted groups of brain regions (network communities) that were coherently active in each time window, with each group is assumed to be responsible for a specific cognitive function. Temporal networks have also been applied to help improving *epidemic modelling* [Salathe et al. (2012), Pastor-Satorras et al. (2015)], the importance of such works cannot be stated enough under the current context of the COVID-19 global pandemic. In the work of [Stehlé et al. (2011a)], the authors considered high-resolution face-to-face interactions data between the attendees at a conference, and used an SEIR (Susceptible, Exposed, Infectious, Recovered) model to simulate the spread of epidemics along these interactions. [Gautreau et al. (2008)] studied the problem of the arrival time of epidemics spread with a metapopulation model on large-scale networks. This work proposed an easily computable quantity that depends only on the links weights and nodes populations to accounts for the average arrival time of the spread in each city. Recently, [Cencetti et al. (2021), Barrat et al. (2021)] used empirical high-resolution contact data collected by digital devices (such as Bluetooth on mobile phone) to build up models to test the effectiveness of different policies to contain the epidemic.

1.2.2.1 Representation of temporal networks

To represent temporal networks, one needs to incorporate information about the time of contacts between pairs of nodes. How this is done depends on the type of contact and the process the temporal network supports. Typically there are two categories: "lossy representation" where some information of the original temporal network is lost, and "lossless" representation, which carries all information. This thesis used mostly lossless representations called "*event lists*". In this representation, temporal networks are defined as $G_t = (V, E_t, T)$, where V denotes vertex set composed of vertices $v \in V$. The edge set E_t is defined as a set of *events*

(or temporal edges), which represent the interactions in a temporal network. To incorporate temporal information, events take the form $E_t \subset V \times V \times [0, T]$, that indicates two nodes have interaction over a time span T. A temporal network is then described as a sequence of such events. Explicitly, a temporal edge is a triplet (u, v, t), indicating an interaction between vertex u and v at time $t \in [0, T]$. Just like static networks, the events (temporal edges) is directed and undirected depending on the context. For many cases the events contains more information than just the event time t such as duration or weight, thus it is necessary to extend the definition of event to $E_t \subset V \times V \times T \times \prod_i A_i^e$, where A_i^e is a set of additional information of an event based on on the actual scenario.

There are also other popular representations, for example *graph sequences* or *snapshots*. These are representations that present temporal network as sequences of static graphs. This representation allows the use of methodology and measures from well-established static network research. By doing this, we could perceive temporal network by looking at characterisations change at different time. One disadvantage is that this representation is effective only for networks with low temporal resolution, which means the dynamic of edges are not rapidly changing, for instance ecological networks. Another representation called *stream graph* or *link stream* is proposed by [Latapy et al. (2018)], where interactions are modelled as $S = (V, W, E, T)$. That is, a node set V and time instants set T are defined like their counterparts in event list. What makes them different is the *temporal nodes* defined by $W \subseteq T \times V$, and a set of links $E \subseteq T \times V \otimes V$ such that $(t, uv) \in E$ implies $(t, u) \in W$ and $(t, v) \in W$, where $u, v \in V$. This representation has the advantage of directly dealing with interactions over time, which includes both temporal and structural nature instead of simply combining the two. Equivalent of most elementary and more advanced network concepts such as density and centrality are also defined for stream graphs and link streams.

1.2.2.2 Temporal structure and characteristics

As outlined above, static networks can be characterised by certain measures of their topology, such as the degree, clustering coefficient and path lengths. Corresponding quantities in temporal networks need to be revised. Here we summarise some concepts and characteristics which are important for temporal network analysis.

Temporal path [Pan and Saramäki (2011)] tells the reachability between nodes, which constrains the dynamics taking place on the network. In a static graph, a sequence of adjacent edges, sharing a common ending node, could be constituted as a path. Such a path indicates all nodes that are connected and reachable (for undirected networks). For the temporal network, temporal paths must be constrained to sequences of events that activate chronologically. For example considering a set of events $\{(v_1^s, v_1^e, t_1), (v_2^s, v_2^e, t_2), \cdots, (v_n^s, v_n^e, t_n)\}$ (here the node index indicates the event number, not the actual node), it could form a temporal path if for $\forall i \in [1, n-1]$, $v_i^e = v_{i+1}^s$ and $t_{i+1} > t_i$, with length defined as $t_n - t_1$. Temporal path is non-symmetric, like path in a directed graph, a path from i to j does not mean the same from j to i. However different from directed graph, temporal paths are neither transitive, which means in temporal network, if there is a path from i to j and from j to k does not necessary indicates there is a path from i to k. What's worth noting is that temporal paths are themselves temporal, and are only valid paths during a specific time interval, this also means there can exist multiple paths from i to j with different starting time t.

Temporal distance Another basic measure in static network is distance, defined as the smallest number of (non-weighted or weighted) links of a path connecting two nodes. In temporal networks there are many ways of defining distances. For example, a commonly used definition is for nodes i and j to define a distance as the shortest time it would take to reach j from i at time t along a temporal paths [Pan and Saramäki (2011)]. This distance is commonly denoted as $\tau_{ij}(t)$. However with the temporal feature, $\tau_{ij}(t)$ changes with t, thus it is useful to characterise temporal distances with their average τ_{ij} over all time $[0, T]$.

Burstiness and interevent time statistics. Recall that we have observed abundant networks with power-law degree distribution. Just like topological measures, certain temporal structures of temporal network are intriguing. A very common temporal structure we could examine is the inter-event time distribution a.k.a. distribution of the gap of starting time between the events on the same pair of nodes. If the probability of an event is time-independent, we would naturally think of Poisson process. In this case the inter-event time τ between two consecutive events follows an exponential distribution [Haight (1967)]. Interestingly, for varies human involved temporal network, it has been discovered that such timings exhibits bursty non-Poisson character [Barabasi (2005)]: substantial events happens during

a relatively short period, then followed by long periods of inactivity. This kind of phenomena are marked with emergent scale-free inter-event distributions, indicating present temporal heterogeneity in their dynamics. Detailed mechanism that forms the bursty phenomenon is introduced in detail in [Barabasi (2005), Goh and Barabási (2008), Karsai et al. (2018)].

1.2.3 Social networks

Studies of social networks are especially focusing on structures where nodes represent people and the edges represent interactions between them such as social or family acquaintances, common interests, friendship or religious affiliation. Social networks are studied ubiquitously, such as in anthropology, economics, sociology, as well as other related interdisciplinary fields.

One interesting characteristic of social networks are is their community structure, which has been analysed in [Bedi and Sharma (2016), Chunaev (2020), Fortunato (2010)]. Correct partitioning of a social network can give important insights to understand network structure and is helpful in studying information diffusion and marketing. Another important research directions in social networks is on how micro-level interactions coded in the network could shape macroscopic phenomena like communities themselves. Many studies focus on the mechanisms in the growth of networks and social tie creation. Georg Simmel in 1908 proposed the concept of *triadic closure*, describing the property that given three people A, B, and C, if the connections $A - B$ and $B - C$ exist, there is a high probability for A and C to be connected. [Bianconi et al. (2014)] showed this mechanism alone is capable of generating systems with community structure. To further investigate, links in network could be assigned with *weight*. More specifically, in social network the strength of a social tie is the combination of time, emotional intensity and intimacy. Dichotomy could divide a social tie into strong or weak by certain criteria e.g. multiplexity. In the work of [Granovetter (1973)], considered that weak ties act as the role of "bridge" that links between tightly connected communities and is vital to global information dissemination. In contrast, strong ties shapes local connectivity. This effect is also observed in mobile phone data [Onnela et al. (2007)], where authors found that removing weak ties could cause social networks to fall apart after a phase transition.

Another social tie creation mechanism depends on node properties, for example a person's age, gender, education, occupation, workplace or socioeconomic status. The tendency of people to associate more than randomly expected with others who are similar to themselves is called *homophily*. This link creation mechanism has been observed in several studies like [McPherson et al. (2001), Newman (2018), Shrum et al. (1988), Aiello et al. (2012)]. It has been found to work together with triadic closure to shape social networks into communities [Kossinets and Watts (2009)]. The effects of homophily, however, are hardly distinguishable from the consequences of another process called *influence*, which describe the mechanism that people adopt feature or habit from their connected peers through social or interpersonal interaction. The influence helps the forming of several macroscopic social phenomena, such as simple information spreading [Barthélemy et al. (2004), Madar et al. (2004)] and the emergence of other collective complex phenomena [Granovetter (1978)]. The core difference between homophily and influence is who comes first, the similarity or the relationship. Regardless of the entanglement between homophily and influence, they are likely to result in significantly different dynamics and purely data-driven observations sometimes struggle to distinguish them. Thus, this question sets one of the most important challenges of computational social science [Aral et al. (2009)].

1.3 Collection of Digital Behavioural Data

Earlier observations of human behaviour were traditionally conducted using methods such as interviews, surveys, direct observation etc. While these methods have achieved significant insights of several key problems, the capacity of data collection is overstretched facing increasing demand with regard to meticulousness, accuracy, timeliness, long-term and large-scale. In recent years however, the rapid growth of Internet access, the maturation of satellite navigation and the increasing penetration of mobile devices removed these limitations [Lazer et al. (2009b), Vespignani (2009)]. Subsequently, enormous datasets have been gathered to fuel the research of human dynamics. In return, the studies of these datasets have revealed in-depth knowledge inducing pioneering works in this direction.

As email became a means of day-to-day communication, email exchange logs formed a network with email address as nodes, and email exchanges as links. [Ebel et al. (2002)]. these networks were found to emerge with scale-free degree distributions and small-world behaviour. [Newman et al. (2002)] studied structure of email

network and found random vaccination (install anti-virus software) of computers has little effect on containing the spreading of virus, while targeted vaccination are more promising. [Eckmann et al. (2004)] discovered that email network develops self-organised structures which arise from temporal correlations when users act in a synchronised manner. This work also observed distribution of the response time roughly approximated by a power-law.

The popularity of mobile communication provided rich spatio-temporal data[Onnela et al. (2007)] exploited mobile phone call records to construct a society-wide communication network which was inaccessible at the societal level before. This work demonstrated a local coupling between tie strengths and network topology, and showed that this coupling has important consequences for the network's global stability if ties are removed. [Candia et al. (2008)] showed that spatio-temporal anomalies can be described using standard percolation theory tools, and the heavy-tail property of inter-event time of consecutive calls was once again observed. Besides the studies that focused on structure and topology, there are also plentiful research in the field of computational social science that benefit from spatio-temporal data, [Blumenstock et al. (2015)] used call detail records of Rwanda as an example to prove that the past history of mobile phone use of an individual is related to his/her *socioeconomic status* (SES). The increasing availability of large human mobility datasets has enabled new systematic studies of mobility patterns [Gonzalez et al. (2008), Noulas et al. (2012), Lotero et al. (2016), Carra et al. (2016)]. Furthermore, compared with traditional data collection methods, the accurate and fine-grained mobile data also provides a powerful weapon against epidemic. Mobile phone data have been proposed to assess potential drivers of spatio-temporal spread, and to support contact tracing using data from cellular base stations or Bluetooth. [Grantz et al. (2020), Budd et al. (2020)]. Except for location acquired from the cellular base station, the finest level of accuracy on locations and trajectories is from *Global Positioning System* (GPS) data, which has better resolution compared with data from base station and is broadly applied in tasks such as navigation, traffic monitoring and congestion prediction [Zheng et al. (2009; 2010), Bazzani et al. (2010)].

The emergence of online social platforms, like *Twitter, Facebook* and *Weibo*, generated data with unprecedented speed and connected people in unprecedented scales. [Leskovec et al. (2007)] analysed blog network and found out that the decline of a post's popularity follows a power-law instead of an exponential decay

as originally expected. This work also studied cascading phenomenon, and it discovered that most rapid cascades of information adoption are tree-like, stars or chains. [Wu et al. (2011)] studied information flow pattern among the members of an online ecosystem. Users were first classified into "elite" and "ordinary" groups and revealed that messages from elite users reached the masses indirectly, but via a large population of intermediaries. The rich messages that users posted on social platform act as a great repository to estimate users' social class and socioeconomic status [Preoţiuc-Pietro et al. (2015), Lampos et al. (2016), Luo et al. (2017)], and could further extend to studies on inequality, discrimination. However, with its openness and accessibility, the online social platforms also facilitated widespread dissemination of misinformation. To tackle this problem, many studies focused on the detection of misinformation by studying its diffusion pattern, message and source. [Shu et al. (2017), Shin et al. (2018), Wu et al. (2019), Shu et al. (2020)].

In contrast to large online social platforms, several experiments to record human proximity has been conducted as introduced in Section 1.2.2. Such experiments focus on relatively small groups with more specialised purposes. [Elmer et al. (2019)] conducted two experiments with RFID badges in order to accurately measure face-to-face interactions.

1.4 Machine Learning Methods

The ever-growing volume of data produced by electronic devices demands automated statistical learning methods [Alpaydin (2020)]. Without aiming a comprehensive summary of the whole field, next I summarise those techniques, which I applied during my work for the solution of various types of classification, inference, or prediction problems.

1.4.1 Logistic regression

Logistic regression dates back to 1830s when Pierre Verhulst proposed *logistic function* to model population growth. The logistic function is

$$f(x) = \frac{L}{1 + e^{-(x-\mu)/\gamma}} \tag{1.14}$$

The standard logistic function, where $L = 1, \mu = 0, \gamma = 1$, is the famous *sigmoid function*.

Logistic regression as a statistical model was proposed in [Cox (1958)]. It normally applies to model the probability of binary dependent variable. Consider a binary regression problem, given a dataset $D = (\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \cdots, (\mathbf{x_N}, y_N)$ where $\mathbf{x_i} \in \mathbb{R}^M$ and $y_i \in \{0, 1\}$, and $i = 1, 2, \cdots, N$. In this model we could construct a decision boundary defined as $\theta^\mathsf{T}\mathbf{x} + b = 0$ where $\theta \in \mathbb{R}^M$. For simplicity we append 1 for each $\mathbf{x_i}$ as corresponding constant term so $\mathbf{w} = (\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(M)}, b)$ to include b. To connect the classification probability $p(y = 1)$ and the input vector x, we apply the 1.14,

$$\begin{aligned} P(y = 1|\mathbf{x}; \mathbf{w}) &= \frac{1}{1 + e^{\mathbf{w}^\mathsf{T}\mathbf{x}}} \\ P(y = 0|\mathbf{x}; \mathbf{w}) = 1 - P(y = 1|\mathbf{x}; \mathbf{w}) &= \frac{e^{\mathbf{w}^\mathsf{T}\mathbf{x}}}{1 + e^{\mathbf{w}^\mathsf{T}\mathbf{x}}} \end{aligned} \tag{1.15}$$

The logistic regression choose the class which gives higher probability in terms of probabilities summarised above. By checking the logit function of classification probability, we could get

$$\log \frac{P(y = 1|\mathbf{x}; \mathbf{w})}{1 - P(y = 1|\mathbf{x}; \mathbf{w})} = \mathbf{w}^\mathsf{T}\mathbf{x}, \tag{1.16}$$

which illustrate that logistic regression actually uses the predicted value of the linear regression model to approximate the log-odds of the true label of the classification task.

The learning process of logistic regression is to estimate the parameters $\mathbf{w}$. It is often solved by using *maximum likelihood estimation (MLE)*, i.e., finding a set of parameters such that the likelihood of dataset sampled from the fitted distribution is maximised. More precisely, assume $P(y = 1|\mathbf{x}; \mathbf{w}) = p(\mathbf{x})$ and $P(y = 1|\mathbf{x}; \mathbf{w}) = 1 - p(\mathbf{x})$, the likelihood function is:

$$L(\mathbf{w}|\mathbf{x}; y) = \prod_{i=1}^{N} [p(\mathbf{x_i})]^{y_i} [p(\mathbf{x_i})]^{1-y_i}. \tag{1.17}$$

To facilitate the computation we take the log form of likelihood. Thus, together with the notation of 1.15, we could derive

$$L(\mathbf{w}|\mathbf{x}; y) = \sum_{i=1}^{N} [y_i(\mathbf{w} * x_i) - \ln(1 + e^{\mathbf{w} * x_i})]. \tag{1.18}$$

In machine learning perspective, $L(\mathbf{w}|\mathbf{x}; y)$ in form 1.17 is equivalent to the *cross entropy loss function*: $J(\mathbf{w}) = -\frac{1}{N} \log L(\mathbf{w})$. A slight difference is that the loss function is usually appended with a regularisation term to prevent overfitting. Finally the likelihood function is optimised using *Newton's method* or *gradient descent* [Alpaydin (2020)].

1.4.2 Hidden Markov Model

Hidden Markov Model (HMM) [Alpaydin (2020)] is a statistical model proposed by Leonard E. Baum in a series of papers [Baum and Petrie (1966), Baum and Eagon (1967), Baum et al. (1970; 1972)]. It is extensively used to model sequential data, especially in pattern recognition [Varga and Moore (1990), Yamato et al. (1992), Schuller et al. (2003), Starner and Pentland (1997)], and in bioinformatics [Krogh et al. (2001), Ernst and Kellis (2012)].

The system being modelled is assumed to follow a Markov process and its generated sequence is the *state sequence* of the system, denoted as X. However this sequence is unable to be observed directly or precisely, which is common in real world systems. Instead, in HMM there is an observation sequence whose outcomes of each step are determined solely by corresponding step in X, denoted as Y as shown in Fig. 1.1. More formally, definition of HMM with discrete time step is given: For two sequence $X = (x_1, x_2, \cdots, x_T)$ and $Y = (y_1, y_2, \cdots, y_T)$ be discrete-time stochastic processes instance:

- X is generated by a Markov process and observation Y doesn't affect state X: $P(x_t|x_{t-1}, y_{t-1}, \cdots, x_1, y_1) = P(x_t|x_{t-1})$ for $t = 1, 2, \cdots, T$, with its behaviours unable to be directly observed, also known as *hidden*.
- Observation at any time depends and only depends on the state at same time: $P(y_t|x_T, y_T, \cdots, x_1, y_1) = P(y_t|x_t)$.

In order to construct a HMM, several other components should be introduced:

- $Q = \{q_1, q_2, \cdots, q_N\}$: *State set* which include all possible states.
- $V = \{v_1, v_2, \cdots, v_M\}$: *Observation set* which include all possible observations.
- $A = [a_{ij}]_{N\times N}$: *State transition probabilities* of Markov process, $a_{ij} = P(x_t = q_j|x_{t-1} = q_i)$.
- $B = [b_{ij}]_{N\times M}$: *Emission probabilities*, which tells the probability of a state q_i to be observed as v_j, $b_{ij} = P(y_t = v_i|x_t = q_i)$.
- $\pi = (\pi_i)_{N\times 1}$ *Start probabilities*, which represents the probability of initial state of system is q_i, $\pi_i = P(x_1 = p_i)$.

There are three basic type of problems with system being modelled as HMM:

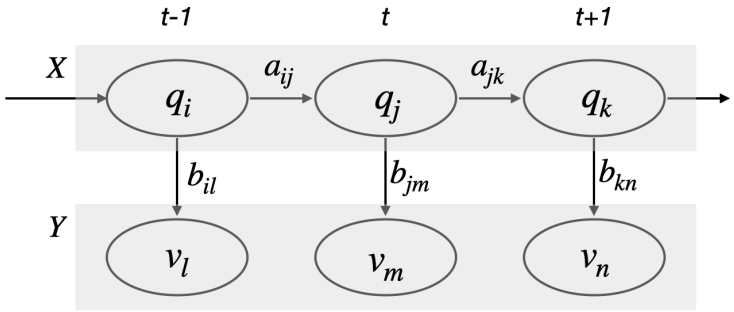

Figure 1.1: An instance of HMM. $t-1, t, t+1, \cdots$ are time steps, upper sequence X is the state sequence which is unobservable and lower sequence Y is the observe sequence, each ellipse present a state (q) or an observation (v). a_{ij} is the transition probability from q_i to q_j, b_{ij} is the emission probability from q_i to v_j, as defined above.

- ***Probability of an observed sequence***. Given model parameters $\lambda = (A, B, \pi)$, this task is to compute the probability of the observed sequence O by adding up over all possible state sequences.
- ***Learning***. Given the the observed sequence X, the goal is to estimate the best set of model parameters $\lambda = (A, B, \pi)$ that maximise the probability of observed sequence X: $P(X|\lambda)$. The task is usually to be solved using the maximum likelihood estimate. The *Baum–Welch* algorithm which is a special case of the expectation-maximisation algorithm. For time-series prediction problems, more sophisticated Bayesian inference methods, such as *Markov chain Monte Carlo (MCMC)* sampling are proven to be favourable over finding a single maximum likelihood model both in terms of accuracy and stability.

  There is also a supervised learning method when both observation and state sequence are given. With this materials we could estimate parameters by MLE, for example to learn state transition probabilities $A = [a_{ij}]_{N\times N}$, A_{ij} is the frequency that state q_i at time t is transited to state q_j at time $t+1$, then $\hat{a_{ij}} = \frac{A_{ij}}{\sum_{k=1}^{N} A_{ik}}$.
- ***Inference problems***. Several inference problems are associated with HMM, one of the most used is the decoding of the observed sequence. Given $\lambda = (A, B, \pi)$ and observed sequence Y, the task is to find the state sequence X that maximise the probability $P(X|Y)$. Speech recognising, for example, where the words of speech is states sequence and corresponding to an observed sequence of voice. The goal would be from voice sequence to decode the words in it. To solve this type of problem, a dynamic programming algorithm called *Viterbi algorithm*

[Viterbi (1967)] could decode the most likely state sequence by obtaining the maximum a posteriori probability (MAP) estimate.

1.4.3 Artificial neural network

Traditional machine learning methods are based on the routine of rigorously defining the problems, followed by designing a feature extractor with engineering and domain expertise, to finally propose an algorithm with calculus, probability and statistical methods to optimise and infer parameters in order to achieving the goal of learning from data. Different from it, *artificial neural networks (ANNs)* take another approach of building a model with structure inspired by the biological neural networks. Two who took first step toward artificial neural networks are *Warren McCulloch* and *Walter Pitts*, whose work in 1943 modelled a simple neural network with electrical circuits [McCulloch and Pitts (1943)]. Different from tradition machine learning methods, ANNs could learn the representations of the data directly from the data itself.

The basic components of ANN are nodes (artificial neurons) and connections. Nodes consist layers - an input layer, one or several hidden layers, and an output layer - and different layers are stacked, as shown in Fig. 1.2. The nodes are linked by connections with associated weights and thresholds, to allow signals (a real number) transmit over the network. An artificial neuron receives one or several signals from other neurons as inputs whose summation is feed to a nonlinear *activation functions*, then the result is propagated as input to its connected neurons, normally with certain direction. When layers are stacked deep enough (known as deep learning), a neural network can composite a very complex functions to approximate the real mapping between inputs and output of sample data. The learning of ANN aims at adjusting the weights and the thresholds of the network to improve the accuracy of the results which is achieved by minimising the observed errors.

Integrating information of the past (backwards) could obviously enhance our ability to learn from data, yet for certain task, the information of future (forward) could also be profitable. For example in handwriting recognition task, when recognising a word, not only the information of previous words, but also the information of following words helps. *Bidirectional recurrent neural networks (BRNN)* proposed in [Schuster and Paliwal (1997)] is the right medicine for this concern. The BRNN group two independent RNNs together as demonstrated in Fig. 1.3. The input sequence is fed to one RNN in normal time order, and to another in a reversed

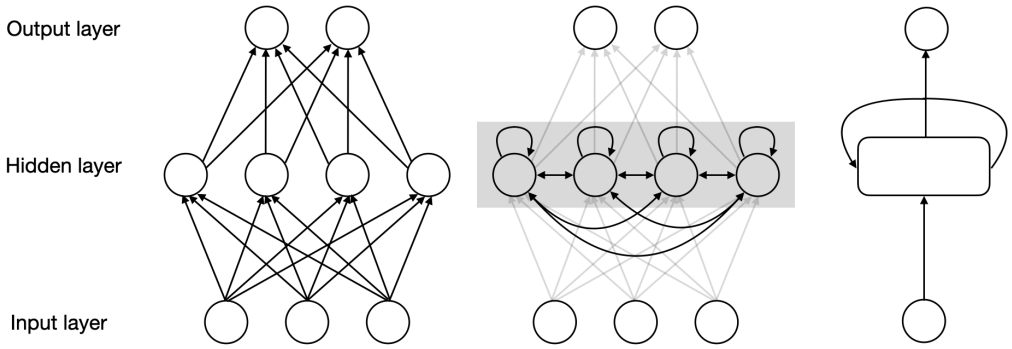

Figure 1.2: Structure of basic ANN (left) and RNN (middle), with recurrent structure of hidden layer highlighted in grey shade. Simplified representation of RNN is on the right.

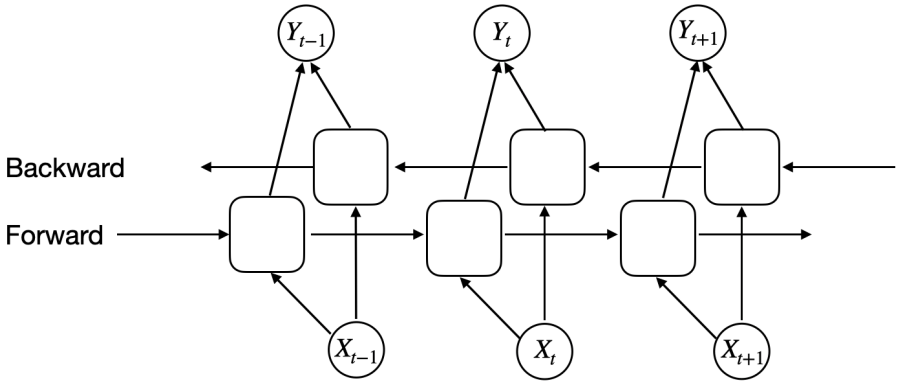

Figure 1.3: BRNN unrolled by temporal logic.

order. The outputs of the two RNNs are combined (usually by concatenation) and then pass to activation function.

There are many such cases where the information of a sequence determines the event itself, for example when we encountered ambiguous phrase while reading, information from its context is beneficial to eliminate the ambiguity. Tradition neural network doesn't provide the architecture to link order dependent data. To tackle this, we focus on a special ANN structure called *recurrent neural network (RNN)*, proposed in [Rumelhart et al. (1986)], that is capable of learning order dependent data. In the traditional artificial neural networks, neurons are connected from the input layer to the hidden layer to the output layer, with no connections within each layer. In the structure of RNN, however, the nodes of the hidden layers are connected so that the input of the hidden layers includes not only the output from the input layer but also the output of the hidden layer at previous moment, which makes RNN capable of keeping the information of previous time step

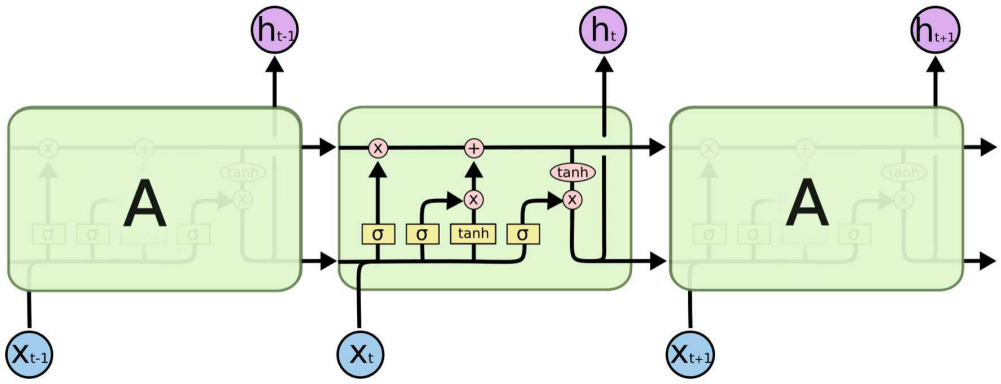

Figure 1.4: Structure of LSTM. The horizontal line in upper part of cell is the cell state. [1]

and join it to the computation of the current time step. This feature makes RNN applicable to tasks with sequential data such as speech recognition, emotion classification or image description generation [Graves et al. (2013), Tang et al. (2015), Ebrahimi Kahou et al. (2015), Karpathy and Fei-Fei (2015)].

In the most basic RNN architecture - fully recurrent neural networks - the hidden layer connects the outputs of all neurons to the inputs of all neurons. Considering the basic RNN with $\tanh$ as activation function, we have

$$\begin{aligned} h_t &= \tanh(W_{hh}h_{t-1} + W_{xh}x_t) \\ y_t &= W_{yh}h_t \end{aligned} \tag{1.19}$$

where t is the time step, x_t, h_t are the input and output of t. W_{hh}, W_{xh} is the weight of hidden layer and input layer respectively. Since RNN incorporates information of the past, a different training method called *Backpropagation through time (BPTT)* is designed for RNN. In vanilla RNNs training, BPTT cannot solve the long-time dependence problem because of the gradient vanishing and exploding. To handle this, a modified RNN architecture called *Long short-term memory (LSTM)* is proposed in [Hochreiter and Schmidhuber (1997)]. Compared with traditional RNN, its core module includes *Memory cells* and *gate units*. Recall that the recurrent hidden layer in traditional RNN holds memories, the gates in LSTM are introduced to "protect the stored memory contents from perturbation by irrelevant inputs" [Hochreiter and Schmidhuber (1997)]. All gates resulting a more complex unit called *memory cell* which holds the memory of past information.

[1] Image from https://colah.github.io/posts/2015-08-Understanding-LSTMs.

1.5 Thesis Objective

Relying on the advancements in complex networks, computational human dynamics and machine learning I introduced, my thesis thesis aims to contribute into the directions of five main objectives:

- *Unique data collection*: The project related to this thesis have collected 3 years of temporally fine-grained longitudinal interaction data for the critical periods of children's socialisation and language development. Along with it we have collected sociodemographic data to assign children with rich attributes for future analysis. Linguistic data has been collected once a year to monitor the linguistic performance. We also collected several ground truth data to facilitate the training of machine learning model which will be established later in the thesis.

- *Data cleaning and activity re-construction*: We have designed a complex scheme of re-constructing the collected interaction data based on the type of error occurred such as equipment failure, accidental drop etc., which has never been achieved in such kind of experiment. Furthermore, due to the different social behaviour under different environments, we segment interaction into two periods and removing signals during transition to make sure the validity of signals.

- *Temporal network reconstruction:* There has always been a lack of comprehensive methods to reconstruct temporal networks from raw RFID. Thus my subsequent aim was to established a pipeline and to investigate several machine learning methods to reconstruct temporal interactions from wireless signal data. With training on our ground truth data, the reconstruction accuracy increased from 77.28% with traditional method to 90.03% of our pipeline with Bidirectional Long Short Term Memory (BLSTM).

- *Multivariate analysis of homophily:* With the reconstructed temporal network, accompanied with sociodemographic data, we search for attributes which exhibit effects of individual and group level homophily. We investigated three homophilic indices and compared empirical aggregated network with fully randomised network. Discovered that gender is the most homogeneous attribute for each age group and its effect increase by age. We have also found

evidence that two linguistic features induce homophilic tie creation effects that increase with age.

- *Visualisation tools of genealogical networks*: Finally I describe a visualisation tool that we developed to visualise multidimensional and dynamical genealogical trees. The tool is demonstrated on a large scale family genealogical history data from China.

As it follows I summarise my achievements in four main Chapters. First I concentrate on the description of the design of the large-scale social experiment, and the data collection pipelines. In the next chapters I introduce my methods for the temporal network and activity reconstruction. Using the reconstructed datasets, I present my results on the homophily analysis. Then I shortly summarise our genealogy tree visualisation tool with case study and finally I close my thesis with a short discussion.

Chapter 2

DyLNet Project and Data Collection

This chapter describes my work, which has been summarised in the paper [Dai et al. (2022)].

2.1 Background and Motivation

The structure of social networks and their dynamics over time strongly influence language usage and change [Labov (2001)]. Conversely, the way in which individuals use language contributes to the way they are judged in society [Giles and Billings (2004)], and therefore influences their friendship choices, modifying the structure of their social network. Despite this recognised relation, the co-evolution of dynamically changing social networks and language dynamics mitigated by social interactions is a largely unobserved phenomenon. Preschool environment provides an ideal place to observe these reciprocal influences. Children's language changes rapidly during the preschool years due to the acquisition process [Conti-Ramsden and Durkin (2012)]. Meanwhile, children integrate and adapt at school via socialisation and increased opportunities to communicate with peers and with the adults in charge. Besides, contacts with many peers cause preschoolers to expand and restructure their social network [Lynn Martin et al. (2005), Schaefer et al. (2010)]. This co-evolution process, through the interactions between language acquisition and socialisation, has societal implications as it may promote or undermine academic success and linguistic skills. A virtuous circle – or a spiral of failure – between children's sociability, oral communication and learning at school may therefore ensue.

Social inequalities are a key factor in this causal chain since, as of age 2, it has been observed that children from families of higher socio-economic status (SES)

have a richer lexicon and use more complex syntax than children from lower-SES environments [Le Normand et al. (2008)]. Children from different backgrounds do not use, to the same extent, the academic language that is encouraged at school [Aarts et al. (2011), Snow (2010)]. The observations of these early differences, which are transmitted within the family [Huttenlocher et al. (2007)], led to numerous studies that have revealed the key-influence of the nature and quantity [Rowe (2012)] of speech addressed to children by their parents in the different social environments. School attendance introduces a new factor into the equation through peers influence, especially when the academic group is socially mixed [Schechter and Bye (2007)], or through the speech produced by the teachers [Bowers and Vasilyeva (2011)]. The linguistic skills of a child will advance more quickly if they are a member of a peer group with a high level of language abilities. This effect has been observed across various indicators of language development: vocabulary, syntax, or narrative skills [Henry and Rickman (2007), Justice et al. (2011), Mashburn et al. (2009), Schechter and Bye (2007)].

The aim of the DyLNet project [dyl] is to observe and characterise the relations between child socialisation and oral language learning during the preschool period by means of an innovative multidisciplinary approach that combines work in the fields of language acquisition, sociolinguistics and network science. This goal has been achieved via a large-scale longitudinal social experiment, where a complete preschool in France was followed, including children from three different grades as well as their teachers and assistants. During the experiment we collected the proximity interactions of about 200 participants (circa 170 preschoolers and 30 adults in charge) in every 5 seconds using autonomous Radio Frequency Identification (RFID) Wireless Proximity Sensors, which were (for a large part) equipped with directional microphones allowing to record continuously the oral interactions of participants too. During the observation period of three years, data collection was conducted monthly with each deployment lasting one week. In parallel, survey campaigns using conventional techniques were carried out to record, on the one hand, the social, economic, cultural, educational, and language background of children's families and, on the other hand, the individual level of language development of the children. All together, these simultaneous data collection efforts culminated in a long, large, and comprehensive children language development study conducted in the context of peer interaction.

This data allow for several types of analysis potentially interesting not only for sociolinguists and researchers in child language development, but also for ex-

perts in social networks, human dynamics, behavioural science, education science or even social anthropology. It may open a cross-disciplinary approach to learn about how children language development depends on age, gender or socioeconomic background and how it co-evolves with the social interaction dynamics on the short and long temporal scales at the level of individuals and groups in a supervised (classroom) or unsupervised (playground) school setting.

2.2 Data Collection Methods

Novel digital technologies enable to follow human social interactions with an unprecedented resolution in time and space [Lazer et al. (2009b)]. Over the last decade, these advancements led to an avalanche of experimental and data-driven studies addressing the precise observations of human interactions [Goffman (2017)] to explain phenomena like social tie formation or group dynamics. One exciting direction involves wearable devices as they allow for tracking dynamically human actions or proximity interactions at the individual level for large populations in various settings. Social studies using wearable technologies have been deployed in multiple settings such as schools [Stehlé et al. (2011b), Fournet and Barrat (2014)], conferences [Isella et al. (2011)] and hospitals [Martinet et al. (2018), Duval et al. (2018)]. These studies commonly relied on some already existing wireless architectures using RFID technology adjusted for the purpose of recording face-to-face interactions between people. Standards as *OpenBeacon* [ope] or *Open badges* [Lederman et al. (2017)] were deployed as centralised communication protocols for data collection about the relative distance and orientation of RFID tags distributed among people moving around in the same space. Our data collection method relies on similar but decentralised technology to record face-to-face interactions of children with the original goal to understand how their social interactions co-evolve with their language acquisition dynamics.

In this section first I introduce the ethics implication of our data collection, then I will present different types of data we collected and how we manage to identify problems followed by the pre-processing techniques. Finally we present the result of our data collection.

2.2.1 Ethics and data protection

The DyLNet project data collection was carried out during three successive years between 2016 and 2020 to gather transactional, vocal, language and socio-demographic

data about children and staff in a preschool in France. Acceptance of the experiment by pupils, parents, educational staff and school authorities, as well as issues of benefit-risk balance and privacy were carefully considered before and during the project.

The choice of the school was made with the involvement of the local and regional education authorities prior to the start of the project. The goal was to find a school with pupils from a variety of cultural, linguistic and socio-economic backgrounds and to obtain official permission for the experiment in the form of an agreement signed by the regional education officer and the university. Before launching data collection, meetings were organised with the parents and school staff to explain the purpose and organisation of the experiment as well as the functioning of the RFID devices. We also offered parents the opportunity to meet individually with the researchers during drop-in sessions. In addition, we created a webpage dedicated to the families accessible from the website of the project. This webpage provided them with every details about the implementation of the project: goals, methods, benefits, risks, and the schedule of data collection [Nardy (2016)].

The issue of children's exposure to radio signals emitted by RFID devices has been carefully considered. We followed the advice of an expert on radio frequency safety who was a member of the French National Council for Public Health. She recommended that we have a Specific Absorption Rate measurement (i.e. rate of energy absorption per unit of mass by a human body) of the RFID device performed by an authorised company. The value obtained for one device was 0.0001683 W/kg under normal conditions of use in contact with the body, which is much lower than the European standard (2 W/kg). In addition, in agreement with the company that designed and manufactured the devices, we made sure that they complied with the European standards on the mechanical and physical properties of objects to be used by young children.

To protect the privacy of participants, we applied the principles of not mentioning participants' names, precise locations and dates in the stored data ahead of analysis and dissemination of the results. The exact dates and timestamps at which data were collected and the child participants' age were notably coded relatively to an arbitrarily defined *T0* set for the research project database. All participants were assigned anonymous numerical identifiers with association keys available only to the principal investigator. Start and end dates of data collection periods were ap-

proximated, and birth dates were replaced with children's ages. Re-identification of participants thus appears to be impossible from the shared datasets.

As a prerequisite to be included in the study, parents (on behalf of their child) and school staff were asked to give a written consent for their participation, while informed that they could nevertheless opt-out from the experiment at any time. Non-participating children were offered to wear empty shells of RFID badges to minimise feelings of envy among classmates. Participants were asked to provide a second written consent to share the collected and anonymised data with the scientific community under the control of the Principal Investigator. The acceptance rate of participation in our experiment was 80.63% (283/351) for pupils and 96.88% (62/64) for school staff over the three years of the project.

The whole project including experimental design, subject recruitment, data collection and processing, data handling, storing and sharing, privacy protection, and all aspects of the involvement of underage children were screened and approved by the ethics committee of INRIA (National Institute for Research in Digital Science and Technology) (favourable opinion, reference 2017-014, IRB00013144) as well as by the Data Protection Officer of the Université Grenoble Alpes (favourable opinion, reference CIL-UGA-2017-0980683).

2.2.2 Data collection

We collected four different types of datasets during the DyLNet project Dai et al. (2022). The main dataset focuses on the dynamical recording of social and oral interactions as transactional and vocal data. These data were collected autonomously using badges installed on children and school staff at the preschool (for details see Section 2.2.2.1). Additionally, we gathered information about the school level of each child, that is being in 1st grade (about three years of age), 2nd grade (about four years of age), or 3rd grade (about five years of age). We also recorded the class in which the pupils and school staff were participating (out of the 7 classes in the preschool). Meanwhile, ground truth (GT) data were collected with the purpose of understanding how distance and relative orientation between a pair of badges influence the Received Signal Strength Indicator (RSSI) of recorded signals. GT data were also essential for training Machine Learning models to classify signal sequences as social interactions (for details see Section 2.2.2.2). In addition, the main data collection was accompanied with survey campaigns. A first type of survey consisted in asking parents to provide information about the socio-demographic, cultural, educational, and occupational background of the family and the daily

out-of-school activities of participating children (for details see Section 2.2.2.3). In addition, a language survey using vocabulary and syntactic skills assessment methods was performed with all participating children once a year throughout the project to follow their linguistic development (for details see Section 2.2.2.4). During the observation period, we followed the interactions among preschoolers and school staff for one week each month (among the 10 months of the academic year). More precisely, we recorded data during five morning and four afternoon sessions each week, as traditionally schools are closed in France on Wednesday afternoons. In practice, children and staff were equipped with a wearable RFID badge every morning of deployment under the supervision of a researcher on site. Badges were then collected before lunch break and re-distributed in the early afternoon (or after nap time for the youngest preschoolers). Badges were collected again in the evenings for charging overnight. Data were extracted from the flash memory card of the badges at the end of each week of deployment. As the RFID badges were autonomous, they were worn not only in the classroom but also during play time when children were moving freely in the open-air yard of the school.

2.2.2.1 Transactional data collection

We employed a decentralised Low Power Wireless technology to collect transactional data between autonomous RFID badges. Each badge could be in two modes, whether broadcasting their own radio frequency ID or listening to signals emitted by other badges. To be more precise, each badge was associated with a unique ID and used the IEEE 802.15.4 low-rate wireless standard to communicate[1]. Since badges in our experiment worked in a decentralised mode, in order to make sure that they shared consistent global time, they were first synchronised with a *synchroniser*, which was connected to a computer and propagated the same time reference to all badges. During data collection, badges broadcast a 'hello' packet with 0 dBm transmission power for 384 μs every 5 seconds. For communication, they used the carrier-sense multiple access (CSMA) protocol. To avoid collision, they first listened to the dedicated channel, then transmitted a packet if the channel was clear. A badge listened to incoming packets from other devices when it was not transmitting a packet. The decentralised architecture means there was no central node to

[1]The employed technology and its implementation meet the requirements of the product standard EN50566 following the basic restrictions of the European Council recommendations 1999/519/EC.

record all traffic. Instead, all badges worked autonomously and recorded incoming packets described by the sender badge ID, the timestamp of reception, and the RSSI. Each received signal was stored in a file locally on the flash memory card of each badge if its RSSI value overreached the minimum sensitivity value of -94 dBm. To facilitate charging of the badges (every night of deployment) and to transfer data to a local computer (at the end of every week of data collection), multi-USB hubs were designed where up to 35 badges could be plugged at the same time.

Our architecture employed three types of badges serving different purposes:

- PROX: These badges were given to participants and were hanged on their chest during data collection days. They were equipped with a battery, a memory card, and about half of them with two directional microphones for voice recordings. As soon as they were unplugged from the charging hub, these PROX badges emitted a radio signal every 5 seconds, as well as listened and recorded incoming signals on their own flash memory card. They stopped collecting data once plugged back on the charging hub.

- RX: These were special sensors, which were installed on the charging hub of each class and left unaltered during the whole week of data collection. Unlike PROX badges, RX badges only listened and recorded incoming signals (even when plugged on the charging hub). Their role was to observe when the participants belonging to a given class (attached to a given RX) were inside their classroom or not.

- FOX: This device was used for the time synchronisation of all badges at the beginning of every week of data collection. This special *synchroniser* device was first plugged onto a PC to catch global time, then moved around the classrooms to propagate time information among the PROX and RX badges.

At the end of each data collection period, data from each badge were transferred to a local computer. Badge IDs were then associated to the corresponding badge bearer (participant) ID within each file of contact data, and finally data were passed through an initial cleaning pipeline. The obtained cleaned raw data served as the input for the data pre-processing and temporal network reconstruction pipeline. All the steps of these pipelines are explained in the Section 2.2.3.

2.2.2.2 Ground truth data collection

In parallel to the autonomous transactional data collection, we occasionally recorded ground truth data about actual social interactions between participants using direct visual observation methods. For two datasets (GT1 and GT2), the researcher focused on a given pair of children for a given period of observation, whereas the third (GT3) was recorded among a complete class group. Besides, while GT1 and GT3 were collected *in-situ* within a classroom with all noise and interference present (i.e. among $20 - 28$ participants wearing a badge), GT2 was recorded in a separate room, away from other badge bearers, in controlled settings.

More specifically, GT1 consists of the recordings of the state of interaction/no-interaction between a given pair of children at a fixed 10 seconds interval (*scan sampling* method [Altmann (1974)]), as well as their relative body orientation. GT3 corresponds to the logs at a fixed 2 minutes interval (*scan sampling* method [Altmann (1974)]) of distances between all the children and adults of one class group during regular activities within their classroom. Both GT1 and GT3 were collected using the Animal Observer application for iPad [Ani]. These observations make possible a direct comparison between the RSSI values collected by the badges and the actual interaction state (GT1) or distance (GT3) between classmates. Finally, to get an even clearer idea of the relation between RSSI values and the actual distance between badge bearers, as well as their relative orientation, in a noise-free environment, GT2 dataset was recorded among pairs of children statically positioned at a given distance (0.1 meters, 1 meter, 2 meters) and orientation (face-to-face, side-by-side, back-to-back) for ~ 10 minutes periods. For a more detailed description on how ground truth data were collected see Appendix A.1.

2.2.2.3 Socio-demographic survey data collection

Each family that consented to let their child participate in the study was asked to fill in a paper questionnaire. This questionnaire aimed at collecting information about the child and their daily family environment. In its first part, we recorded basic socio-demographic information about the participating child: gender, date of birth, birth rank and number of siblings. Other questions aimed at gathering information about the places of socialisation frequented by the child before entering school (nursery, childminder...) and within the school (daycare, canteen). We also asked the parents about their child's level of sociability, talkativeness, favourite out-of-school activities (e.g. sports, drawing, imagination/construction games...)

and customary activities before sleep (e.g. story telling, cartoons, music...). Finally, two questions aimed to identify the child's language environment and language practices at home (i.e. whether French and/or other languages are spoken within the family, and whether or not they are understood/spoken by the child).

In the second part of the questionnaire, we collected information about the child's family environment and living place(s) (i.e. the composition of the household, and, in the case of separated families, the child's habitual place(s) of residence). We also recorded the parents' geographical origin, their employment status, their area of professional activity and their level of education. That questionnaire was fully completed when each child participant entered the study, then parents were asked every year to fill in a shorter version of it in order to update some information likely to change over time (e.g. the child's favourite activities, the parents' professional status). The variable sample is presented in the Section 4 later.

2.2.2.4 Linguistic survey data collection

Individual tests were administered to participants at school in order to assess the children's level of language development at several points throughout the longitudinal 3-years follow-up. Tests were performed at the beginning of every school year for participants entering the study, then at the end of every school year for children already enrolled in the project. Children were evaluated individually, in a separate room, by a member of the research team. Test sessions were designed to last no more than 15 minutes, hence each child took part in two short sessions a few days apart: one during which their receptive lexical skills and short-term memory span were evaluated, and the other for assessing their receptive syntactic skills.

Language tests aimed at evaluating the participating children's level of comprehension of words and utterances. Four versions of the tests were designed to be appropriate for each stage of preschool education, that is for pupils 1) entering 1st grade, 2) completing 1st or entering 2nd grade, 3) completing 2nd or entering 3rd grade, and 4) completing 3rd grade. The evaluation of lexical skills included 40 items (words), 30 of which being 'test items' specific to each version (i.e. items chosen to be adapted to the children's school level) and the remaining 10 being 'anchor items' shared across all four versions (i.e. items systematically presented to the children whichever their grade, and chosen to be rather adapted to 3rd grade pupils). Similarly, the evaluation of syntactic skills contained 20 items (utterances), namely 10 'test items' and 10 'anchor items'. In both cases, for each item, the child

was presented with a plate of four pictures. The experimenter then produced the item (a word or an utterance), and the child had to point to the corresponding image with their finger.

Besides, as memory span is known to be closely linked to language skills [Blake et al. (1994)], we gave each child a memory span test as a control measure. More precisely, we asked the child to repeat after the experimenter a series of digits of increasing length. Each level contained two trials (i.e. two different series of digits of the same length). We started with a series of two digits, then proceeded with increasingly longer series (of three, four, etc.), and stopped the evaluation when the child had failed the two consecutive trials for a given level (i.e. two consecutive series of a given length).

2.2.3 Transactional data pre-processing

2.2.3.1 Initial data cleaning

The data directly recorded by the badges appeared with several trivial corruptions, which were corrected during a pre-cleaning pipeline. In this pipeline, we first converted the recorded binary data files to human readable format. Then we systematically removed signals that were recorded during out-of-school time, and pushed it through an individualised cleaning process. In fact, a researcher on site during each deployment week continuously followed which badge was attributed to which participant, whether a defective badge had to be replaced, whether it was dropped by a child, or if a participant was absent over a given period of time. Also, in such recorded situations, all signals emitted or received by unused, broken or missing badges were removed, not only from their own data sequence but also from the data sequence of other badges. At the end of this initial cleaning pipeline, we applied a file merging procedure to ease further analysis of the data. According to the schedule of the observed preschool, each day was divided into two periods: a morning session from 8:30 to 11:20, and an afternoon session between 13:45 and 15:50. To follow the same schedule, we merged the raw files of each participant into two files each day. This merging procedure allowed to retain only one file per participant per half-day, notably in cases where a participant wore different badges during that period (when a defective badge had to be replaced) or when file fragmentation had been caused by system errors. Consequently, each individual appeared with nine half-day files for each week of data collection (i.e.

every mornings and afternoons from Monday to Friday, except for Wednesday afternoon) in the cleaned raw transactional dataset, from now on called the *raw data*.

2.2.3.2 Data pre-processing

Taking the raw data as input, we developed a data pre-processing pipeline that we present below and whose corresponding source code we share [Nardy et al. (2022)]. This pipeline contains some methodological choices that we made for the purpose of our very own study. However, the shared raw data allow other researchers to develop their own data pre-processing pipeline with their own constraints and needs, potentially requiring other methods and parameterisation than those we used. During data pre-analysis, we identified multiple issues that we corrected during this data pre-processing stage.

Issue 1 In the raw data, we identified some rare corruptions induced by a few badges: occasionally, some badges ('silent badges') only received signals without emitting any data packet or, on the contrary, some other badges ('deaf badges') only sent out signals without receiving any. To solve this issue, we copied the signal sequence that silent badges received and reversed the direction of senders and receiver, then added these reconstructed signals to the corresponding half-day files of the senders. Meanwhile, all signals received by badges that detected the deaf badge were gathered to reconstruct its incoming signal.

Issue 2 There were other situations where the recorded signals were meaningless, that is when badges were turned on but not worn by the participants. This happened notably when data collection was on hold and badges were retrieved and gathered before being plugged back on the charging hub. To deal with this issue, we exploited the signals received by the special RX badges which were located on the charging hub of each classroom, and thus received strong and stable RSSI from any unused badge located in the vicinity of the hub. To detect such situations, we used sliding time windows over the signal sequence of each RX badge (assigned to a given class) and computed the average and standard deviation (STD) of RSSI values for each PROX badge used in that classroom within each window. We used a relatively long time window of 3 minutes with a 1 minute step to avoid spotting situations where equipped children only approached the hub for short periods. Also, we considered only time windows with at least 80% of the expected number of signals present (i.e. at least 29 signals for a time window of 3

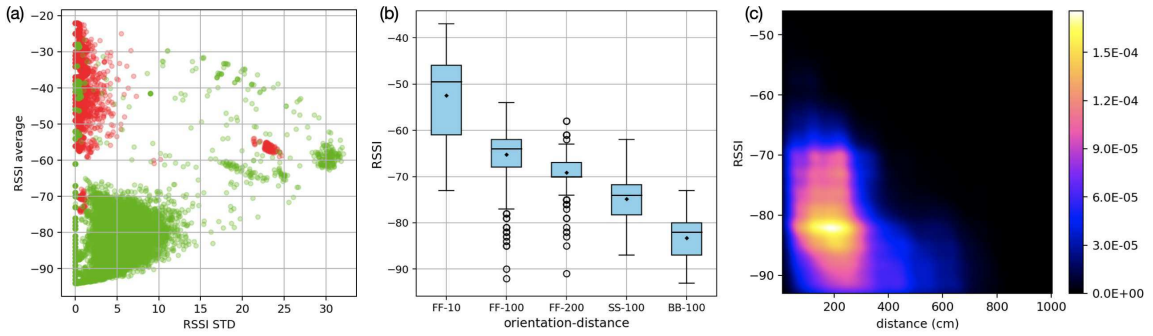

Figure 2.1: Signal strength statistics during interactions and no-interactions. (a) Standard deviation (x-axis) and average (y-axis) of RSSI values recorded by RX devices and coming from badges either worn (green dots) or not worn (red dots) by a child participant. Annotation is based on ground truth collected at the school (i.e. known absences). Within this uncleaned set of data, cases of misclassification are likely due to wrong manipulations. Red points in the green zone would be badges of absent children approached or manipulated by other classmates, but such cases will be solved during the Initial Data Cleaning procedure based on recorded absences. However, green points in the red zone would be badges unused or dropped by present children (notably at the beginning and end of a half-day of data collection), and these are such cases that will be filtered by the described cleaning procedure in Issue 2. Settings: window size = 3 minutes, step = 1 minute. Plot based on observational data from 7 classes, i.e. 163 children and 7 RX badges, during 1 week of data collection. (b) Distributions of RSSI values shown as box-plots for pairs of children observed at different experimentally-fixed distances and relative orientations (from GT2 dataset), and used to parameterise the cleaning procedure described in Issue 3. Black diamonds indicate the average value and bar is the median value. Position-distance (x-axis): letters indicate relative orientation as *'FF'* face to face, *'SS'* side by side, *'BB'* back to back; and the following number indicates the distance in centimetres. Plot based on observational data from: FF-10: 1 pair, 120 seconds, 48 data points ; FF-100: 5 pairs, 2765 seconds, 1106 data points ; FF-200: 2 pairs, 720 seconds, 288 data points ; SS-100: 1 pair, 240 seconds, 96 data points ; BB-100: 1 pair, 335 seconds, 134 data points. (c) Correlation between distance (x-axis) and RSSI values (y-axis) shown as a density plot (from GT3 dataset). Plot based on observational data from: 1 class with 28 participants, 8 observation sessions for a total of 302 minutes, 62965 data points.

minutes), as otherwise it meant that the RX had temporarily lost contact with the badge (e.g. when the participant had left the classroom to get to the yard).

To set an appropriate RSSI threshold for distinguishing between worn and unused badges, we analysed the average and STD (i.e. the strength and (un)stability) of signals collected by RX devices and coming from unplugged active badges during the known presence or absence of the corresponding children (see Fig. 2.1a). Based on this analysis, we decided to set the threshold for RSSI average at -62 dBm and for STD at 2.5. These two thresholds together allowed us to achieve 93% accuracy as compared with ground truth.

To further refine data cleaning, we added two auxiliary treatments:

1. Two spotted consecutive inactive periods were concatenated if they followed each other by less than 2 minutes, i.e. if *period 1* goes from t_{s1} to t_{e1}, *period 2* goes from t_{s2} to t_{e2}, and t_{s2}-$t_{e1} < 120$ seconds, then we considered as inactive the period going from t_{s1} to t_{e2}. This step was necessary in order to remove isolated residual signals induced by noise between periods spotted as inactive.

2. Anytime the script identified an inactive period between t_s and t_e, data were in fact deleted from $t_s - x$ to t_s and from t_e to $t_e + x$, with x being a safety margin set at 30 seconds. This helped solving cases where badges were in the midst of being transferred from the charging hub to the participants (or vice-versa), hence recording meaningless data that the sliding time window method failed to detect because of the high STD of the RSSI values received by the RX in such situations.

These two treatments together with the sliding window method effectively removed the relative long periods of inactivity in the cases of unworn badges before equipment of the participants or after retrieval, badges dropped on a table, or badges inadvertently lost in the school for instance.

Issue 3 It also happened that PROX badges could not be detected by the RX badge assigned to the corresponding class, either because the RX badge was not working properly or because a PROX badge had been dropped somewhere far from the charging hub, hence too far from the RX. To circumvent this issue and still manage to detect periods of inactivity for PROX badges, we applied a similar sliding time window method but, this time, on the signal sequence of each PROX badge. This allowed us to identify relatively long periods of inactivity when two or more PROX badges were laid together, e.g. after retrieving all the badges from a class group at the end of a day or during certain sport sessions when wearing a badge could be uncomfortable. Meanwhile, we tried to avoid deleting any signal

corresponding to a potential social interaction. An actual social interaction typically involves just a pair of mutually observing badges for a short period of time and with largely varying RSSI values. To prevent spotting (and therefore deleting) such valuable series of signals, more restrictive thresholds were applied, namely $-55dBm$ for average RSSI and 1.5 for STD. From the ground truth dataset collected in controlled settings (GT2), we could assess that such a high signal strength can be reached only when badges are 10 centimetres away from each other or closer, and that such a low STD is very unlikely in natural settings with badges worn by children who typically never stand still (see Fig. 2.1b). In natural settings, such high RSSI values are in fact almost never observed among participants during regular class activities (see Fig. 2.1c). Here again, we considered only time windows with at least 80% of the expected number of signals present, and we applied the two above-described auxiliary treatments (i.e. concatenation and safety margin) whenever an inactive period was spotted.

Issue 4 Once long inactive periods had been discarded from the signal sequence of each badge, we often observed some remaining unrealistically strong signals. Situations where badges had been gathered in close proximity during equipment or retrieval process (e.g. in a box, or in the hands of an adult) resulted in particularly strong, though unstable, signal exchanges that should not be considered as actual social contacts. Contrary to the periods of inactivity, these periods were short and with large RSSI variation. They were likely to be found at the edges (beginning and end) of each data collection period, and also occasionally within the day (e.g. when retrieving badges ahead of a sports session then equipping the children again), so sequences of signals were analysed both forwards and backwards. During this procedure we applied a 1 minute window with a 20 seconds offset, thereby avoiding to delete longer periods because of a single strong RSSI value.

In the forward analysis, we first spotted the first signal (with the earliest timestamp) to define the 'initial time' t_0. Within a 1 minute window starting from t_0, we scanned signals chronologically to detect signals with RSSI > -45 dBm. That threshold was chosen based on the observation that this level of signal intensity is rarely observed even when badges are 10 cm away from each other (see Fig. 2.1b), and actually never observed in natural settings during regular class activities (see Fig. 2.1c). Then:

– if none were found, this initial time t_0 was kept untouched (it meant that the badge was already in use)

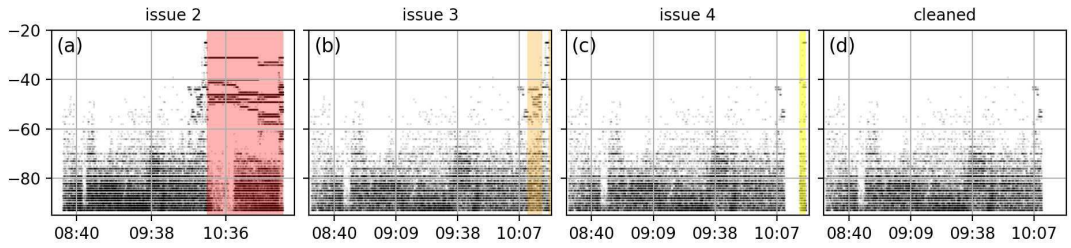

Figure 2.2: Illustration of the issues detected during pre-processing. On all these panels, each black dot represents a signal, and colour shaded areas show the problematic signals spotted by each procedure: (a) Issue 2 detected (red area) on a half-day file after initial data cleaning. The signals in the red area were subsequently removed; (b) Issue 3 detected (orange area) within the remaining signals from subplot a; (c) Issue 4 detected (yellow area) within the remaining signals from subplot b, and (d) remaining signals from subplot c constituting the final half-day file with cleaned signal sequence (i.e. pre-processed data).

- if a signal with RSSI > -45 dBm was found, then its timestamp was defined as 'strong signal time' t_s. The script continued to look for signals with RSSI > -45 dBm, and updated t_s every time a new strong signal was found within the 1 minute window.
- once we reached the end of this 1 minute window, we continued to update t_s only if a new strong signal was found not later than 20 seconds from the previous t_s.

We applied this procedure the same way for the backwards analysis, but the other way round: a final 1 minute window was defined at the end of each data clip, signals with RSSI > -45 dBm were searched to define t_s, and a step of 20 seconds was applied when the analysis continued backwards outside of the final 1 minute window.

These four issues were detected and treated in the order presented above. The effects of the data pre-processing pipeline are evident from Fig. 2.2, where issues are detected step-by-step on an example of a half-day signal sequence. After solving all the issues we listed above, we obtained pre-processed data sequences for 164 children and 32 adults over 50 days during 10 weeks in the 10 months of the academic year. We used these pre-processed data as input to reconstruct the temporal network of real social interactions and to identify class-time and free-playtime periods during each day for each class.

2.3 Summary

In this chapter first I introduced the DyLNet project with its motivation, aim and experiment settings. This was followed by the detailed explanation on the collection procedure and results of four types of data: transactional data, ground truth data, socio-demographic survey data and linguistic data, which provided materials of studies in following chapters. Furthermore, four types of identified anomalous were discussed in the transactional data after comparing it with data recorded in regular situation. For each anomalous behaviour we developed corresponding methods to remove meaningless and corrupted signals to minimise their impact. These work has been summarised in [Dai et al. (2022)], with sample data of one year and corresponding code available on request.

Chapter 3

Network and Activity Reconstruction

In this chapter I describe my work summarised in the papers [Dai et al. (2020; 2022)]. In this work we first explored several pipelines of event reconstruction using machine learning methods on sequential data. Then we simulated information spreading on the reconstructed networks obtained by different reconstruction methods. The results showed that network reconstructed in different ways may lead to significantly similar spreading dynamics, even if their network reconstruction accuracy was very different, which demonstrated the importance of precise network reconstruction and the careful choice of the reconstruction method. Furthermore, after we reconstruct the temporal networks, we designed a method to distinguish data collected in free- and class-time, which is essential due to their different interaction patterns. Finally in order to conduct homophily analysis, we developed algorithms to group class according to their mutual presence of free-time, and remove fragmentary groups with small duration.

3.1 Temporal Network Reconstruction

The precise observation of the dynamics of face-to-face interactions of people have been a major challenge in social studies [Goffman (2017)]. Such observations were commonly limited to small-scale observations for short periods of time [Duncan and Fiske (2015)]. Recently developed new technologies of wearable wireless devices made possible a giant leap in this direction, as they allowed for large-scale experiments to observe offline interactions in multiple settings. However, using the collected data streams, the reconstruction methods of temporal interactions were commonly based on naïve assumptions [Cattuto et al. (2010), Elmer et al. (2018)], which may seem convenient at first, but have indisputable consequences on the reconstructed event structure and any observed process taking place on it.

To bridge this shortcoming, by focusing on the proximity data collected by *DyLNet*, and relying on ground truth data recorded simultaneously in controlled settings, we explore several supervised reconstruction methods of the temporal social interactions. As demonstrated in Fig. 3.1, first we build a binary interaction sequence from the raw data of packet exchange between the LPW badges of a pair of individuals, and then use it to reconstruct the time and duration of the mutual interactions among the participants in order to obtain a temporal network representation of the social interaction dynamics. As we explain in Section 3.1.2.1, we translate the first level problem to a regression task, while in Section 3.1.2.2 we explore multiple naïve and advanced statistical learning methods to solve the final reconstruction problem of the dynamical interaction sequences. Further, in Section 3.1.3, via data-driven simulation of spreading processes, we demonstrate that while commonly used naïve reconstruction methods consistently overestimate the number of interactions, using advanced statistical learning methods, even a minor improvement in the reconstruction performance can have radical effects on the dynamics of an ongoing process.

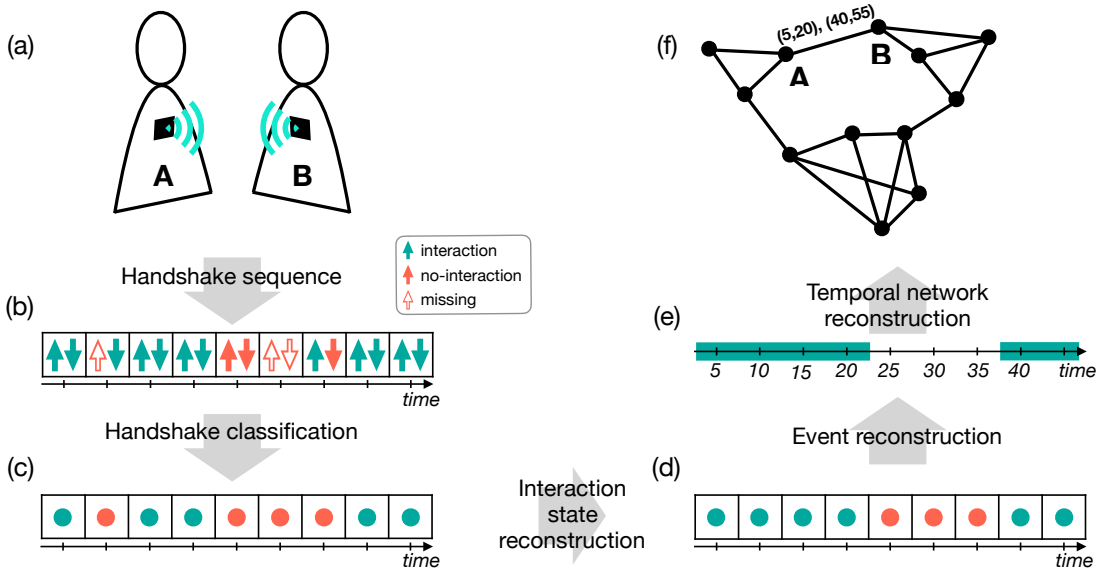

Figure 3.1: Temporal network reconstruction pipeline. Starting (a) from mutually observed packets of pairs of LPWD tags, (b) we train models using the raw and annotated data to (c) reconstruct interaction and non-interaction periods between individuals to (d) reconstruct events and ultimately (e,f) build a temporal network.

3.1.1 Environmental dependencies and parameters

During the experiments, each badge recorded a time-stamped sequence of packets, which were broadcast by other badges in its vicinity. More precisely, a sequence recorded by a given badge consists of (*t, ID, RSSI*) tuples, where t is the time of observation, *ID* is the unique identifier of the observed badge, and $RSSI$ is the received signal strength indicator of the observed packet transmitted as a radio signal. In Fig. 3.2a we show the distribution of RSSI values recorded over a week (24 hours, 196 badges). Values were observed between -24 and -94 dBm with a bimodal distribution. One peak, above -45 dBm, corresponds to the situation when badges are stored in a box close to each other thus communicating with strong radio signals. The other peak, below -75 dBm, corresponds to any other observations, including ones capturing real social interactions with an -94 dBm ceiling value hardwired in the badges configuration. Observed RSSI values can depend on external factors like distance, body orientation, battery status, or humidity conditions. While battery status should not be an issue here as badges are charged overnight, and we can control for distance and orientation (as explained below), we cannot account for changing weather conditions, which can cause some fluctuations in our measurements. In addition, the potential conflict of signals within $1\mu s$ may induce accidental loss of observed packet in case of interactions within large social groups.

Social as well as verbal interactions depend upon the relative distance and orientation of the participants, which should be reflected by the RSSI values of captured packets in our experiments. The strength of transmitted radio signals depends on distance and orientation as they are effectively absorbed by the water of the human body. These dependencies are demonstrated by the measurements depicted in Fig. 2.1c and 3.2b. There, in panel Fig. 2.1c, the density plot of RSSI values (y-axis) of captured packets shows a non-linear negative correlation with the distance between participants (x-axis). This measure suggests, for a realistic distance of maximum $1.5 - 2$ meters for verbal interactions, a corresponding RSSI range between $-70 \cdots -75$ dBm, while high intensity regions for lower RSSI values are due to noise and situations of close, non-interacting proximities. This is verified by other measures based on GT2 dataset (see panel Fig. 3.2b), where RSSI values remain within this range for several orientations and radically change only when participants are back to back to each other. However, only visually inspecting these results, it is very difficult to determine a precise RSSI threshold separating real and

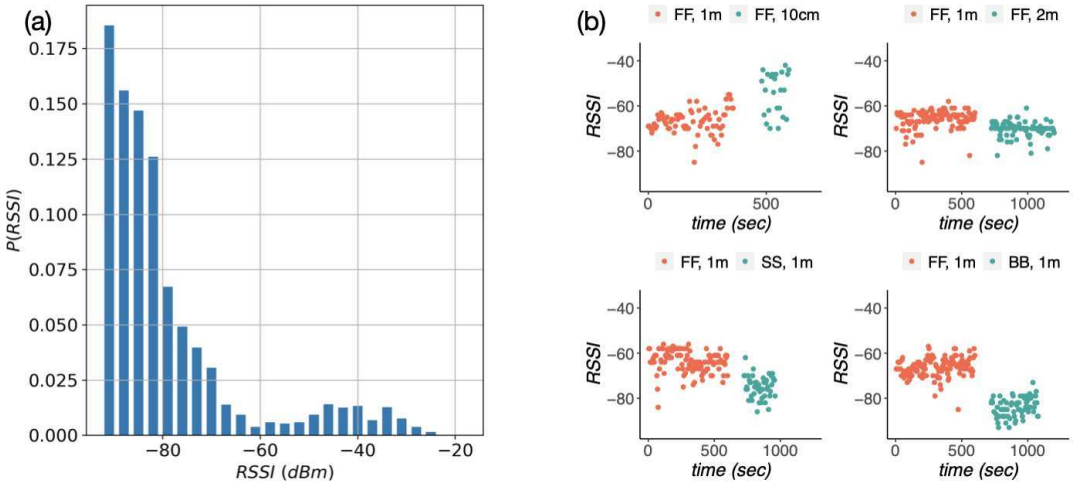

Figure 3.2: LPWD packet statistics. Panel (a) shows the distribution of RSSI values (bin-width $= 3$) observed by single badges over a week (24 hours, 196 badges); (c) Recorded RSSI values as a function of time in settings where the relative orientation of participants has been changed (based on GT2 dataset ; position FF : face to face, SS : side by side, BB : back to back ; distance : 10cm, 1m, 2m.)

false social interactions. To better solve this task, next we frame this question as a classification problem to distinguish between packets indicating real and false social interactions that we can use then to reconstruct the temporal network.

3.1.2 Temporal network reconstruction

In our pipeline, we are going to reconstruct the temporal network from raw data in five main steps, as demonstrated in Fig. 3.1. First, we discuss how to arrive from the recorded data to a handshake pair sequence, whose items indicate mutual handshakes between interacting badges. Then we perform a binary state classification process, turning handshake pair sequence into the binary sequence where each item indicates mutual social interaction state. Finally, we propose several methods to reconstruct the real dyadic temporal interactions with duration, which in turn provides us with a temporal network capturing time-varying interactions between a larger group of individuals.

We separate the binary state classification step from event reconstructions as our first goal was to create a binary sequence of interactions from the raw data that we can apply earlier defined methods on. In addition, we found this approach necessary as we identified two potential sources of errors effective dur-

ing the reconstruction process. One is due to the fluctuations of recorded signal strengths of transmitted packets, while the other is caused by packet loss or inferences, which induce uncertainties in present or absent handshake pairs. This second type of errors makes it difficult to reconstruct events with longer duration, a problem for which earlier studies provided overly simplified heuristic solutions only with limited precision. We will explore various methods relying on this two-steps approach, but show also a method, which solves the problem at once by taking packets with signal strength values as input and directly reconstructing temporal interactions with duration.

3.1.2.1 Binary signal reconstruction

Physical proximity between two participants, A and B, within the right distance range and orientation should appear as a sequence of consecutive mutual 'handshakes' of badges for the duration of their interaction. To obtain the sequence of these handshakes, we take the sequences of packets observed by the badges of A and B and match those packets, which correspond to the mutual observation of the two participants. In other words, since packets are transmitted every 5 seconds, we match two packets into a single handshake event (see Fig. 3.1b) if they appear within ± 2.5 seconds to each other and they refer to the opposite ID (for A from B and for B from A). Missing packets are also recorded in the handshake sequence (see Fig. 3.1b empty red arrows) and are assigned a default RSSI value, -95 dBm, out of the possible RSSI range, that we can easily distinguish from observed packets. To clearly distinguish 'fake signal' from 'real signal', we appended one item called '$pair\_state$' to each handshake RSSI pair to indicate the number of 'real signals' in the pair. This variable can take values 0, 1 or 2, and it enables to code for the presence of 'fake signals' while keeping the normalisation of RSSI values possible for the coming reconstruction methods. Thus the encoding of each handshake pair becomes a vector $(RSSI_A, RSSI_B, pair\_state)$, forming a sequence of handshake events recording all information for the reconstruction task. A handshake signal with appended fake signal is shown in Fig. 3.3.

In order to determine if a handshake pair should be considered as a state of social interaction, we use GT1 where we recorded the start and end time of each social interaction so that we could mark each handshake pair as interaction or non-interaction event. This is shown in Fig. 3.4a, where we plot handshake pairs using their RSSI values as coordinates. Colours code a handshake being an interaction (green) or non-interaction (red). Since different handshake pairs could appear with

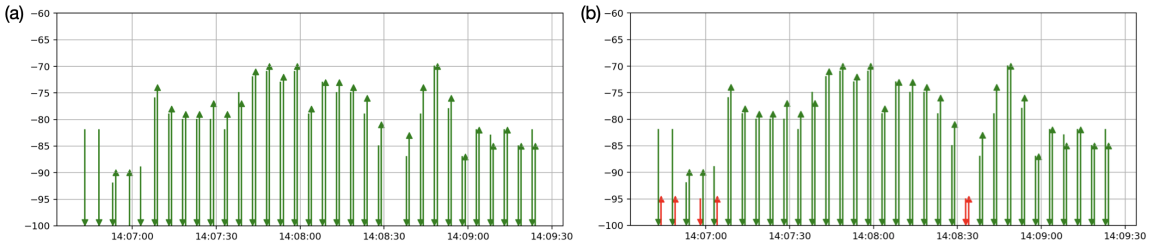

Figure 3.3: (a) is handshake pairs sequence, where signal of different direction marked by arrow's point. The y-axis shows signal strength (RSSI). (b) presents the same raw signal, but with patched "fake signal" marked by red arrow.

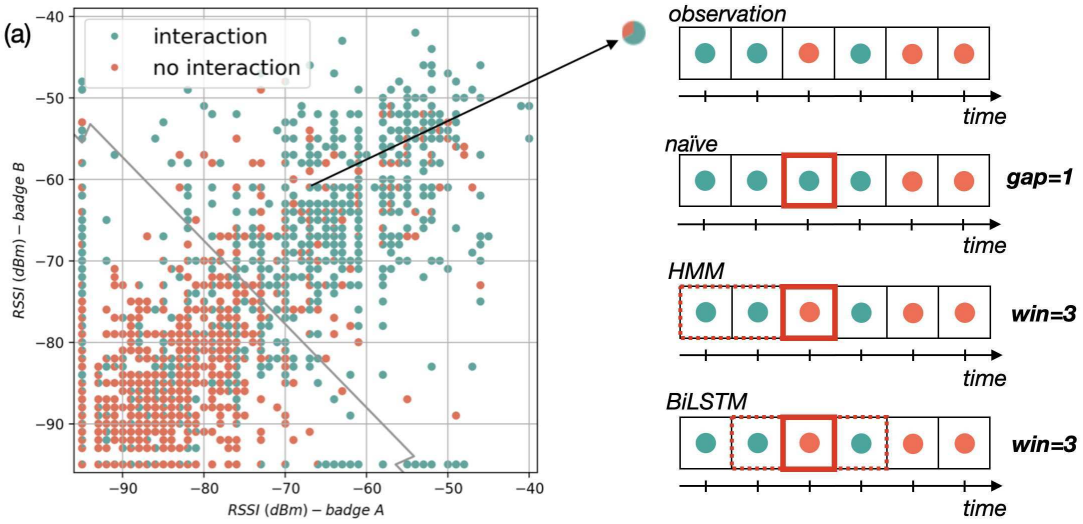

Figure 3.4: Reconstructing binary sequences and interactions. (a) Scatter plot of RSSI values of pairs of interacting (green) and non-interacting (red) badges observed in GT1, with decision boundary presented as grey line; (b-d) Demonstration of reconstruction strategies of (b) an observed binary sequence of interactions using (c) a naïve method with $gap = 1$ threshold, (d) a HMM with window size $win = 3$ and (e) a BiLSTM with window size $win = 3$.

the same RSSI values but different interaction states, in Fig. 3.4a we represent with a small pie chart their fraction at a given location (magnified example pointed by black arrow). The strong diagonal component indicates that the RSSI values of mutual observations are very similar to each other as expected, while the interactions seem to separate from non-interactions around ~ -70 dBm, which corresponds well to the earlier estimated threshold range. To solve this classification problem in a more systematic way, we trained a logistic regression model on the annotated GT1 dataset. As input we gave vectors of handshake pairs and we used their an-

notated labels for the training task. As output we received a probability for each state to be a real social interaction and we thresholded this probability at 0.5 to assign $0/1$ states to each handshake pair. The obtained decision boundary is shown as a grey line in Fig. 3.4a, which appears to be linear, except close to the boundaries where saw-teeth appears due to the two dimensional projection of a three dimensional decision surface. With this method we reached a 77.28% accuracy with 10-fold cross validation (for further details see Table 3.1) to classify a handshake pair as real social interaction or not. This way we can turn our sequence of handshakes into a binary signal (demonstrated in Fig. 3.1c), by assigning $1/0$ to interaction and non-interaction events in every 5 seconds.

Table 3.1: Confusion matrix with accuracy for logistic regression to reconstruct binary signals.

acc.=0.7728	contact	no-contact
contact	0.7728	0.2272
no-contact	0.2272	0.7728

3.1.2.2 Interaction state reconstruction methods

Using the obtained binary sequences, what we call now on un-reconstructed sequences, our next task is to reconstruct the real interactions, which appeared between pairs of participants. The general problem here is to identify false interaction events, which were induced by interference and thus should appear as actual non-interactions, and reconstruct true ones, which were missed due to packet loss. As this is the most challenging task in our methodological pipeline, we are going to follow three different methodological tracks. We will start with a naïve approach commonly used in the literature, then we will explore variants of the Hidden Markov model (HMM) and the Long Short-Term Memory (LSTM) model to find the best solution for this dynamical reconstruction task. Note that while the naïve method only reconstructs interaction periods, the two learning methods naturally adapt to the inverse problem and also reconstruct non-interacting periods with falsely observed interactions in the middle.

Naïve reconstruction model

Consecutive binary signals in a sequence (following each other in 5 seconds here) can be merged into long interaction periods (we call them events) with duration equal to the length of the continuous interaction. These events are separated

by non-interaction gaps. If such a non-interaction gap is induced by an accidental packet loss, it is assuming to be very short. On the other hand, if it is due to a real break of social engagement, it may occupy a longer period. Based on this assumption we can design a very simple reconstruction method, where we merge two interaction periods if they are separated only by a sequence of non-interaction events shorter than a given gap threshold value. This naïve reconstruction method is demonstrated in Fig. 3.4c, where we assume with $gap = 1$ to reconstruct the sequence observed in Fig. 3.4b. This method has been used conventionally in most of the RFID social experiments so far, typically choosing the threshold to be $gap = 0$, thus merging only directly consecutive interaction packets (what we call non-reconstructed method here) or $gap = 1$ corresponding to a gap smaller than 40 seconds. This choice has been challenged recently by [Elmer et al. (2018)], who identified the optimal threshold being 75 seconds for the best reconstruction accuracy.

Hidden Markov model

The second reconstruction method we chose is the Hidden Markov Model that has been introduced in Section 1.4.2. To set the parameters for HMM with supervised learning method we used the GT1 dataset with the annotated states of handshake sequences as states sequence, and the sequence of binary states (explained in Section 3.1.2.1) as the observations. After training on the annotated data (GT1), we determined the values of the conditional transition probability of hidden states as transition matrix, conditional emission probability from hidden states to observation states as emission matrix and the initial states probability as start matrix. In turn, we used these as parameters for the *Viterbi* algorithm to solve the most likely sequence problem and use the output as the reconstructed sequence.

To enrich the information coded in the input sequence, instead of providing a sequence of binary values for each time step, we define a backward window, which contains some short term information before the actual state being reconstructed. More precisely, as demonstrated in Fig. 3.4d, we define a tuple of win number items (there $win = 3$), where the last one is yet the state to reconstruct while the others are the previous states in the sequence. Applying these envelop definitions to a unit of binary state sequence, we create an envelop with backward signals for each signal, thus transforming a binary state sequence into an embedded envelop sequence. Subsequently, we use these transformed envelops instead of binary signals to define the hidden states, observation states, as well as determining all the

matrices. Finally, as an output of the Viterbi algorithm we obtain a sequence of envelops, with last item of each envelop as the predicted interaction/non-interaction state of each time step. Note that we tried multiple other envelop methods (not reported here) coding different distance information between actual and last interaction packets but received worst performance than in the actual case.

Bidirectional LSTM model

The Hidden Markov Model has two limitations in terms of reconstructing the real interaction signal. First, it can only consider states from the past, while states in the future may be also important for the actual state to predict. Second, it is a Markov model thus it can consider only short-term temporal correlations between the actual and previous states. We tried to overcome this shortcoming by introducing longer observation windows for each state, which helped to learn longer temporal correlations yet they were very limited to the actual window size.

Bidirectional recurrent neural networks propose simultaneous solutions to these two problems as they can be trained using input information in the past and future of a specific time frame [Schuster and Paliwal (1997)] (for demonstration see Fig. 3.4e). Especially the Bidirectional Long Short-Term Memory (BiLSTM) model has been shown to perform well on dynamical signal reconstruction (for more detail in Section 1.4.3). This model was initially adopted in speech recognition and showed to improve model performance on sequence classification problems. In practise, it trains two LSTM models on a complete input sequence from opposite directions, one on the input sequence as it is, and another on a reversed copy. The output of each time step from the two LSTM models are merged and passed to the next layer, this way providing some additional context to perform the learning task better.

We applied this model in three different settings to find the best performing one. In one case, that we call BiLSTM-bin, the input of the model was a sequence of binary states we obtained from the classifier's binary output as explained in Section 3.1.2.1. In the second setting, that we call BiLSTM-logi, the input sequence was also generated from the classifier but, instead of binary states, it was a sequence of probabilities obtained as the direct output of the logistic regression before thresholding it. Finally, the third case, that we call BiLSTM-RSSI, is not relying on the sequence of classified states, but instead it takes directly sequences of encoded handshake pair vectors ($RSSI_A, RSSI_B, pair\_state$). This solution has the advantage to skip one step of the reconstruction pipeline and to use a more complex set

of information, but it needs to solve the same problem using noisy RSSI signals without pre-processing.

3.1.2.3 Event reconstruction

To train all these models we used GT1 since it was recorded in the most realistic setting. These data were built from 7 observation clips of 1290, 3060, 3200, 1230, 1740, 1350 and 1030 sec, covering 3 hours 35 minutes combined. For training and validation purposes, we divided evenly observations longer than 3000 seconds (2 clips) into 3 shorter periods, and retained 10 observation clips all with length between 1000 and 1740 seconds (for a total of 3 hours and 18 minutes since we discard small clip with corruption). To determine the best hyper-parameter set for each method, we applied a nested cross validation strategy. In the outer loop, we selected one clip (each of them once) for testing purposes thus we kept it out from the training of the model at this round. From the remaining 9 clips we perform the traditional 9-fold cross validation. Considering all combinations, we could compute the average accuracy over 9 possible divisions of training-validation sets in order to screen hyper-parameter dependencies. Subsequently, we could repeat it 10 times to obtain the average test accuracy with the selected best hyper-parameters. Note that while computing averages we took into account the variance in length of the actual clips used for validation or testing.

Hyper-parameters

All BiLSTM models have two hyper-parameters to define their architecture, the number of hidden neurons and hidden layers. In our computations, we decided to use all architectures with a single hidden layer, which was a sufficient choice for the relatively small training data we have. At the same time, with grid search we explored the dependency of the models on the number of hidden neurons. The results summarised in Appendix A.2 suggested that the performance of the models were weakly depending on this hyper-parameter, but suggested different optimal values for their best performance as summarised in Table 3.2.

The most important hyper-parameter controlling the performance of all of the methods was the window size, which determined the length of temporal correlation a given method could consider. For the naïve model, this window can be associated with the gap parameter (see Fig. 3.4c). In case of the HMM model, as shown in Fig. 3.4d, it is the size of the window that the model considers from the past to infer the actual state. Finally for the BiLSTM models, this window was

Table 3.2: Selected optimal hyper-parameters as number (No.) of hidden neurons, optimal (Opt.) window size, average test accuracy values, and the corresponding standard deviations for each model.

	unrec	naïve	HMM	BiLSTM-bin	BiLSTM-logi	BiLSTM-RSSI
No. hidden neurons	-	-	-	64	16	32
Opt. window size	0	6	6	27	25	27
Average accuracy	**77.28 %**	**83.36%**	**84.25%**	**88.34%**	**89.02%**	**90.03%**
Standard deviation	14.91%	12.11%	11.67%	9.55%	9.36%	8.99%

defined as an envelope of equal number of states before (past) and after (future) relative to the actual state to reconstruct (see Fig. 3.4e).

To choose the best window size, we took it as a parameter to compute the dependency of average accuracy values over the validation sets. As results in Fig. 3.5a depict, the reconstruction accuracy of each of the models shows strong dependency on the selected window size. First, in the case of the naïve method, by increasing the filled non-interaction gap size the accuracy reaches a maximum at $gap = 6$[1]. This corresponds to a gap length of 35 seconds, which is somewhat smaller than the gap window size of 75 seconds reported by [Elmer et al. (2018)] on another RFID dataset. In the case of the HMM model, the best performance corresponds to the same window size $win = 6$. For the BiLSTM models, the accuracy increases with the window size but reaches a plateau at window size $win = 27$ for the BiLSTM-bin, $win = 25$ for the BiLSTM-logi, and $win = 27$ for the BiLSTM-RSSI, after which the reconstruction accuracy decreases.

Performance of network reconstruction

After computing the average accuracy values over the test sets, surprisingly all methods performed relatively well the reconstruction task (see Table 3.2 and Appendix A.2 for the confusion matrices). Even the unreconstructed sequence (abbr. as "unrec" in the tables and figures) reaches a surprisingly high accuracy of 77.28%. On the other hand, the naïve method, commonly used in other studies, performs significantly better with 83.36%, closely matching the performance (84.25%) of the considerably more complicated model of HMM. It is evident, however, that from all tested models, the BiLSTM methods perform the best to solve the temporal network reconstruction. They all provide accuracy at least 4% better than any other method reaching 88.34% for the BiLSTM-bin method, closely matching the values

[1]Note that in case the threshold was $gap = 0$, we obtain the unreconstructed sequence.

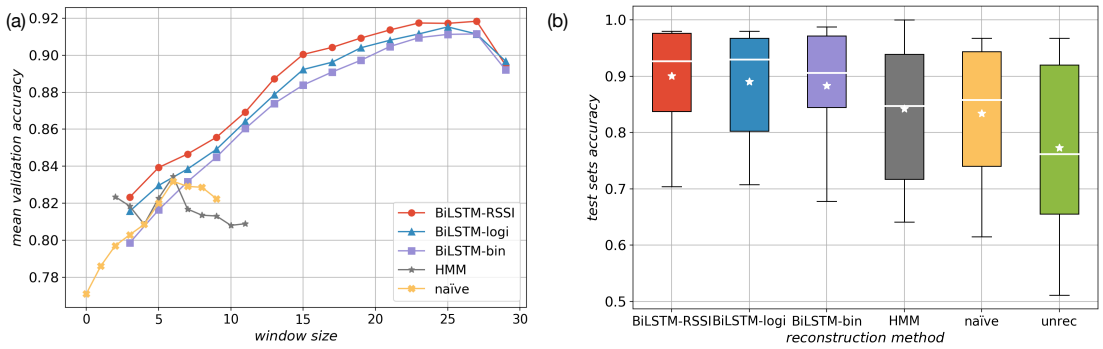

Figure 3.5: Accuracy of temporal network reconstruction (a) Average accuracy values as a function of window size for the naïve, HMM, BiLSTM-bin, BiLSTM-logi and BiLSTM-RSSI models, with fixed number of hidden neurons for BiLSTM models as summarised in Table 3.2. (b) Distribution of accuracy values shown as box-plots for the different models with optimal hyperparameters values summarised in Table 3.2. Horizontal white bar inside box is median and white star is average.

of 89.02% and 90.03% for the BiLSTM-logi and BiLSTM-RSSI methods respectively. More importantly, the best performing BiLSTM methods are also the ones providing accuracy values with the smallest fluctuations over the different test cases. This is reflected by the standard deviation values reported in Fig. 3.5b and Table 3.2 where all performance measures are summarised. In summary, these results suggest that the pipelines with the binary classification and logistic regression provide one of the best performances, but the BiLSTM-RSSI model trained directly on the RSSI values of interaction pairs provide just as good but simpler solution.

The reconstructed temporal network

The different models we introduced may reconstruct the temporal network with different characteristics. First of all, difference may arise as some models would label the same event to be present and some others as being absent interaction. This can be easily demonstrated by looking at the rates of reconstructed interactions by each model, as shown in Fig. 3.6a for a single morning period (3 hours). There, evidently, the highest event rate appears for the unreconstructed signal (naïve method with $gap = 0$) where we only merge consecutive packets labelled as interactions by the binary classifier. Relative to the unreconstructed sequence, each reconstruction method reduces considerably the rate of identified interaction events. The naïve method, being still a very simple model, which merges events maximum 30 seconds apart, appears with the second highest event rate.

Subsequently, the HMM method provides a lower event rate while BiLSTM-logi, BiLSTM-bin and BiLSTM-RSSI methods are closely grouped with lowest rates, reconstructing about four times less events than in the unreconstructed case. The detailed data is listed in Table 3.3.

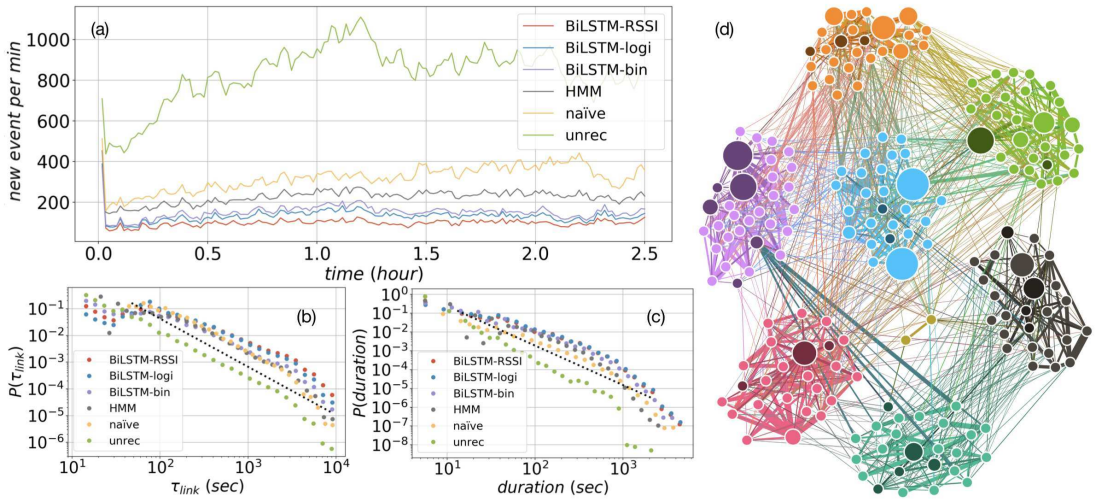

Figure 3.6: Characterisation of the reconstructed network. (a) Number of events per minute reconstructed by different methods (for colours see key). (b) Inter-event time distribution between original and reconstructed event on single links. (c) Distribution of the duration of interactions. Dashed lines on panels (b) and (c) depict approximating power-law functions with exponents 1.8 and 2.1 respectively. (d) Weighted static representation of a reconstructed network using the BiLSTM-RSSI method. Here link widths are proportional to the time spent together when both nodes were present and node sizes are proportional to node degrees. Nodes are coloured according to the original class partitions with darker nodes indicating adults (teachers, assistants or interns). For better visualisation we removed links which correspond to the weakest 3% of weights.

Despite these large differences in the reconstructed volume, the $P(\tau)$ inter-event time distributions between interactions on single links (shown in Fig. 3.6b) and the $P(dur)$ distribution of duration of interactions (Fig. 3.6c) appear with very similar shapes. These distributions all depict broad tails ranging over several orders of magnitudes and can be approximated well with power-law functions with exponents of $\alpha = 1.8$ and $\gamma = 2.1$ respectively. Interestingly, this scaling is very similar to earlier observations in independent RFID studies [Zhao et al. (2011), Cattuto et al. (2010)]. In one way, this match verifies our experimental setting and observations, and at the same time it suggests that heterogeneity present in

Table 3.3: Event numbers of temporal networks reconstructed by different methods. The percentages shows the ratio of reconstructed temporal networks' event numbers to the unreconstructed temporal network's event number.

	unrec	naïve	HMM	BiLSTM-bin	BiLSTM-logi	BiLSTM-RSSI
number of events	715,702	271,567	195,908	130,903	115,230	87,592
% of unreconstructed event number	100%	37.94%	27.37%	18.29%	16.10%	12.23%

the interaction dynamics of face-to-face interactions may be universal with similar characteristics in independent systems.

To demonstrate the structure of the reconstructed network, we chose the BiLSTM-RSSI method as it was one of the best performing models with the smallest variance in accuracy. Using this method, we reconstructed the events recorded in five consecutive mornings (15 hours observation combined) for 165 children and 25 adults (teachers, assistants and interns), and aggregated the obtained interaction sequences into a static network structure. Link weights in this representation were defined as per hour interaction rates between participants. This network is visualised in Fig. 3.6d where we draw links with width proportional to the time the connected nodes interacted during periods when they were both present. The size of the nodes reflects their degrees, while their colours are associated to the class they belong to (with darker colours indicating teachers, assistants and interns). This network structure appears with several interesting characteristics. First of all, the network is heterogeneous in degree, which is a common characteristic of social networks. Second of all, it well recovers the expected community structure where children of the same class connect densely together including the teaching staff in charge of that group.

3.1.3 Spreading processes on reconstructed networks

Temporal social interactions are far from being random but highly correlated in time and structure. They are characterised by heterogeneous bursty dynamics [Karsai et al. (2018)], which potentially appear due to causal correlations between events. Such causally related adjacent events, sharing at least one person in common, build long time respecting paths [Holme and Saramäki (2012), Kivelä et al. (2018)], which are extremely important as they determine how information/epidemics/influence can flow in the temporal network.

Consequently, the precise inference of temporal interactions in a network is extremely important not only to study the emergent structure but any ongoing process, like language evolution or information spreading or epidemics. To demonstrate this issue, here we take all the different event reconstruction methods we explored, and study how the temporal networks reconstructed in different ways influence the dynamics of a simple information spreading process. More precisely, we use the susceptible-infected (SI) model [Barrat et al. (2008)] as one of the simplest prototypical models of information spreading, which in turn can be used to simulate the fastest possible spreading under certain conditions. This model, defined on temporal networks, assumes that each node in the network initially is in susceptible state except for a single randomly selected node, which was set to be infected initially at a randomly selected time. Infection can be transferred with rate β from an infected to a susceptible node (i.e. $S \overset{\beta}{\rightarrow} I$) only at the time and direction of their temporal interactions. In case $\beta = 1$ the model is equivalent to a breadth-first-search process realising the fastest possible information spreading scenario with given initial conditions in the actual temporal network. However, if $\beta < 1$, the process would be arguably less sensitive to local fluctuations in the temporal networks, as it could take alternative routes than the shortest paths to reach nodes, thus would spread slower on the same network. Note that due to the finite observation period of temporal interactions, in our simulation we divided a 150 minutes long observation period into a 30 minutes and a 120 minutes time windows. We selected 800 random seeds from the first window and simulated the SI process for 120 minutes in each case. This way we obtained simulated spreading curves with the same length that we could easily average.

To depict our simulation results, in Fig. 3.7a we show the average spreading curves for each model for $\beta = 1$ case, while in the inset for lower β values the average times the process reached 90% infection on each reconstructed networks. Fig. 3.7b shows the corresponding distributions of time to reach 90% infection in each case, again when $\beta = 1$. All these results indeed demonstrate large differences between spreading dynamics simulated on temporal networks reconstructed with the different methods, despite they all relied on the same raw observation sequences. Not surprisingly, in general, the speed of spreading is largely determined by the overall number of events that the different models reconstructed, as already shown in Fig. 3.6a. Larger number of interactions means larger number of possible transitions between the same set of nodes and over the same period. Following this logic, not surprisingly the unreconstructed network spread the infection the

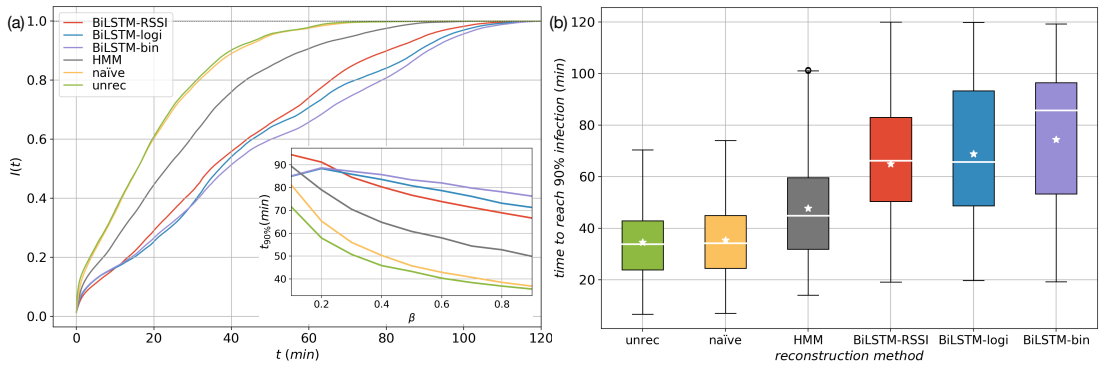

Figure 3.7: Characterisation of information spreading on reconstructed networks. (a) Average spreading curves of susceptible-infected processes with $\beta = 1$ simulated on different reconstructed temporal networks as indicated in the key. Inset shows the β dependencies of the speed of spreading processes measured as the average $t_{90\%}$ 90% infection time. (b) Distributions of 90% infection times as box-plots for the different reconstruction methods in case of $\beta = 1$. Distribution averages are represented by stars. For the parameters of the SI simulations see the main text.

fastest, while BiLSTM models were the slowest. However, there is an important exceptions, which reflects our main conclusion here. The naïve method reduced by more than $\sim 90\%$ the event rates as compared to the unreconstructed sequence, but when it turns to disseminating information, this seems to make no difference. It is suggested by the corresponding spreading curves in Fig. 3.7a, which are almost indistinguishable, and by the distributions of 90% infection time which appear with almost the same average and standard deviation (see Fig. 3.7b). At the same time, these results seem to be consistent over a range of β values. From Fig. 3.7a inset it is evident that at small β values the spreading is strongly stochastic, fluctuations are very large, and the process takes a long time to spread. However, as we increase β the spreading becomes faster on each network. More importantly, after an initial β regime, the spreading processes evolve with similar relative speeds on the different structures as observed in case of the deterministic $\beta = 1$ case. This indicates that even for stochastic settings ($\beta < 1$) the dynamical process is sensitive to the precise reconstruction of the underlying temporal network.

In conclusion, when using wireless proximity sensors to capture temporal interactions, (a) it is very important to carefully reconstruct events from the raw data and not only rely on simplistic intuitive conditions, otherwise the constructed temporal network will be biased by noise and overestimated event rates and will lead to unreliable outcomes of simulated dynamical processes; and (b) it is not enough

to choose the best reconstruction method by its final accuracy, but it is crucial to choose carefully the reconstruction pipeline, which balances between good reconstruction performance and matching the purpose of the actual system under study.

3.1.4 Discussion

The goal of this work was manyfold. First, we developed a filtering and temporal network reconstruction pipeline to obtain the best approximation of temporal social interaction sequences from proximity data recorded via wearable wireless devices. We used ground truth data recorded in various settings and explored different reconstruction strategies involving supervised methods of classification and sequence reconstruction. We found that, while all tested methods provide reasonable performance, naïve methods commonly used in the literature show the worst performance. At the same time, bi-directional LSTM methods, which take into account information from the past and future of the actually predicted state, solve the reconstruction task the best, with accuracy up to $\sim 90\%$.

Furthermore, we wanted to highlight the importance of precise reconstruction of temporal interactions from raw data. Over the last few years, experiments using wearable wireless devices provided an ideal way to study collective social phenomena through the precise recording of temporal social interactions of people/animals in various settings. At the same time, these datasets became inductive resources to study ongoing dynamical processes such as epidemics [Stehlé et al. (2011a), Chowdhury et al. (2009)], opinion dynamics [Maity et al. (2012)], etc. evolving on the temporal social fabric. However, without the careful reconstruction of social interactions, any study addressing the dynamics or structure of the evolving networks or any ongoing collective dynamics would risk to draw wrong or inaccurate conclusions. We demonstrated the sensitivity of this issue by simulating susceptible-infected processes on the reconstructed networks, which in turn follow significantly different scenarios depending on the actual method used for event reconstruction, even for those with comparable accuracy.

As any data-driven study intending to predict or infer human behaviour, our study has also limitations. First of all, the collected data contains certain noise, which cannot be reconstructed with any actual method. Noise is also inevitably present in the ground truth data, which at the same time code only a finite set of configurations used for training, while rare and exceptional scenarios may remain unobserved. These limitations together with the stochastic nature of human behaviour lead to an always perfectible reconstruction of human traits of actions

or interactions. Finally, although we payed special attention on de-noising, pre-filtering, model selection, and the exploration of the hyper-parameter space of each model, surely the optimal inference pipeline we identified is not universal but may be different in the case of data from other wireless proximity sensors.

Beyond scientific merit, our results highlight the importance of the careful design of event reconstruction in studies using wireless sensors. We demonstrated this in the case of LPWD based experiments recording social interactions, but it is important more generally in any study relying on similar data collection methods. This way, we hope that our study contributes not only to the better design of coming scientific studies but also to future emerging technologies.

3.2 Free/class-time Activity Separation

The school schedule assign free-time in the middle of each half-day, during such period the restriction on behaviour and partners during class-time period are unrestricted. Moreover, children from different classes will have the chance to mix. Thus making it essential to distinguish the free-time and class-time once we have reconstructed the temporal network. First I utilised the dynamic of signals to distinguish the period of free- and class-time, then in order to conduct homophily analysis on different population, we designed an algorithm to keep track of class mixture.

3.2.1 Free/class-time periods

While we distinguish between morning and afternoon periods in our data collection according to the school's schedule, we can further divide these periods into sub-periods when the children are in the classroom or in the schoolyard. These two settings allow characteristically different interaction patterns for the participants. During *class-time*, children are limited to interact only with peers from their own class. Moreover, they are commonly seated by the teacher in formations or around a table, in which settings potential social interactions are even further restricted. On the contrary, during *free-time*, children are free to choose to interact with anyone else from their own or any other class, as long as they are present in the schoolyard during the same free-time period. This allows for more spontaneous social interactions, and potentially larger mixing between children from different classes, grades, genders and sociodemographic groups. As we explain

next, we inferred for each social interaction whether it took place during a class- or a free-time period and we share this information in the data as a feature of each temporal network event.

To distinguish between class- and free-time for a given class, we define two types of interactions, the first type is interaction that children had with peers belonging to the same class, which we call *intra-class interactions* and denoted as a. The second type is interaction that children had with peers not belonging to their own class, which we call *inter-class interactions* and denoted as b. For every 10 seconds, we counted the number of a and b, therefore resulting in two sequences:

$$A = \{a_1, a_2, \cdots, a_T\};\ B = \{b_1, b_2, \cdots, b_T\} \tag{3.1}$$

The fraction of these two counts for time t provided the *inter/intra ratio*, denoted as $r_t = \frac{b_t}{a_t}$, and by dividing the two sequence we could get the sequence of inter/intra ratio R:

$$R = \{r_1, r_2, \cdots, r_T\},\ r_t = \begin{cases} \dfrac{b_t}{a_t}, a_t \neq 0 \\ 2 * C_r, a_t = 0 \end{cases} \tag{3.2}$$

where C_r is the ratio threshold which we will explain later. This sequence R sensitively reflects whether the children were in the classroom (the ratio takes low values) or having free-time (the ratio takes larger values). Indeed, the school schedule was organised in such a way that pupils went out on free-time simultaneously with age-peers (of similar grade), thus class-groups successively met in the yard in group of 2 to 4 classes. This is demonstrated in Fig. 3.8a, where the *inter/intra ratio* sequence R is shown as curves for the seven classes in the school during a usual morning period.

As can be seen on that plot, spikes appeared frequently in the observed contact numbers, which could cause false detection of free-time or class-time. To overcome this, before computing the inter/intra ratio sequence R, we smoothed the number of interactions using a Gaussian kernel with $\sigma = 3$:

$$K(x^*, x_i) = exp(-\frac{(x^* - x_i)^2}{2\sigma^2}) \tag{3.3}$$

Therefore the a_t after smoothing will be:

$$\hat{Y}(a_t) = \frac{\sum_{i=1}^{N} K(t,i) a_i}{\sum_{i=1}^{N} K(t,i)} \tag{3.4}$$

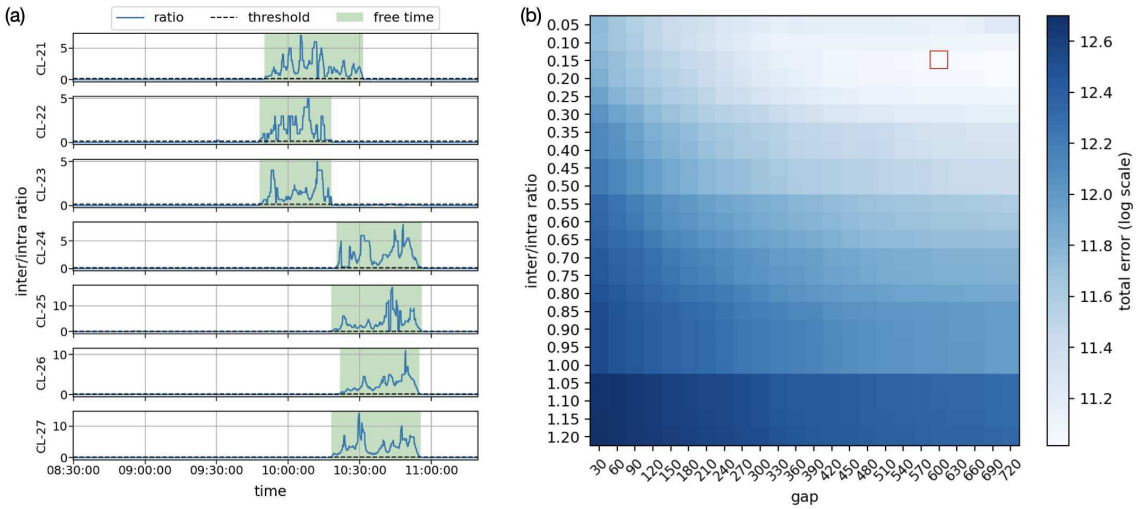

Figure 3.8: Free and class-time detection. (a) Examples of inter/intra ratios (as curves) for one morning period for each of the seven classes with detected free-time periods (green shaded areas) using the optimal hyperparameters. (b) Grid search of the optimal hyperparameters of gap size and inter/intra ratio minimising classification error (coded in blue colour). Optimal parameters correspond to the smallest classification error and were found at gap = 600 seconds and ratio = 0.15, respectively (marked by a red square).

We then developed a method which uses a *ratio threshold*, denoted as C_r to segment every half-day into class- and free-time periods for each class. For time-steps with zero division problem, a default value was set at $2C_r$. Sometimes, the ratio signal fluctuates for short periods in the middle of long free- or class-time periods. This could happen when children from one class in the schoolyard played close to a classroom where other children were taking class. Similar fluctuations appear when a class was leaving the schoolyard and another class stayed there alone for a short period of time until being joined by yet another class. To filter these short periods, we decided to bridge consecutive free-time (or class-time) periods of any given class if these periods were separated by a fluctuating period shorter than a *gap threshold*, denoted as C_g. To determine the best values of these two threshold parameters (C_r and C_g), we used a grid search over an extensive range of the parameter space, as shown in Fig. 3.8b.

To estimate the precision of a given parameter set, we compared the identified free- or class-time periods to the ground truth data. We associated the ground truth free-time period with its start and end time $FT = [t, t']$, and duration $\tau_{FT} = t' - t$. Meanwhile, the corresponding inferred free-time period was between $\hat{FT} = [\hat{t}, \hat{t}']$

with duration $\tau_{\hat{FT}}$. If $\hat{FT}$ and FT overlapped, the overlapping period will be:

$$\text{diff}(\hat{FT}, FT) = [max(t, \hat{t}), min(t', \hat{t}')] \tag{3.5}$$

and its duration denoted as τ_{diff}. Then we could defined error of free- and class-time identification as:

$$err(\hat{FT}, FT) = \tau_{\hat{FT}} + \tau_{FT} - 2\tau_{\text{diff}} \tag{3.6}$$

The minimum error of free- and class-time identification (as compared to ground truth observations, i.e. the recording of the actual schedule of the 7 classes on site, over a four months sample of the data for morning periods) appeared for a ratio threshold C_r of 0.15 and a gap threshold C_g of 600 seconds. The total error computed for all classes with these two parameter values was $60,768$ seconds, which is only 3.13% as compared to the total observation time of $1,941,000$ seconds for the four-month period considered. The identified free-time periods for a typical morning session are demonstrated as shaded areas in Fig. 3.8a.

In most cases, this method resulted in the identification of free-time periods shared by two or more classes, who were simultaneously in the yard, as expected based on our knowledge of the school typical organisation. In few occasions though, only one class was detected as being on free-time. A high *inter/intra ratio* could indeed occur when one class was on the move to the yard, children being spread out in the corridor and passing by the other classrooms, thus having numerous short contacts with pupils from other classes. During such periods where only one class is out of its classroom, interactions among children can still be considered to be restricted to peers from the same class. Thus, we applied a final correction that consisted in aligning the detected free-time periods for the first two classes out in the yard, and the last two classes leaving it, on every half-day. By doing so, we retained only free-time periods where at least two class-groups were together out in the playground, that is periods where pupils could possibly interact with peers from other classes.

3.2.2 Reconstructed temporal network data with free/class-time annotation

As introduced in Section 3.1.2, annotated ground truth dataset GT1 was used to train the model to reconstruct events based on the pre-processed data. Furthermore, since we distinguished between free-time and class-time for each class-group, we annotated each event with four state labels S_t^i $S_{t'}^i$ S_t^j $S_{t'}^j$ and a binary

flag X. State labels are denoted $S^i_{\{t,t'\}} \in \{F, C\}$ where 'F' and 'C' indicate free- and class-time respectively. S^i_t indicates whether i was in class-time or in free-time at time t, while the binary flag X indicates whether i, j are from the same class or not. For example, the event $\theta = (10, i, j, 40)$ and its corresponding label $FCFC1$ would refer to an interaction between two children i and j, belonging to the same class (indicated by $\cdot\cdot\cdot 1$ at the end of the label). In this case, both of them were in free-time at $t = 10$ when the interaction event started, as indicated by the first and third letters (respectively $F \cdot F \cdot$). However, they were in class at the end of the interaction event (at $t' = t + \delta = 50$) as indicated by the second and fourth letters (respectively $\cdot C \cdot C$).

Table 3.4: Free- and class-time events annotation statistics.

Type of interaction	Count	Count (%)	Duration (sec)	Duration (%)
CCCC1	604,065	81.32	66,472,625	88.04
FFFF1	63,009	8.48	4,011,285	5.31
FFFF0	50,878	6.85	2,716,530	3.60
CCCC0	15,347	2.07	1,180,295	1.56
FCFC1	2,641	0.36	343,315	0.45
CFCF1	2,567	0.35	481,575	0.64
FFCC0	1,650	0.22	70,735	0.09
CCFF0	1,007	0.14	39,095	0.05
FFFC0	350	0.05	37,130	0.05
CFFF0	308	0.04	37,795	0.05
FCFC0	213	0.03	31,955	0.04
FFCF0	208	0.03	30,240	0.04
FCFF0	196	0.03	22,030	0.03
CFCF0	146	0.02	12,860	0.02
FCCC0	90	0.01	5,405	0.01
CFCC0	67	0.01	4,695	0.01
CCCF0	52	0.01	3,490	0
CCFC0	45	0.01	1,935	0
CFFC0	5	0	2,205	0

In Table 3.4 we indicate the total count of each type of interaction that we inferred for a sample dataset corresponding to one academic year (i.e. 10 weeks of data), and the total duration of dyadic interactions they represent. Over the sample dataset, most frequently observed types of interactions were $CCCC1$ (81.32% of the total number of interaction events, accounting for 18,465 hours of dyadic contacts) corresponding to classmates interacting during class-time, and $FFFF1$

(8.48%, i.e. 1,114 hours) corresponding to classmates interacting during free-time. Less frequent patterns were $FFFF0$ (6.85%, i.e. 755 hours) corresponding to children from different classes interacting during free-time, $CCCC0$ (2.07%, i.e. 328 hours) corresponding to children from different classes interacting during class-time (e.g. visiting each other class, meeting in the corridor or in the lavatory), and $CFCF1$ (0.35%, i.e. 134 hours) and $FCFC1$ (0.36%, i.e. 95 hours) corresponding to classmates starting an interaction in class that continued in the yard or vice-versa.

3.2.3 Free-time class grouping

Once the free-time and class-time periods for each class are split, in order to construct temporal network for free-time and class-time periods, the following step should be allocating social interaction to the period it belongs to. Each interaction has four elements: $child\_A, child\_B, start\_time$ and end_time, where time information and kids involved should be used to assign them to different periods. The easier task is to allocate interaction to class-time period: for a class C, an interaction belongs to C's class-time period if $start\_time$ and end_time are within the range of its class-time period and both $kid\_A \in C$ and $kid\_B \in C$. Allocating interaction to free-time period is however more subtle. Recall that multiple classes, denoting the set of classes as C_{ft}, could have free-time together, therefore an interaction should be allocated to free-time period if both kids are from class $C \in C_{ft}$, which requires us to gather free-time periods of all classes and group them into ranges with unique associated classes. For instance, if free-time period of class C_a is $[0, 5]$ and free-time period of class C_b is $[3, 8]$, ranges after splitting will be [0,3] with $\{C_a\}$, [3,5] with $\{C_a, C_b\}$ and [5,8] with $\{C_b\}$. This is also necessary when analysing homophily during free-time since we need to consider the population with different class compositions.

To split our morning and afternoon observation periods into shorter periods associated to the actual list of classes on free-time together, we first defined a data structure called $action$ with three items: [*type*, *tsp*, *class*], where type is the start (S) or end (E) of a free-time period, *tsp* is the timestamp of the action, and *class* indicates to which class the action refers to. With the definition of action, each free-time period could thus be represented as two actions: *start action* $[S_a, tsp_{S_a}, C_a]$ and *end action* $[E_a, tsp_{E_a}, C_a]$ for class C_a for example. In addition, we needed to compute the *active_list* to record active class(es) for each period (between a given *S* and *E*). Actions of all classes were then sorted chronologically (note that end action should appear prior to start action if their timestamps collide). Next, we

went through the list to spot any action that would cause a change in the active list. We could then define the range between the last action and a new action, as well as their associated class(es). Detailed process is described in algorithm 1.

Algorithm 1: merge-split of free-time

input : chronologically sorted actions of all free-time periods
output: split ranges *range_list*, with associated class(es) of each range *classes_list*

```
active_list ← ∅; range_list ← ∅; classes_list ← ∅
range_start ← 0; range_end ← 0
for act in actions :
    if act.type = S then
        if size(active_list) > 0 then
            range_end ← act.tsp
            if range_start ≠ range_end then
                range_list.add([range_start, range_end])
                classes_list.add(active_list)
            end
        end
        active_list.add(act.class)
        range_start ← act.tsp
    else
        range_end ← act.tsp
        if range_start ≠ range_end then
            range_list.add([range_start, range_end])
            classes_list.add(active_list)
        end
        active_list.remove(act.class)
        if size(active_list) > 0 then
            range_start ← act.tsp
        end
    end
end
```

3.2.4 Boundary corrections for free-time class group

We noticed that some fragmented ranges, which we call *unstable ranges*, that would be generated after aforementioned process, mostly due to the case that start/end of free-time periods of different classes are very close. Social interactions during this period is likely to be inadvertent so it is more surefooted to remove interactions

within these unstable ranges. Also, in order to preserve more meaningful interactions in unstable ranges, we don not simply remove all interactions in these small ranges. We use a scheme as shown in Fig. 3.9a where A, B are two classes, S, E is start and end action respectively, marked by vertical dashed line. Horizontal lines represent free-time period of corresponding class. For unstable range formed by : 1. two start actions, we postpone the tsp of the first start action to tsp of the second one; 2. first a start action then an end action, we bring forward the tsp of the end start action to tsp of the start action; 3. two end actions, we bring forward the tsp of the second end action to tsp of the earlier one. We sort unstable ranges by size and iterate from short to long until no more unstable range exists. Like the problem in Section 3.2.1, here we have to decide how short is the range to be considered as an unstable range, the curve of duration and number of left ranges when threshold changes is plotted in Fig. 3.9b. It could be seen that there are not much ($< 10\%$) data lost even with threshold of 300 seconds (less than 20000 seconds lost out of 205000 seconds), so we took a middle value of 3 minutes so that when we will not risk deleting many real social interactions.

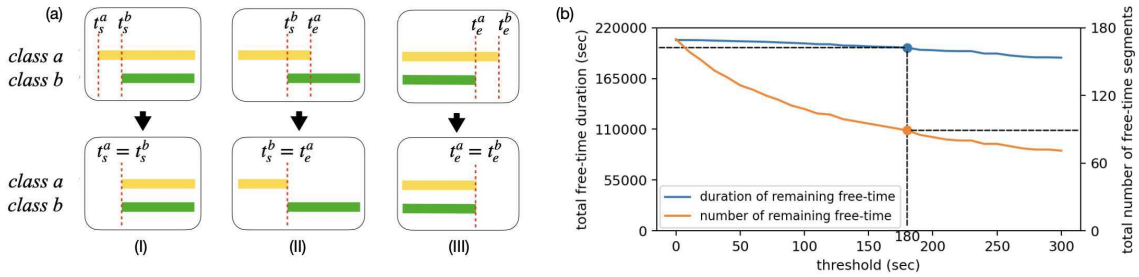

Figure 3.9: Illustration of boundary corrections. (a) Strategies for removing short overhanging segments indicated between the two red dashed lines in the top panels: panels I present the case of non-matching beginnings, panels II show the case of overlapping beginning and end times, and panels III present the case of non-overlapping endings of free-time periods and their corrections. (b) Loss of total free-time duration and number of free-time segments. The dashed line indicates the selected time threshold of 180 second corresponding to $\sim 50\%$ segment number loss and only $\sim 5\%$ total duration loss.

After process above, free-time periods and class-time periods are separated, classes are also properly grouped. Since the social behaviour and language development is a process of gradual change, hence we do not need to keep a very high temporal resolution and it is advisable to merge meaningful interactions into aggregated networks. For each free-time period, interactions within period of each class group could be merged into a network that includes all children from the

group of classes, the link weight is defined as the summation of duration of interactions since it contains more information. Interactions within class-time periods on the other hand, should be divided into 7 different networks since each class during class-time is a network with its children as nodes. An example of class grouping and boundary corrections result is shown in Fig. 3.10.

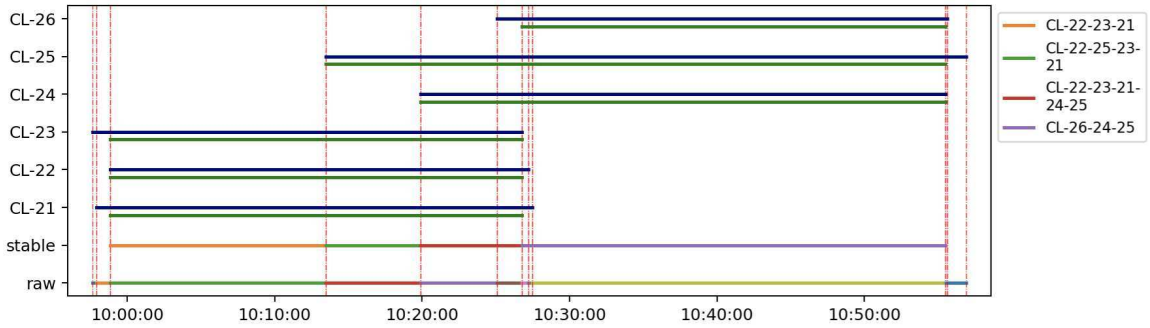

Figure 3.10: Example of boundary correction result. Blue lines are the original free-time for each class, green lines are free-time after boundary correction. *Raw* is the result after free-time class grouping. *Stable* is the result after boundary correction, different colours represent different class grouping, which is labelled on top-right. The vertical red dashed lines mark the start and end of free-time of each classes

3.3 Technical Validation

To validate the structure and dynamics of the inferred social temporal network, we present here some frequently used statistics about the unreconstructed and reconstructed temporal structures. Note that although some of these network characters have been discussed in Section 3.1.2.3, but here we discuss with more focus and additional information. For this presentation, we randomly chose a week in the dataset and aggregated its morning periods into a sample network. First, the distribution of event duration, presented in Fig. 3.11a, appears with fat tails for both unreconstructed and reconstructed networks, similar to earlier observations made in independent systems Zhao et al. (2011). However, while the distribution corresponding to the unreconstructed network decreases monotonously, the distribution for the reconstructed network starts with a plateau (corresponding to the range of 5 - 30 seconds). Moreover, probabilities for the first three duration values (5 - 15 seconds) are lower for the reconstructed network than for the unreconstructed one. That reduced probability of short duration is due to the reconstruction method merging short events separated by short inter-event times

into longer interactions, making low values less frequent. The reconstruction process also accounts for the different scaling of the tails of the distributions. Indeed, longer events appear with a higher probability and over a larger range in the case of the reconstructed network.

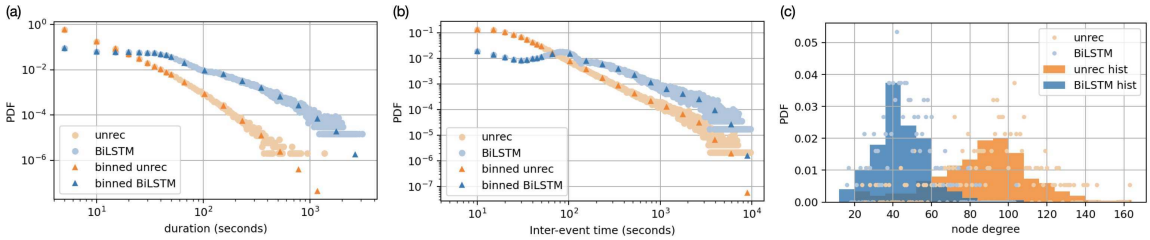

Figure 3.11: Statistical characteristics of the observed social temporal networks. (a) Distribution of event duration; (b) Distribution of inter-event times on links; (c) Distribution of node degrees. Dark symbols and bars indicate the logarithmic binned distributions of the corresponding probability density functions (PDF) (light symbols). Results are shown for the unreconstructed (orange) and the BiLSTM reconstructed (blue) structures.

Similar differences can be observed for the inter-event time (IET) distributions measured on links, which are shown in Fig. 3.11b. This metric measures the length of time between the end of an event and the beginning of the next one for each pair of interacting participants. Its scaling indicates how heterogeneous an activity sequence is in time. If it appears with a broad tail, it demonstrates that the observed dynamic is bursty, thus characterised by short periods with events of short inter-event times separated by long periods of inactivity. As it has been observed in similar systems Zhao et al. (2011), the inter-event time distribution appears with a long tail both for the unreconstructed and reconstructed temporal networks. Similar to the duration distributions, short inter-event times become underrepresented after reconstruction as compared to the unreconstructed case. For larger values, both distributions scale close to a power-law, but with longer IETs more likely in the reconstructed temporal network. These differences can again be explained by the merging process, which commonly bridges short inter-event times between events, this way creating longer interactions and removing short inter-event times. Nevertheless, the scaling of the tails of both distributions is similar since the reconstruction model does not connect two events separated by a very large gap, thus not affecting the frequencies of long IETs.

To verify some structural character of the aggregated network, we measured its node degree (number of neighbours) distribution, plotted in Fig. 3.11c. Similar to

earlier observations Zhao et al. (2011), the degree distribution of both of the networks indicates certain heterogeneity. The unreconstructed network appears with a broader range of node degrees, ranging from 24 to 163 with an average of 90.4, while the reconstructed structure is more homogeneous with degrees between 16 and 104, for an average of 47.1. This difference is evidently caused by the reconstruction method, which filters a number of falsely observed interactions induced by noise and other external effects. Furthermore, the reconstructed network, by observing preceding and succeeding signals, gives a more accurate evaluation of whether exchanged signals between two participants indicate a real interaction or just an accidental encounter. Therefore, the number of neighbours is reduced in the reconstructed network.

Finally, to present directly the structure and long term evolution of the observed social network, we show three successive networks in Fig. 3.12 during class-time (upper panels) and free-time (lower panels) periods, which were aggregated over three distinct weeks of observation to illustrate the beginning, middle, and end of the academic year (i.e. September, February and June, respectively). Here, node sizes indicate the total duration of interaction for each participant, the width of links scale with the total duration of interaction between peers, and dark shaded nodes are assigned to adult participants. Classes coloured in different ways are easily identifiable from the structure, especially during class-time observations (see Fig. 3.12 a-c). On the other hand, free-time observations (Fig. 3.12 d-f) evidently demonstrate the different, potentially inter-class mixing patterns between children.

3.4 Summary

In this chapter, we first presented a pipeline of reconstructing signal data to temporal network. After applying different techniques from the field of machine learning in the pipeline, we achieved 90.03% accuracy with BiLSTM, which is significantly better than conventional methods with 77.28% accuracy, and outperforms other methods we tested. Then we used a spreading model to simulate information propagation on networks reconstructed by different methods. The results are generally consistent with the event rate of reconstructed network: higher event rate means faster spreading. However, even though the *naïve* method significantly outperforms the *unreconstructed data* in reconstruction accuracy, their reconstructed

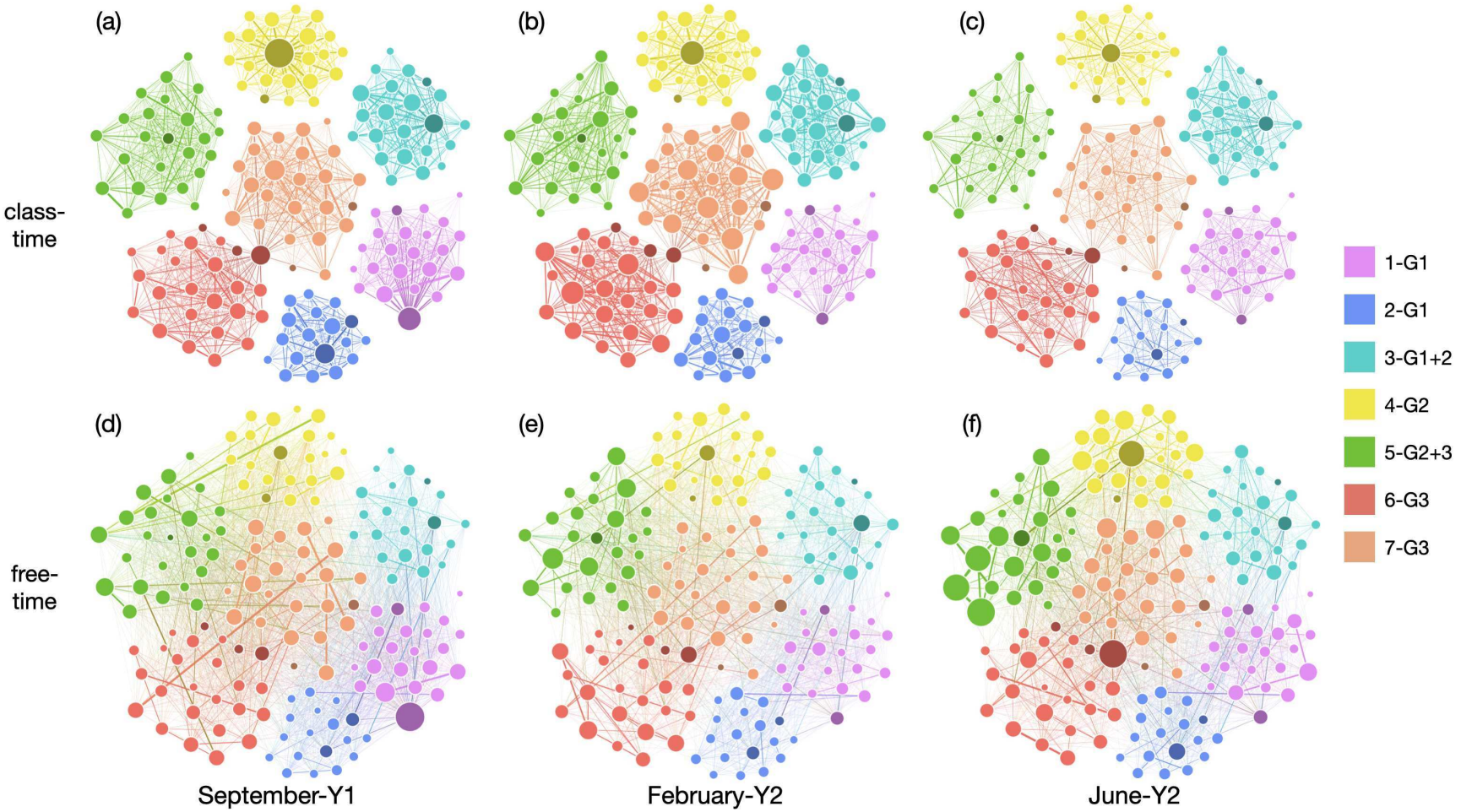

Figure 3.12: Aggregated network for morning periods over a week in September, February and June of the academic year. Networks are shown in panels (a-c) for class-time interactions, and in panels (d-f) for free-time periods. Node colour indicates the class, node size scales with the total duration of interaction time for the corresponding participant, and edge width represents aggregated duration of interaction between two participants. Nodes with darker shaded colours are adult participants. In the caption, next to the class number, the class-group composition is given in terms of grade(s): G1 for 1st grade, G2 for 2nd grade, G3 for 3rd grade.

network lead to similar spreading dynamics. Moreover, *naïve* and *HMM* reconstruct networks with similar accuracy, but with significantly different spreading speed. These cases demonstrated the importance of precise network reconstruction and the careful choice of the reconstruction method. These results are published in [Dai et al. (2020)]. In next task of activity reconstruction, we used number of intra- and inter-class interaction to separate free- and class-time with 3.13% error rate. Finally in order to conduct homophily analysis, we developed algorithms to group classes according to their mutual presence of free-time. This method lead to only $\sim 5\%$ free-time lost but reduced largely ($\sim 50\%$) the number of different class group combinations. In the end we validated our methods with several examples. The results of the second part of this chapter were summarised in [Dai et al. (2022)] and the data has been shared in [Nardy et al. (2022)]

Chapter 4

Network Homophily Analysis

As the profound proverb "birds of a feather flock together", the presence of homophily has been discovered and studied in plentiful networks, with the most attention being paid to social networks. This link creation mechanism, that we already discussed briefly in Section 1.2.3, induces more likely social interactions between peers of people similar in terms of gender, religion and organisational role or social status, etc. [McPherson et al. (2001)]. However, seldom does researchers have access to a longitudinal network data with considerable change to observe the dynamic of homophily. Also, the entanglement between homophily and influence is still not fully resolved. Here in this chapter of the thesis, with the network and sociodemographic data collected by DyLNet projects, we try to bring some new insights of the aforementioned topics.

To achieve this, first of all, social contact data collected of 4 consecutive months are selected as study sample. Then we selected 8 most fundamental attributes as node features, as well as 3 different homophily indices to compute the homophily effect. We used a reference model to fully permute empirical network as the baseline to represent the homophily in randomised cases. Furthermore, due to the natural difference of class-time and free-time, social contacts are separated into respecting class/free-time periods thus we could further compare mixing in free- and class-time by observing their difference of homophily. Finally, we conducted cross-sectional observation through grouping participants by their grades to demonstrate the development of homophilic preferences with age.

4.1 Sample Data and Attributes Description

To analyse homophily, we focus on six sociodemographic attributes, namely: 'sex', 'child_lang' (dominant language of the child), 'mother_occ' (occupation category

Table 4.1: explanation of attributes and their values

attribute	description	value	value explanation
sex	gender	*f*, *g*	'*f*' means girl, '*g*' means boy
child_lang	does the child understand or speak other language(s) other than French	*oui*, *non*	'*oui*' is yes and '*non*' means no
mother_occ	category of mother's employment, in 3 levels	1, 2, 3	1: basic profession; 2: intermediate skilled profession; 3: High skilled profession
father_occ	category of father's employment, in 3 levels	1, 2, 3	1: basic profession; 2: intermediate skilled profession; 3: High skilled profession
mother_dip	category of mother's diploma, in 3 levels	1, 2, 3	1: basic degree (below bachelor); 2: intermediate degree (between bachelor and master); 3: Higher degree (master and higher)
father_dip	category of father's diploma, in 3 levels	1, 2, 3	1: basic degree (below bachelor); 2: intermediate degree (between bachelor and master); 3: Higher degree (master and higher)
voc	performance of vocabulary in language test	1, 2, 3	1: lowest 25%; 2: between 25% to 75%; 3: Top 25%
synt	performance of syntax in language test	1, 2, 3	1: lowest 25%; 2: between 25% to 75%; 3: TOP 25%

of mother), 'father_occ' (occupation category of father), 'mother_dip' (education level of mother) and 'father_dip' (education level of father). Additionally, we have two linguistic performance attributes that we are especially interested: 'voc' (vocabulary size) and 'synt' (syntactic development level). More details about these attributes see Table 4.1. Among these attributes, 'sex' and 'child_lang' are binary attributes and the values of other attributes are categorised and digitised into 3 groups: 1 (low), 2 (middle) and 3 (high). The distribution of each attribute's value is shown in Fig. 4.1, from where the heterogenous background of children is evident. This endows the data with potential of observing the behaviour and dynamics of children from different social backgrounds. Besides, as illustrated in Section 2.2.2, the observed participants are from of 7 classes and 3 grades, hence enabling us to conduct cross-sectional analyse with different age groups involved. In this end, in order to exclude interference between different age groups and to

reach equilibrium in group size, we selected 2 classes of 1st grade (PS) and 2 classes of 3rd grade (GS) for our research focus.

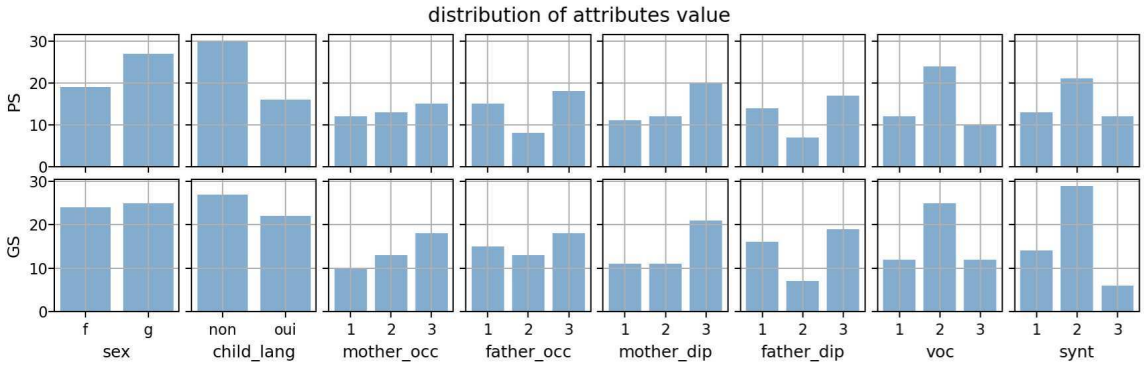

Figure 4.1: distribution of each attributes, grouped by grade (PS - 1st grade (petite section (in French), GS - 3rd grade (grand section)).

After the selection of features and participants, it is also vital to choose the identify precisely the social ties which connect individuals in the temporal social network. As described in Section 2.2.3.1, the data collection is comprised of two parts: morning (8:30 to 11:20) and afternoon (13:45 to 15:50) part. For our analysis we decided to choose the morning part due to two reasons. First, in the morning there are both types of activities in the schedule of the preschool, class-time and free-time. Their different behavioural patterns and social mixing could lead to more discovery. Second, children are more active during morning part, in contrast, children of 1st grade (PS) usually had a nap during afternoon parts.

4.2 Homophily Analyses

While socialising, people exhibit the tendency of choosing similar partners (e.g. in terms of race, gender and socioeconomic status) more frequently than random from the same population. However the ubiquitous presence of homophily may lead to critical social problems such as inequality and segregation [Moody (2001), Blau and Schwartz (2018)]. Studies has found that there are many reasons accounting for people's preference of forming ties with similar others. On one hand, the similarity is beneficial to the communication and understanding, and helps to establish trust and solidarity with the counterparts [Festinger (1957), Werner and Parmelee (1979), Portes and Sensenbrenner (1993), Mollica et al. (2003)]. On the other hand, the cost of maintaining a social ties would be lower between similar than between dissimilar partners [Felmlee et al. (1990)].

In reality, however, an individual's social relations are not entirely decided by preference, but constrained by the environments where he/she lives, such as schools, workplaces or neighbourhoods. The population of these environments are not uniformly distributed in terms of race, gender or wealth. Therefore, the exhibited homophily arises from two theoretically distinct mechanisms: the first is *induced homophily*, which originates from the homogeneity of structural opportunities for interaction, for example in neighbourhoods and workplace; the second is *choice homophily* which reflects people's preference in choosing whom to interact with [McPherson and Smith-Lovin (1987), Kossinets and Watts (2009)]. Furthermore, in order to include the effects of biased sorting of individuals among groups, *inbreeding* and *baseline homophily* are introduced in [McPherson et al. (2001)] where baseline homophily is defined as the level of homophily expected from random mixing in the population, and inbreeding homophily is defined as the level of homophily in excess of that baseline. In this chapter we focus on the *inbreeding* and *baseline homophily* of our data.

4.2.1 Random network construction

In network science, randomisation techniques shuffle the given network to produce a random structure. Randomised networks serve as reference when trying to assessing whether a character of an empirical network is significant. Randomisation techniques on graphs have been studied and applied in considerable amount of literature for various fields [Milo et al. (2002), Gionis et al. (2007), Hanhijärvi et al. (2009), Ray et al. (2012)]. In our study, the observed homophily in the empirical network is the accumulation of baseline and inbreeding homophily, and the homophily in randomised network could be considered as baseline homophily. To observe the difference of these two types of homophily, the key of randomisation of our network is to keep the properties of each individual, for example node degree (number of peers) or strength (total duration of interactions), while rearranging the network structure to eliminate the correlation caused by attributes of node. There are several randomisation schemes we have considered, namely:

- *weight randomisation*, which keeps the edges as where they are but shuffles their weights.

- *attribute randomisation*. Attribute randomisation shuffles attribute of nodes while preserving network structure and edges weight.

- *link swap*. Select two dyadic pairs randomly (no overlapping of nodes), rewire the edges such that could form two new dyadic pairs. Repeat this process for enough times.

The weight randomisation could preserve the connectivity and neighbour structure of network, however this randomisation possess lacks the capability of bridging disconnected pairs in empirical data, in another words the structure is not shuffled enough. With attribute randomisation, all structures are preserved, but the connection of attribute and structural properties of nodes will be completely destroyed. Finally with link swap, we are able to achieving our goal of randomisation. By repeating rewiring in Markov process sufficiently, the network loses its original topography. However in our network, the edges (interactions) are weighted with aggregated duration time, which we need to take into consideration when rewiring if we want to preserve the node's strength. DCWB (equal-weight link-sequence shuffled) method proposed in [Karsai et al. (2011), Gauvin et al. (2018)] allows links to be rewired only if they are with same weight. Unluckily in our data, the edge weight is sparsely distributed within the range of weight, so we have to make a compromise proposal to rewire edges with approximate weights. More precisely, we first group all edges to 5 linear bins by their weights, then rewire edges within each bin. The network shuffled in this way is therefore fully randomised and almost preserve the strengths of each node.

4.2.2 Homophily analysis with EI index

E-I index proposed in [Krackhardt and Stern (1988)] is a straightforward and classic ego-centric individual level homophily index. Its simplest definition expresses the homophily with EI $= (EL - IL)/(EL + IL)$ where EL, IL represents number of external and internal link respectively. Since children have their own attributes, in our case internal means interaction happens between children with same value for a given attribute. The E-I index range from -1.0 to 1.0, with values representing homophilic (-1) to heterogeneous (1) patterns, while 0 indicates neutral contacts with the same number of external and internal links in one's egocentric network. Given the fact that duration is a pivotal feature of an social interaction, we modify the EI index to weighted form:

$$\text{EI} = \frac{EW - IW}{EW + IW} \tag{4.1}$$

where EW and IW indicates weights of external and internal links respectively. EI index applied on our data indicates the homophily tendency of a child when choosing friends. We compute average EI index of each individual on every attributes for our empirical network and randomised network. To be more precise, recall that in Section 3.2.3, we have segmented the freetime into group of classes, denote all free-time class groups during our selected 4 months as Ω. Then for child i, its average E-I index is defined as:

$$\mathrm{EI}_i = \frac{\sum_{f \in \Omega(i)} \mathrm{EI}_i^f}{|\Omega(i)|} \tag{4.2}$$

where $\Omega(i)$ is set of free-time class groups that child i participates, $|\Omega(i)|$ is the size of $\Omega(i)$ and EI_i^f is EI index for kid i in free-time class group $f \in \Omega(i)$. In practice, when randomising network we rewire each network 50 times (i.e. $50\times|E|$ rewiring, where $|E|$ is the edge number) to sufficiently remove any residual correlation from the structure. For each free-time class groups we iterate aforementioned randomisation procedure 100 times to create 100 randomised sample hence 100 EI_{rand} scores, trying to cover enough possible randomised network topology.

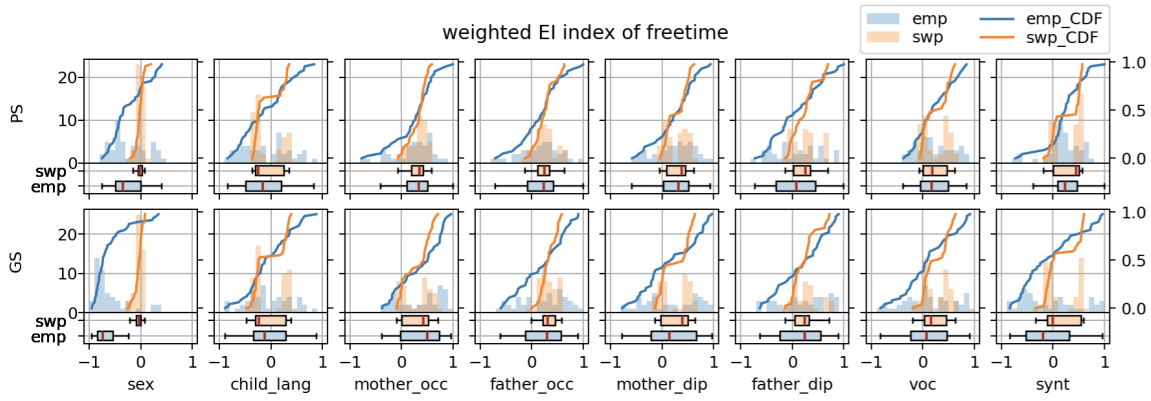

Figure 4.2: Probability density and cumulative distribution functions of E-I index for free-time periods. In the legend 'emp', 'swp', 'hist' and 'CDF' means 'empirical', 'swapped', 'histogram' and 'Cumulative Distribution Function' respectively, same notation applies below.

Firstly in Fig. 4.2 we have the distribution of EI index of free-time periods, we divided the EI indices by grade so that we could also make cross-sectional comparison between different age groups to discover the dynamic of homophily tendency changing along with age. In the figure we have histogram (y-axis on the left) and eCDF (empirical Cumulative Distribution Function) curve (y-axis on the right). From the figure we could judge intuitively by the median value (red line

in boxplots), for the EI index distribution of gender (sex) there is significant difference in observed data and randomised data for both grades, which shows overwhelming homophily phenomenon, and the tendency of inbreeding homophily is getting even stronger with growing of grades if we compare different age groups. While for randomised network, for both grades the EI index are centred around 0, the neutral value. Its small variation of 100 randomisation also proves that our randomisation process is sufficient to permute the network. Since network after rewiring is entirely randomised, the deviation from 0 should be owing to *baseline homophily*, for attribute *sex* is the imbalance of gender in group. The difference of EI index of empirical and link swapped network should be attributed to *inbreeding homophily*. Other attributes with significant inbreeding homophily includes *mother_occ* for GS, *father_occ* for PS, as well as linguistic attribute *voc* for GS and *synt* for both grades.

Table 4.2: KS-test for EI index in free-time

		sex	child_lang	mother_occ	father_occ	mother_dip	father_dip	voc	synt
PS	KS	0.59	0.37	0.23	0.27	0.21	0.42	0.24	0.26
	p-val	0.000	0.003	0.231	0.086	0.269	0.002	0.123	0.072
GS	KS	0.88	0.18	0.32	0.37	0.40	0.29	0.35	0.49
	p-val	0.000	0.346	0.025	0.003	0.002	0.052	0.004	0.000

To analyse the inbreeding homophily more accurately, we use *KS test* to measure the difference of EI index distribution in empirical and randomised cases. KS test is a non-parametric test to determine the probability that a sample is drawn from a reference probability distribution, or two samples are drawn from the same (unknown) distribution, which is our case. For two given probability distributions, the KS test quantifies the absolute maximum distance between the empirical distribution functions of two samples. KS test results of free-time EI index distribution in empirical and randomised cases are shown in Table 4.2. From the table, combined with corresponding Fig. 4.2, we could see that gender (sex) exhibits the largest inbreeding homophily in empirical network with significant difference from randomised counterpart, moreover, the inbreeding homophily grows with age. For *child_lang*, we could observe significant difference of empirical and randomised EI distribution in PS (given by the small p-value), while not in GS. However no homophily effect is observed from Fig. 4.2. The attribute *father_occ* does not show inbreeding homophily in PS but homophily appeared in GS, same situa-

Table 4.3: KS-test for EI index in class-time

		sex	child_lang	mother_occ	father_occ	mother_dip	father_dip	voc	synt
PS	KS	0.18	0.18	0.31	0.35	0.12	0.19	0.18	0.13
	p-val	0.437	0.437	0.039	0.011	0.910	0.479	0.437	0.790
GS	KS	0.41	0.14	0.12	0.17	0.16	0.17	0.16	0.16
	p-val	0.001	0.665	0.902	0.452	0.580	0.564	0.494	0.494

tion happened for *mother_dip* and two linguistic attributes *voc* and *synt*. However *father_dip* showed the opposite trend with homophily fades with age growth.

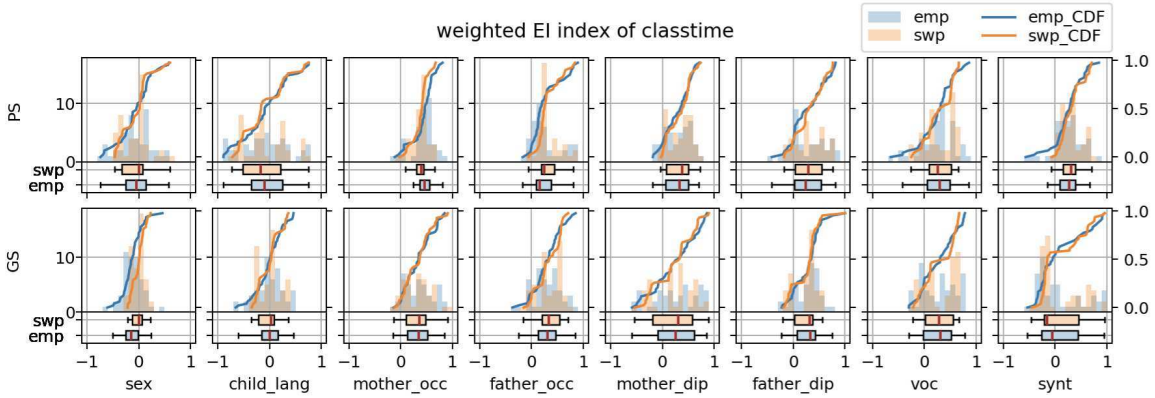

Figure 4.3: distribution and eCDF of EI index for class-time

For comparison, the distributions of EI index of class-time periods are shown in Fig. 4.3. Network during class-time is in essence a "randomised" network because children are randomly assigned to fixed seats and do not have much choice of whom they talk to, as we demonstrated in Section 3.2.1. Therefore the distribution of EI for empirical and randomised cases in class-time intertwines together for most attributes and grades, with exception of *sex* in GS as well as *mother_occ* and *father_occ* in PS. Same conclusion could be derived from the very large p-value in Table 4.3.

4.2.3 Homophily analysis with Coleman index

Different from EI index, which measures homophily in ego-centric fashion, the index proposed by Coleman [Coleman (1958)] measures group homophily. It concentrates on links of a homogeneous group of people in excess of their population shares and then normalise it by the group proportion. Coleman index is popular in various fields [Currarini et al. (2009; 2010)] due to its simple definition given as:

$$\text{Coleman}_L = \frac{H_L - w_L}{1 - w_L}. \tag{4.3}$$

Here $w_L = N_L/N$, which is the population share of group L, N_L is the population of group L, N is total population. w_L is also the expected proportion of homophilic matches that would result if links in the network would be randomly matched. $H_L = m_{LL}/m_L$ is called *group homophily* of group L, with m_{LL} being the link number inside the group L and m_L is the total link number that all members of group L is involved in. The numerator of Coleman index is called excess homophily, which compares H_L with w_L. Coleman index facilitates the comparison of different groups by normalising the excess homophily of group L by its maximal possible value $1 - w_L$.

The range of the Coleman index is between $(-\infty, 1]$, corresponding to a heterogeneous or to a homophilic network (respectively). 0 value implies that for the given group, proportion of internal links are equal proportion of the group over the whole population, which could be interpreted as neutral. Since the Coleman index is defined on the group level, in our case we define a group L_v^g including children at the grade g with attribute values equal to v. For the same reason in Section 4.2.2, we take link weights (total duration) into consideration, therefore instead of link number, the definition of m_{LL}, m_L is replaced by the total weight of links. As with the average scheme in Section 4.2.2, we use the average Coleman index over all free-time of a class to represent the Coleman index of a group.

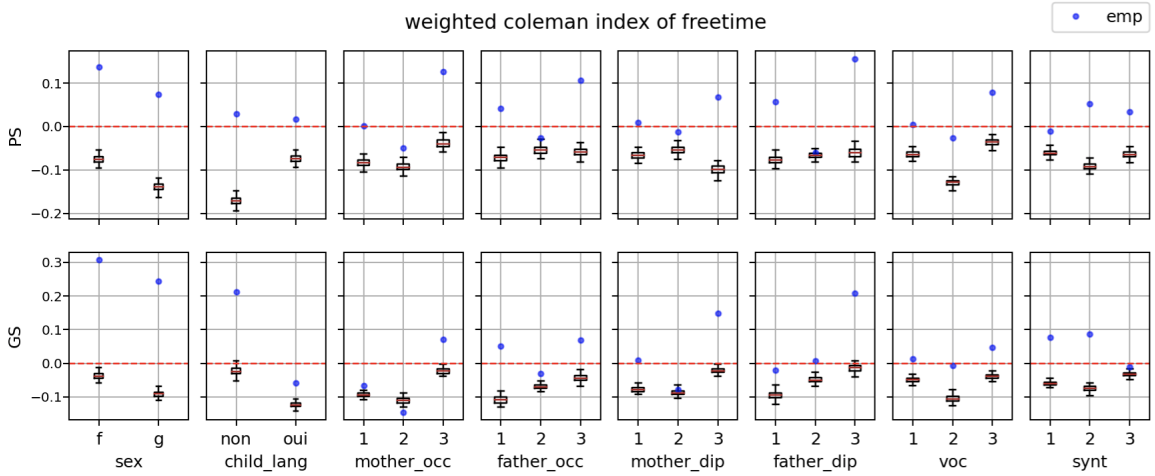

Figure 4.4: Coleman index of freetime, neutral index value (0) is marked as horizontal red dashed line. Blue points are the average index value for all freetime intervals of empirical network. Boxplots are distribution of average index value of 100 instances of link swap reference networks.

The Coleman homophily indices of free-time period are shown in Fig. 4.4. It indicates that many groups show homophilic preferences in our empirical net-

work (blue dots) and these empirical index values significantly deviate from the distribution of reference indices measured in randomised networks, which represents the level of baseline homophily of each group. Just like in case of the EI index, the difference between the Coleman index of randomised and empirical networks tells us the inbreeding homophily level of each group. However, the consideration of the homophily values for the different value groups of each dimension give us more detailed information as compared with EI index. Overall, all groups show heterogeneous (smaller than 0) patterns for randomised network and appear with more homophily (larger index value) in empirical network. This indicates ubiquitous inbreeding homophily among all attributes. In the empirical network, 'sex' showed unquestionable homophilic phenomenon, with girls (f) being more homophilic than boys (g). In addition, most high-valued (3) groups of sociodemographic attributes show homophily effects and they deviate from their randomised homophily value. In contrast, most middle-valued (2) groups in the empirical network show heterogeneity or appear neutral with close values to their randomised homophily average. Nevertheless, the situations for low-valued (1) groups are more irregular without evident trends regarding present homophilic, neutral and heterogeneous patterns. Differently from EI index, when comparing between PS and GS, the homophily of *mother_occ*, *father_occ* show decrease with the grow of age, while homophily of other sociodemographic attributes grow with age. For the linguistic attributes, low and middle level (1 and 2) groups of *voc* are close to their neutral values, while children with well developed vocabulary show homophilic network preferences among themselves. As comparison, most groups in *synt* show homophily which grows with age. To check the difference of empirical and randomised network, we computed the Z-score for every groups as:

$$Z = \frac{Coleman_L - \mathrm{E}[Coleman_L^{rand}]}{\sigma(Coleman_L^{rand})}, \tag{4.4}$$

The results of Z-score are shown in Table 4.4. Results here complement our visual evaluation. Beyond our earlier observations they suggest that linguistic homophily in terms of vocabulary decreases from 1st grade to 3rd grade, while for syntax the contrary is true (except for the value group 3).

In contrast, the Coleman index measured during class-time (see Fig. 4.5) for empirical network show a very different picture. Empirical values in this setting commonly fall close to the range of randomised network and show no clear trends

for any group due to the random assignment of children during class-time, as described in Section 4.2.2. This conclusion is well supported by the corresponding Z-score values summarised in Table 4.4.

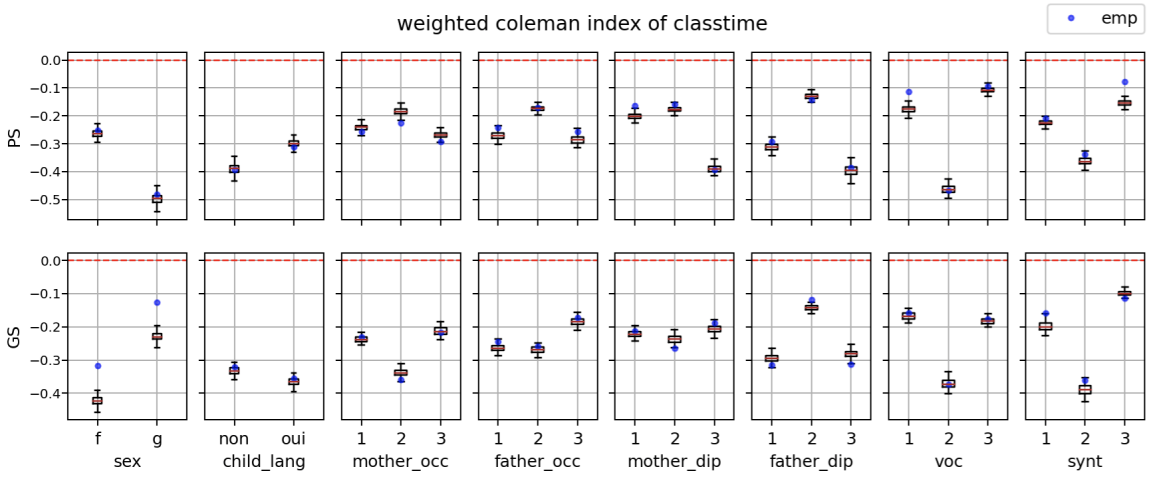

Figure 4.5: Coleman index of classtime, with same groups and notation in freetime.

4.2.4 Homophily analysis with Naive index

Finally, we consider another group homophily index, that was proposed in [Asikainen et al. (2020)] as a more precise measure as compared to the Coleman index. It measures group homophily similar to the Coleman homophily index, but corrects for a disproportionate amount of links potentially observed within large groups as compared with small groups even if there is no intrinsic bias. This feature leads to a different size-correction than the Coleman homophily index as:

$$\text{Naive}_L = \frac{n_O * T_{LL}}{n_L * (1 - T_{LL}) + n_O * T_{LL}}, \tag{4.5}$$

where n_L is the population share of group L and $n_O = 1 - n_L$ is the population share of other groups except for L. T_{LL} has the same meaning as m_{LL} in Coleman index. The range of Naive index is [0,1], indicating heterogeneous to homophilic patterns (respectively).

The Naive homophily indices computed for free-time periods are shown in Fig. 4.6. These results, however, closely resemble the result derived from the Coleman index. As earlier observed, attribute like 'sex' is significantly different from the reference values and they grow with age. Children who speak only French becoming more homophilic with age, while multilingual children getting more

Table 4.4: Z-score of Coleman index in freetime

attr.	sex		child_lang		mother_occ			father_occ			mother_dip			father_dip			voc			synt		
val.	f	g	non	oui	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
PS	22.12	22.68	18.15	10.69	9.97	4.89	15.91	12.39	2.82	16.72	9.93	4.36	16.61	13.48	1.01	18.91	8.90	13.33	14.19	6.14	16.29	11.28
GS	36.87	37.34	19.06	8.52	4.14	-3.75	10.11	14.76	4.68	11.12	11.41	0.93	19.56	6.32	6.22	19.83	8.82	10.52	12.33	20.20	19.67	3.08

Table 4.5: Z-score of Coleman index in classtime

attr.	sex		child_lang		mother_occ			father_occ			mother_dip			father_dip			voc			synt		
val.	f	g	non	oui	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
PS	0.89	0.77	-0.14	-0.99	-1.29	-2.83	-1.82	1.95	0.28	1.81	3.46	1.64	-0.25	1.32	-1.25	0.70	5.16	-0.17	1.64	1.33	1.67	6.71
GS	7.40	7.47	0.83	0.79	0.90	-1.92	-0.29	1.72	1.04	1.03	0.82	-2.37	1.56	-1.72	3.05	-2.49	0.85	-0.08	0.77	3.02	1.88	-2.29

Table 4.6: Z-score of naive index in freetime

attr.	sex		child_lang		mother_occ			father_occ			mother_dip			father_dip			voc			synt		
val.	f	g	non	oui	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
PS	15.58	17.26	14.72	8.33	7.92	4.56	12.98	11.25	4.47	12.97	9.33	4.23	13.42	12.68	1.28	14.04	8.28	10.48	9.47	6.21	11.83	9.31
GS	18.45	23.22	12.42	7.35	4.31	-3.35	9.58	13.06	6.81	11.08	8.91	0.59	13.45	6.16	7.17	11.80	8.84	8.46	10.67	15.40	11.70	4.21

Table 4.7: Z-score of naive index in classtime

attr.	sex		child_lang		mother_occ			father_occ			mother_dip			father_dip			voc			synt		
val.	f	g	non	oui	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
PS	1.42	0.01	-0.45	-0.69	-1.24	-3.03	-1.68	1.62	-0.11	1.94	3.38	1.80	-0.46	1.22	-0.79	0.77	5.29	-0.10	1.64	1.45	1.60	6.93
GS	7.41	7.33	0.52	1.14	1.80	-2.00	-0.15	1.76	0.55	0.92	1.52	-2.19	1.74	-1.53	3.06	-2.39	0.68	-0.12	0.49	2.21	1.81	-2.13

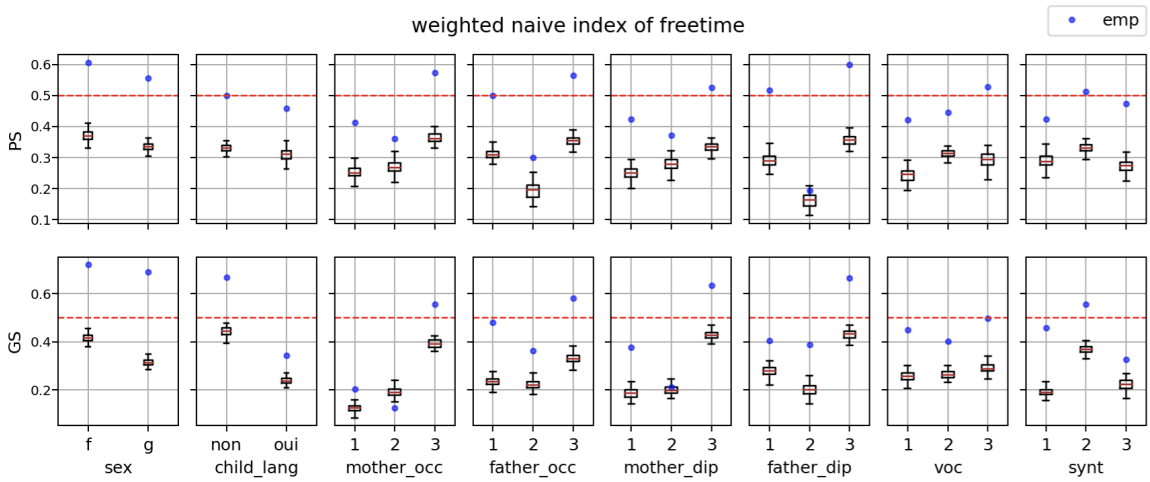

Figure 4.6: Naive index of freetime, with neutral index value (0.5) is marked as horizontal red dashed line.

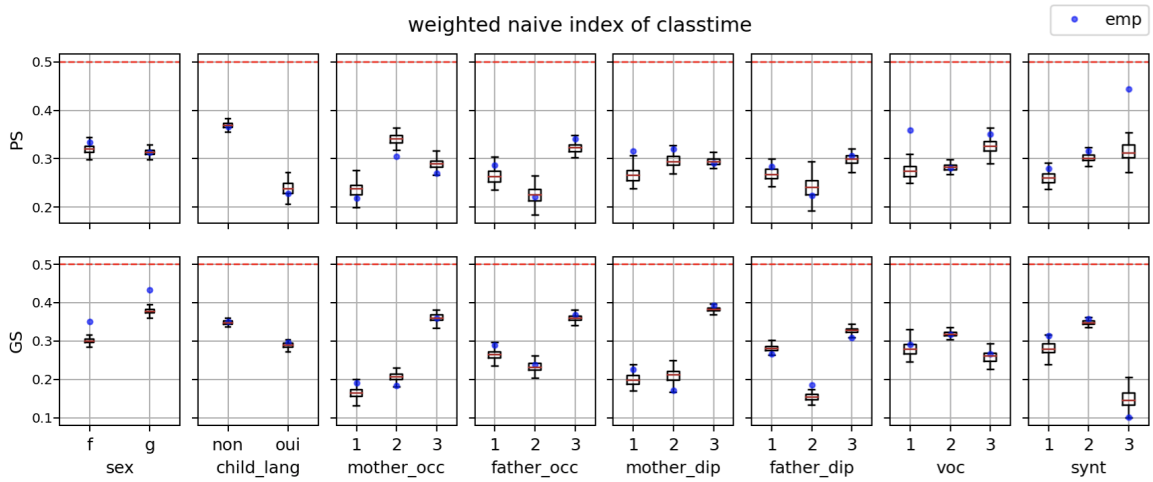

Figure 4.7: Naive index of classtime, with neutral index value (0.5) is marked as horizontal red dashed line.

heterogeneous. Using this index, groups show slightly less homophily as compared with the Coleman index, but still indicate considerable level of inbreeding homophily present. The Naive index in classtime are in Fig. 4.7 together with the corresponding Z-score for free- and class-time are summarised in Table 4.6 and Table 4.7 respectively. Their values reinforce our conclusion we have drawn already from the Coleman index analysis, indicating way less evident homophily patterns in classroom settings due to the the restrictive interaction possibilities.

4.3 Summary

In this chapter we conducted a homophily analysis for aggregated social networks of children collected during free-time and class-time periods. We first came up with a scheme to randomise the network to fully remove structural correlations, while keeping some node properties invariant. Homophily measured in the randomised networks indicate the baseline level of homophily. Homophily observed in empirical networks could be identified as the effects of inbreeding homophily compared to these reference values. We used three different homophily indices on the individual and group level and witnessed strong inbreeding homophily on free-time network with almost every sociodemographic attribute in all cases. On the contrary, no significant inbreeding homophily was observed in class-time networks. In terms of different attributes, the most significant was sex, which exhibits the strongest homophily effects and homophily grows with age. Other attributes behaves were also found important, including linguistic development features of the observed children.

Chapter 5

Visualisation System

Finally, in this last chapter I introduce a data visualisation system, which serves for the purpose of visual analysis of genealogy data. Although this chapter falls somewhat far from the main focus of the previous chapters, it has a relevance in visualising complex attributed network in a way that can be potentially applied for temporal networks in the future. The work summarised in this chapter has been published in [Liu et al. (2017)]. I participated in this work to carry out case studies of migration pattern analysis, as we demonstrate some of them in the end of this chapter.

5.1 Brief description of data and visualisation system

Genealogical datasets provide a great opportunity for sociologists, historians and the public to study a wide variety of topics in demography, family, household, kinship, stratification, and health. Nevertheless, These types of data are usually large scale, hierarchical, record spatio-temporal information and multi-dimensional details thus they pose special challenges for effective data analysis. "*GenealogyVis*" is a visualisation system developed to analyse family history and evolution using the China Multi-Generational Panel Dataset-Liaoning (CMGPD-LN) [Lee et al. (2010), Lee and Campbell (2016)], which has more than 1.5 million observations and provides socioeconomic, demographic and other information for more than 260,000 residents from 698 communities in China.

The design study was conducted with a research group led by a domain expert of humanities & social sciences in an iterative manner over half a year. Several in-depth case studies, involving the research group, are described to demonstrate the usefulness of GenealogyVis and to discuss new findings. GenealogyVis provides

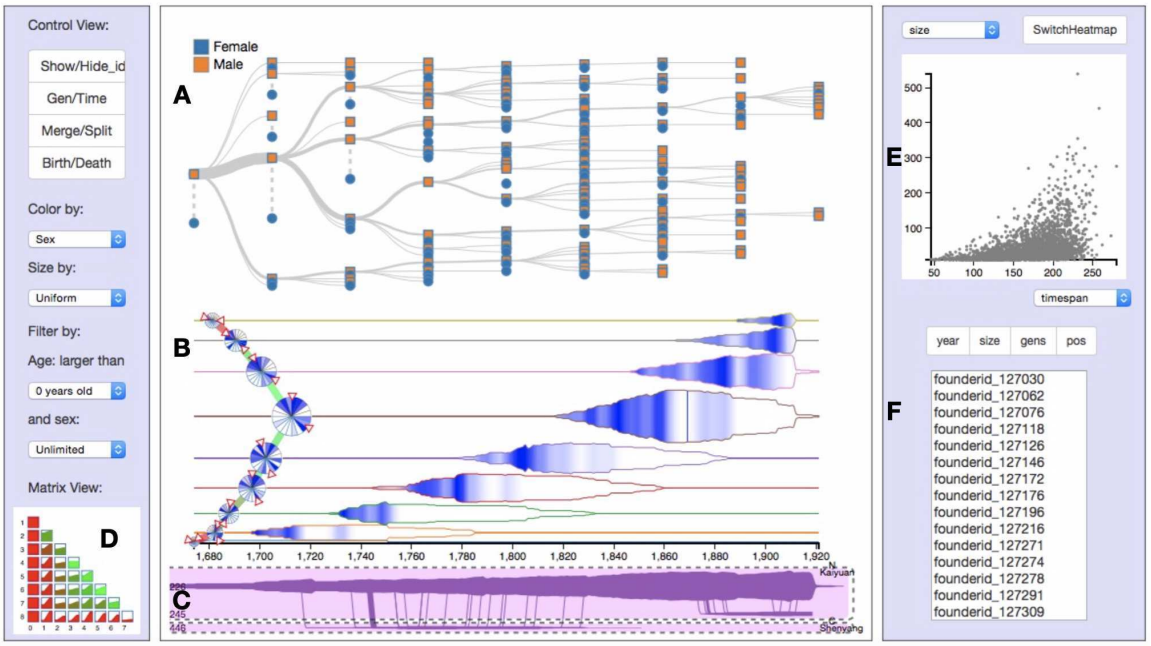

Figure 5.1: Overview of GenealogyVis system.

various perspectives on the CMGPD-LN data through multiple visualisation modules called "views". The system includes a control panel on the left side, and five main linked views:

- Tree View (A) to show the family structure and details of individuals. It focuses on the family tree structure and original data attributes, which can help the users check the details after applying other visualisation views to solve the analytic tasks from the high level
- Stream View (B) to show various statistical information such as population, birth and mortality rate. It depicts various statistical information such as demographic information and temporal information, which combined with Tree View to present demographic characteristic and evolution pattern.
- Migration View (C) to present the genealogical migratory behaviours.
- Matrix View (D) to analyse the reproduction pattern between two generations.
- Scatterplot View (E) to provide overview of the data and further explore the correlation analysis.
- List View (F) to order the families by different attributes.

Tree View for structure and original Data

The genogram is a conventional way to present family trees, which is very useful for people to research the genetic and interpersonal family-household data. In this system we use square and circle to represent male and female respectively. The CMGPD-LN data, due to patrilineality in that time, emphasise the males with records and most of the patterns related to males. Therefore the design in GenealogyVis links the child nodes to their father nodes instead of the middle position of the parent nodes. In the design, all males of the same generation are ordered top-down by the birth year. Each female node is placed under its husband node.

Stream View for demographic and temporal information

The Stream View present the demographic information of each generation or the whole family. Each divided stream shows the population changing over time of the corresponding generation, where the width of the stream encodes the number of people alive in the current year. Thus, users can know the changing trend of population in each generation. Streams could be merged to show the total population changing over time by *Merge/Split* button in the Control Viewy. Those streams are filled with different opacity of colour, which represents the number of births or deaths.

Migration View for genealogical migratory behaviours

For a family in pre-industrial period, the migration is usually a slow process that may last for several decades, and its scale is limited with short distance. Traditional visualisation scheme of migration usually label the itinerary on a map, which is bothersome to encoding the temporal information, for example to know the time and scale of migration. To solve aforementioned problems, a storyline-like visualisation is proposed to show the process of migration, which eliminates the overplotting that would occur on geographic maps. We use dotted-border rectangle to represent a region, and within which the colour area represents a district. Each stream in the colour area represents the population of this family living in the village. Migration is represented by links between the corresponding streams, with link's horizontal positions represent the time and width encodes the size. Moreover, the links is designed to be slant, with horizontal position from left to right to show the starting village to destination.

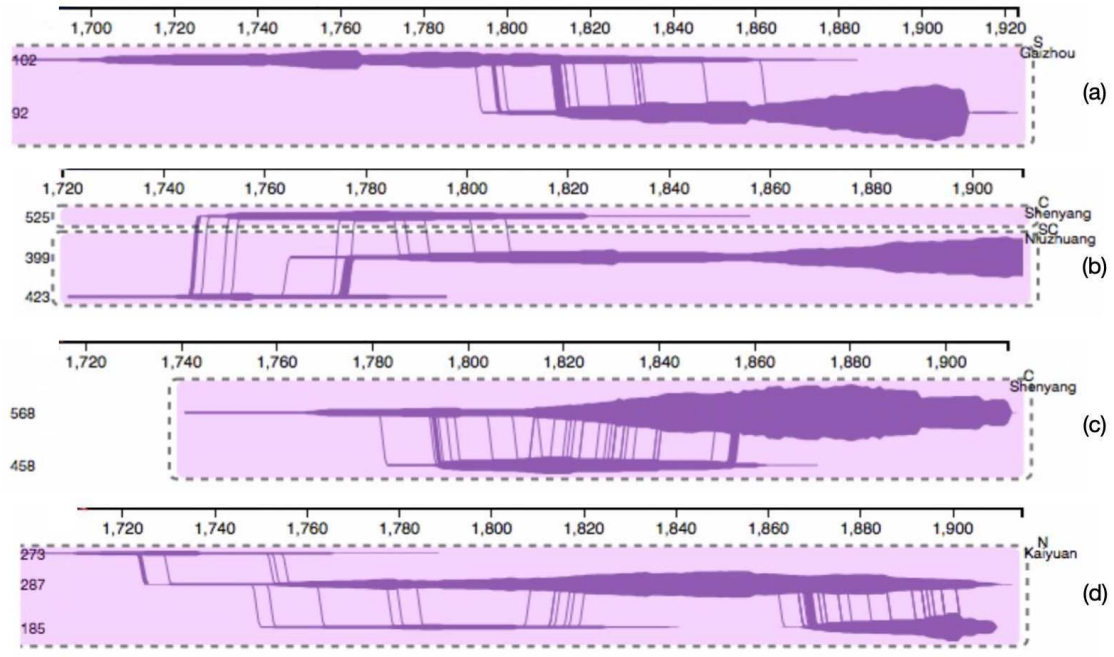

Figure 5.2: Patterns of migration.

5.2 Case Study of Family Migration Patterns

Migration View is particularly useful in discovering migration pattern. In Fig. 5.2 we show migration patterns of four different, only for demonstration purposes. Fig. 5.2a show a relatively common one-way migration, moving from the ancestral home to other villages. The figure shows that the family initially lived in village 102, then family members moved to village 92, where the family settled down and population increased through reproduction, at the same time the population of this family in village 102 gradually demised. In contrast, the family shown in Fig. 5.2b moved to two different villages. Some family members moved farther away to village 525 which located in another district Shenyang. While some other members moved to a closer village 399 which was still the same district. The group of village 525 and original village 423 gradually disappeared and the group moved to a closer village 399 thrived.

The family shown in Fig. 5.2c showed the pattern of reverted migration. Parts of the family in village 568 migrated to village 458 in 1790s. But starting from 1810s the same group of people started moving back to ancestral village 568, with peak in 1850s, left with only limited family member. In the end the bloodline of this family faded in village 568. Fig. 5.2d showed a complementary situation. A family from village 273 first moved to village 287, then several member again moved to

village 185, whom entirely moved back to village 287 and no more member stayed there. However after 50 years later, when the family again chose to migrate, they choose the same village 185, following the path of their predecessors.

5.3 Summary

In this chapter, I gave a brief introduction of visualisation system GenealogyVis, with slightly detailed description of three most informative parts and designs and in the end, we discussed some cases of migration patterns of families using this visualisation tool. The way of processing, organising and presenting data with such rich attributes is inspiring and with prospective of porting to other dataset with temporal feature such as DyLNet. This system is made possible by the collaboration with team from Hong Kong University of Science and Technology.

Chapter 6

Conclusion and outlook

My main goal in this Thesis was to summarise my contributions to the design, collection, and analysis of a large-scale behavioural dataset that we gathered in a unique social experiment involving hundreds of pre-school children and their teachers over several years. This experiment, one of the largest of its kind, was carried out to follow the social and oral interactions of children to understand how socialisation, influence, and language acquisition co-evolves in children groups in a pre-school setting.

This data collection, reconstruction, and analysis, that was in the main focus of my thesis, were carried out with wearable wireless RFID tags. These autonomous devices could detect the distance and orientation of proxy individuals and they recorded their voice too. They provided a massive amount of sequential data, that could be used for the reconstruction of real social and oral interaction with high spatial and temporal resolution

I started my Thesis with an introduction to the research landscape of Human Dynamics, where most of my works landed. I did it in relation with the corresponding fields of Network Science, Machine Learning and Digital Data Collection Methods. Subsequently, in Chapter 2, I introduced in details the goals of the DyLNet project, its empirical setting, the data collection processes, and most importantly how I arrived from the recorded raw RFID signals to a meaningful dataset that was used for further analysis. In Chapter 3, I summarised methods and pipelines I developed for two purposes: One of my goal was to precisely reconstruct meaningful temporal interactions between pairs of participants. This highly non-trivial task required to adapt a pipeline of classification methods and sequential learning models, but in turn provided a temporal network verified against several expected network and dynamical characters. My other goal in this Chapter was to describe a method I developed to reconstruct free and class time

periods, which correspond to significantly different activities and behavioural patterns of children in a school. I developed algorithmic methods to solve this issue and demonstrated their capacities to precisely reconstruct different activity periods simultaneously for several school classes, that I verified against the ground truth data. In Chapter 4, I combined the reconstructed temporal networks with individual level socio-demographic and linguistic data, which was collected in parallel using offline questionnaires. This combined dataset allowed for a multivariable data analysis to understand, which dimensions of homophily (tendency of similar people to be socially connected) are determinant for the formations of social ties between children and staff. I carried out this homophily analysis at the individual and group level by comparing children from different age groups. This analysis provided evidences about the effects of gender, demographic and linguistic homophily and their effects in socially restricted (class-time) and non-restricted (free-time) settings. In Chapter 5, I summarised some of my earlier contributions to develop a visualisation tool for genealogy trees. Finally I closed my Thesis with some concluding words.

There are several promising research directions I can foresee in the future using the presented datasets. Topics from social group dynamics, co-development of language and social networks, or the data-driven modelling of epidemic processes in a school setting are only a few. Nevertheless, the work presented in this Thesis marks a milestone in this process, as it demonstrates a complete methodological circle from experimental design, data collection, reconstruction, analysis and the achievement of new fundamental results. This way it provides an example how cutting-edge digital technologies and methods borrowed from Computer Science can be used to learn about the behaviour of people to answer long-lasting questions in remote fields.

Appendix A

Supplementary computation results and experimental details

A.1 Experimental design of ground truth data collection

While designing the ground truth data collection, special attention has been paid to the feasibility and reliability of our observation method, to decrease human errors during observations while obtaining a meaningful ground truth dataset. To meet with all these requirements we followed the following logic and conditions:

- To record the ground truth data for GT1 and GT3 we had a researcher in the classroom who was monitoring the behaviour of children to record their interaction state, their relative orientation or position at regular intervals. To quantify interactions, distance and orientation, we chose a *scan sampling* strategy (i.e. observations of states at predetermined time intervals) [Altmann (1974)]. Another option was *focal sampling* (i.e. continuous recording of interaction events) [Altmann (1974)], however social interactions (and kids position/distance too) were sometimes too short and fluctuating for continuous observation, thus it was impossible to record the beginning and end of each interacting event without using video-recording.

  In addition, for scan sampling, we made trials to find the appropriate time intervals between two observations (scans) that allowed us to record data without loss (i.e. the shortest step possible for recording the data of interest without taking the risk to miss one observation point). In the end, we chose 10 second steps for pair observations (GT1) and 2 minute steps for group observations (GT3).

- For GT1 observations, we worked during free play time to be able to observe spontaneous play interactions. To decrease the possible noise due to fleeting behaviours and random moves, observations were carried out on older children from the middle (4-5 years old) or the grand class (5-6 years old). Importantly, we focused only on one pair of kids at a time to record their state (interacting or not) and relative position every 10 seconds. This way we reduced to the minimum the possible human observation errors in this setting.

  Regarding the scoring of the state of interaction/no-interaction, we set the criteria of interactions based on the literature [Santos et al. (2008)] and the field expertise of the participating researchers from earlier similar experiments. Specifically, we considered two children "interacting" if they were within arm's reach (i.e. less than 1 meter from each other), either playing together (e.g. cooperatively manipulating construction blocks, kitchen toys, puzzles...) or playing alongside (e.g. making a drawing next to each other). These situations typically involved talking to each other at times.

- Regarding GT3, we selected the most appropriate and stable conditions to observe distance and orientation of as many children as possible. In practice, we performed observations only with older children (grand class) when their movements inside the classroom were limited for an extended period of time. This was possible during specific activities that involved standing still (like collectively sitting on a bench to listen to the teacher reading a book, sitting around tables in small groups to do written work...) rather than freely moving around (like during free play time).

- Measuring distances between children is very much subject to inter- (and even intra-) individual variations. To minimise such fluctuations in recording the distances in GT3, we used a customised behavioural observation app originally developed to record the positions and distances of animals in a fixed environment (Animal Observer application for iPad [Ani]). This app projects a scaled map of the classroom with indicated reference objects (furnitures, doors, windows, etc.) and allows to record by an observer the positions of individuals on this map as the function of time. Using this temporal location dataset inter-individual distances of children have been computed afterward. Relative to the objects on the classroom map the location of children could be estimated precisely with a very small error margin (around 10cm),

something that would have been impossible to accurately assess through “naked eye" estimations and “hand" recording.

To further check the impact of potential errors due to human encoding, we conducted a randomisation experiment where we induced noise in the already collected data. More precisely, we randomly selected the 5%, 10%, 15%, and 20% of observation points from the ground truth data sequences and flipped the annotated flags from interaction to non-interaction or vice versa to add noise to our original observations. Through remeasuring the accuracy change of the BiLSTM-RSSI method on these randomised data we found that the average and variance of accuracy is rather robust against such small induced noise, only having the average to decrease slightly as summarised in Table A.1.

Table A.1: Average and standard deviations of BiLSTM-RSSI reconstruction method trained on GT1 after introducing random noise of various levels. Average values are computed over 20 independent realisations.

Flipped random fraction	0.05	0.1	0.15	0.2
Average accuracy	90.02%	89.98%	89.91%	89.84%
Standard deviation	9.14%	9.24%	9.53%	9.26%

A.2 Parameters and performance of reconstruction methods

In this appendix, we summarise the parametrisation of the different reconstruction methods together with the confusion matrices corresponding to their best performing parameters, as summarised in the Section 3.1.2.3 in the main text.

A.2.1 Classification model choice

When choosing method to classify the handshake pairs, we have tested two methods: Logistic regression (LR) and Support vector machine (SVM). We chose LR because it achieved slightly better accuracy compared with SVM, also considering the data is a simple 3-dimension vector, LR could handle it well enough, although SVM could better avoid overfitting. The decision boundary of two models on our ground truth data is shown in Fig. A.1

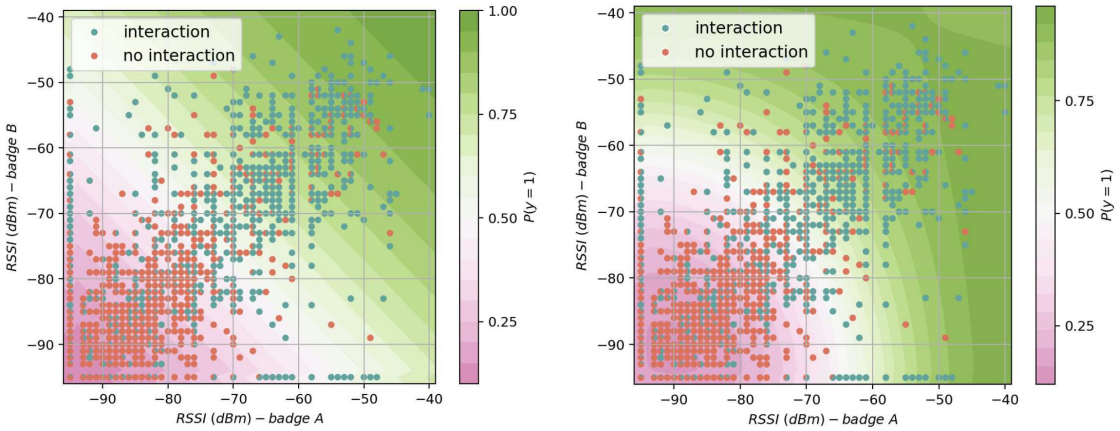

Figure A.1: Decision boundary of RL and SVM

A.2.2 Naïve method

The naïve method has a single parameter gap, which determines the maximum length of non-interaction gaps between two interaction events to be filled automatically with interaction states. $gap = 0$ is a special case as it belongs to the non-reconstructed signal where no state has been filled. We explored gap from 0 to 9, corresponding from 0 to 45 seconds of non-interaction gaps, with 5 seconds incremental step size. We found that in our setting the best reconstruction can be reached with $gap = 6$ with accuracy reaching 0.834 as summarised in the confusion matrix in Table A.2. Consequently, if a longer than 30 second gap appears in the interaction sequence of two individuals, the two participants most probably broke their actual social interaction, thus events before and after the gap should be considered separately.

Table A.2: Confusion matrix with accuracy of the naïve event reconstruction method with $gap = 6$.

acc.=0.8336	contact	no-contact
contact	0.8942	0.1058
no-contact	0.2361	0.7639

A.2.3 Hidden Markov model

When parametrizing the HMM with annotated data, we used maximum likelihood estimation to compute three matrices. Take transition matrix for example, if the frequency of hidden state i at t transiting to hidden state j at $t + 1$ is A_{ij}, then the

estimated transition probability $\hat{a}_{ij}$ is computed as follows: $\hat{a}_{ij} = A_{ij} / \sum_{j=0}^{N-1} A_{ij}$, where N is the number of hidden states. Same method is applied for computing emission matrix and initial matrix.

With embedded envelop sequence, we first pad 0 (indicating non-interaction states) at the beginning of each sequence, with size of $windowsize - 1$. We then use transformed envelop signals instead of binary signals to define the hidden states, observation states, as well as determining all the matrices. Finally, as an output of the Viterbi algorithm we obtain a sequence of envelop, the last item of each envelop being the predicted interaction/non-interaction state of each time step.

Table A.3: Confusion matrix with accuracy of the HMM model with window size $win = 6$.

acc.=0.8425	contact	no-contact
contact	0.8917	0.1083
no-contact	0.2176	0.7824

The reconstruction accuracy and the confusion matrices of the HMM methods are shown in Tables A.3 for window size $win = 6$.

A.2.4 Bi-directional LSTM methods

For each BiLSTM method we used an envelop with size of $winsize$ located symmetrically on the middle state which we wanted to reconstruct (as demonstrated in Fig. 3.4e). We pad void signals at the beginning and at the end of each sequence, with size of $\lfloor winsize/2 \rfloor$ on each side. More precisely, the padded void signal for BiLSTM-RSSI is a vector $(-95, -95, 0)$, for BiLSTM-logi is a vector $(1, 0)$ and for BiLSTM-bin is a single number 0.

We merged the outputs of the two LSTMs using concatenation, which provided double size of outputs to the next layer. For the training, we split our labelled data into 10 clips with each around 25 mins then use nested cross validation to select best hyper-parameters and examine the performance of each reconstruction method. The confusion matrix of the three BiLSTM reconstruction tasks are shown in Table A.4, Table A.5 and Table A.6. The accuracy of the BiLSTM-RSSI reached 0.9003, which is the best among all the tested methods.

Table A.4: Confusion matrix with accuracy of the BiLSTM-RSSI for event reconstruction.

acc.=0.9003	contact	no-contact
contact	0.9165	0.0084
no-contact	0.1226	0.8774

Table A.5: Confusion matrix with accuracy of the BiLSTM-logi for event reconstruction.

acc.=0.8902	contact	no-contact
contact	0.9067	0.0933
no-contact	0.1333	0.8667

Table A.6: Confusion matrix with accuracy of the BiLSTM-bin for event reconstruction.

acc.=0.8834	contact	no-contact
contact	0.8919	0.1081
no-contact	0.1294	0.8706

A.3 Linguistic data validation

In both the lexical (vocabulary) and syntactic skills (syntax) tests, we used two types of items. Ten identical so-called 'anchor items' were presented to children whichever their grade, every time they took a test (at every period of testing all along the project). These anchor items had been meticulously chosen to allow the researchers to evaluate linguistic skills development over time. Results shown in Fig. A.2a confirm that subjects' scores indeed increase with age. Besides, different so-called 'test items' were used in the four different versions of the tests adapted to the subject's grade, namely 30 items in the vocabulary test and 10 in the syntax test. These test items had been carefully selected to allow to evaluate the level of linguistic abilities in children of similar school level. Results presented in Fig. A.2b show that scores for these items are, as expected, centered around the average ($\approx$15/30 for vocabulary, $\approx$5/10 for syntax). Moreover, the scores within each group (i.e. each test version) are widespread, thus meeting the requirements for achieving the goal to distinguish between high-, medium- and low-skilled pupils. These results attest to the relevance of the linguistic tests we designed specifically to achieve our objectives, which were two-fold: first to assess the improvement of linguistic skills in pupils over time, and second to measure the heterogeneity in language ability

levels within same-grade groups.

In addition, the memory span test was used as a control measure, given that linguistic skills are known to be constrained by working memory capacities in developing children Blake et al. (1994). Children were asked to repeat the exact same series of digits whichever their grade, allowing the researchers to evaluate memory skills improvement in subjects over time. Subjects' scores indeed increase with age (Fig. A.2c). More interestingly, the fact that language test scores correlate with memory span scoring (as shown in Fig. A.2d) is an additional way of validating the relevance of the linguistic tests we designed specifically to achieve our research goals.

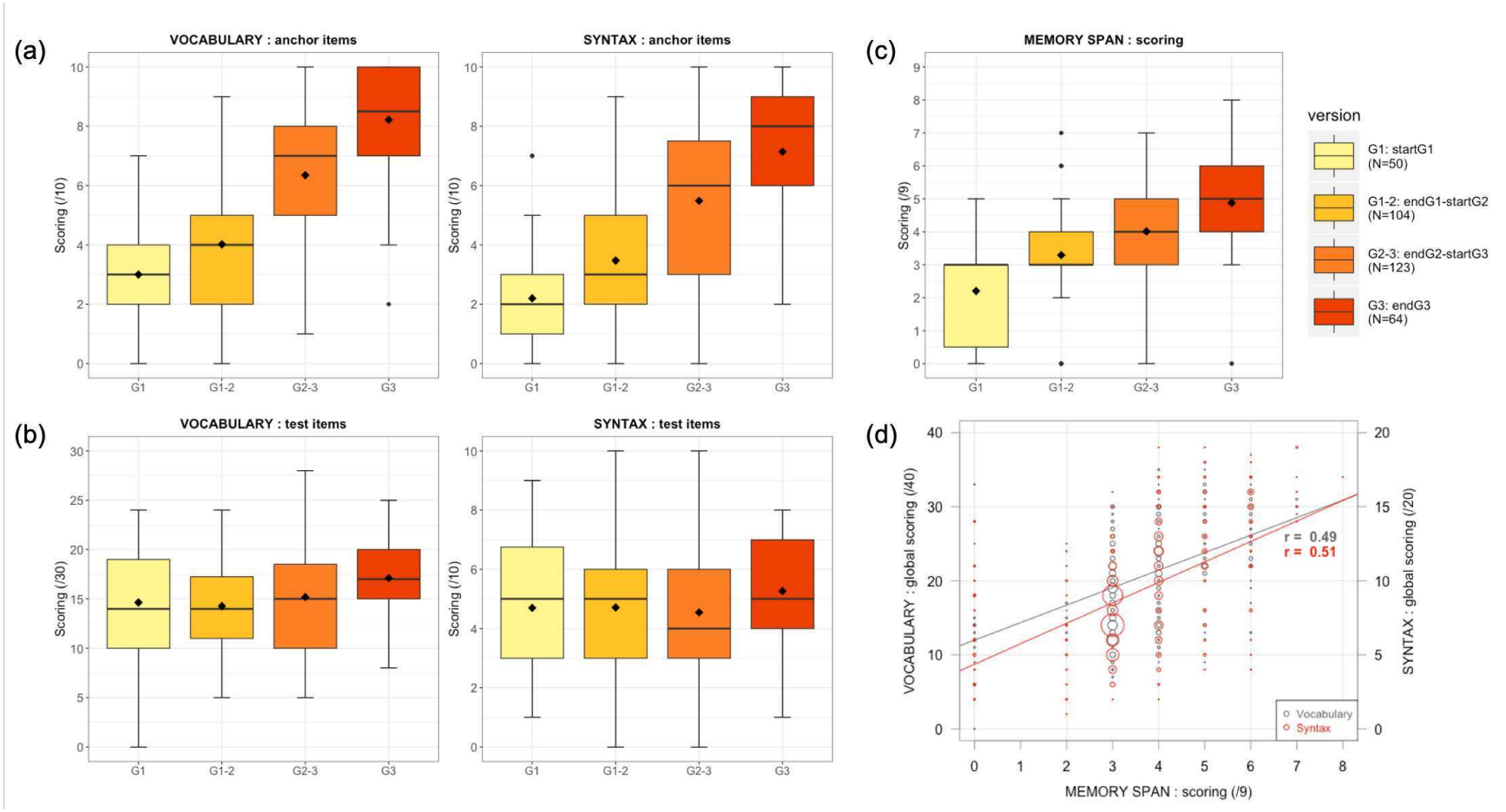

Figure A.2: Language test results, shown as box-plots, for children of different grades tested at the beginning and end of the academic year during which transactional data collection took place. Four versions of the tests were designed to be adapted to subjects' grade (G1: *startG1* version of the tests designed for children entering 1st grade, G1-2: *endG1-startG2* version for children completing 1st grade or entering 2nd grade, G2-3: *endG2-startG3* version for children completing 2nd grade or entering 3rd grade, G3: *endG3* version for children completing 3rd grade). For both vocabulary and syntax tests, children were presented with two types of items: (a) anchor items, shared across versions, and (b) test items, adapted to the subject's grade. (c) Scoring for the memory span test during which identical series of digits to repeat were used whichever the subject's grade. In every box-plot, black diamond indicates the average value and bar shows the median value. (d) Correlation between memory span and performance in linguistic tasks for children of all school levels combined. Circle size is proportional to the number of data points at each (x,y) coordinates. Each of the six plots was drawn from 341 points originating from 174 children tested twice in the academic year (or only once for just 7 of them who left or entered the school over the course of the year).

Appendix B

DyLNet supplementary information

Ethics

Experiments were approved by relevant ethics committees, COERLE (*Comité Opér ationnel d'Evaluation des Risques Légaux et Ethiques* of INRIA institute, favourable opinion no. 2017-014) and CNIL (*Commission Nationale de l'Informatique et des Libertés*, favourable opinion Avis CIL_UGA-2017_0980683). The study was conducted within a French primary school with permission from the relevant authorities of *Education Nationale*. Written informed consent was obtained from all adults of the teaching staff and from the parents of all the children who took part in the study.

Usage Notes

Note that the data presented in this paper were recorded in 174 children as study subjects, however observations of only 164 children are shared as parents of 10 children opted out to be included in data shared with researchers outside of the project team. In addition, data about 32 adult study subjects are shared. Furthermore, as the data contain sensitive information on human subjects, it cannot be shared fully openly and used freely. Access to the data can be granted by the Principal Investigator (Aurélie Nardy - corresponding author), after submission of a short research proposal, via email, on the planned use of the recorded data. Access to the data is conditional to the prior signature of the Data Access Agreement. The data can be exclusively used for scientific purposes.

Code availability

The program codes for data cleaning and temporal network reconstruction are shared along the dataset in an open repository. The codes have been developed in Python language using only standard or open licensed packages, and they are shared as iPython notebooks at Nardy et al. (2022).

Acknowledgments

We are especially grateful to all the children, their family, and the members of the school teaching staff who took part in our experiment. We thank Christophe Braillon and Jean-François Cuniberto from INRIA Rhône-Alpes for their support and technical help during the data collection. We are grateful to Isabelle Rousset, Laurence Buson, Sophie Loiseau and Emmanuelle Roman for their help in designing the language tests and in conducting the language surveys on site. We are also thankful to Sophie Loiseau, Lisa Paillussière, Beatriz Villa and Charlotte Chauvet for their assistance in entering the socio-demographic or linguistic surveys data into the database. This work has been supported by the DyLNet ANR project (ANR-16-CE28-0013). MK acknowledges support from the DataRedux ANR project (ANR-19-CE46-0008) and the SoBigData++ (H2020-871042) project. SD benefits from sponsorship by China Scholarship Council (CSC) for his PhD program.

Bibliography

Animal Observer: An iPad app designed to collect animal behavior and health data. https://fosseyfund.github.io/AOToolBox/.

Dylnet. https://dylnet.univ-grenoble-alpes.fr/?language=en. Accessed: 2019-06-26.

Openbeacon. https://www.openbeacon.org. Accessed: 2019-06-26.

R. Aarts, S. Demir, and T. Vallen. Characteristics of Academic Language Register Occurring in Caretaker-Child Interaction: Development and Validation of a Coding Scheme. *Language Learning*, 61(4):1173–1221, 2011. ISSN 1467-9922. doi: 10.1111/j.1467-9922.2011.00664.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9922.2011.00664.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9922.2011.00664.x.

L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer. Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)*, 6(2):1–33, 2012.

R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the world-wide web. *nature*, 401(6749):130–131, 1999.

E. Alpaydin. *Introduction to machine learning*. MIT press, 2020.

J. Altmann. Observational study of behavior sampling methods. *Behavior*, 49:227–267, 1974.

R. M. Alvarez and J. Nagler. A new approach for modelling strategic voting in multiparty elections. *British Journal of Political Science*, 30(1):57–75, 2000.

S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

A. Asikainen, G. Iñiguez, J. Ureña-Carrión, K. Kaski, and M. Kivelä. Cumulative effects of triadic closure and homophily in social networks. *Science Advances*, 6 (19):eaax7310, 2020.

A. Badie-Modiri, M. Karsai, and M. Kivelä. Efficient limited-time reachability estimation in temporal networks. *Physical Review E*, 101(5):052303, 2020.

Y. Bar-Yam. General features of complex systems. *Encyclopedia of Life Support Systems (EOLSS), UNESCO, EOLSS Publishers, Oxford, UK*, 1, 2002.

A. Barabási and M. Pósfai. *Network Science*. Cambridge University Press, 2016.

A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

A. Barrat and C. Cattuto. Temporal networks of face-to-face human interactions. In *Temporal Networks*, pages 191–216. Springer, 2013.

A. Barrat and M. Weigt. On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13(3):547–560, 2000.

A. Barrat, M. Barthelemy, and A. Vespignani. *Dynamical processes on complex networks*. Cambridge university press, Cambridge, UK, 2008.

A. Barrat, C. Cattuto, M. Kivelä, S. Lehmann, and J. Saramäki. Effect of manual and digital contact tracing on covid-19 outbreaks: a study on empirical contact data. *Journal of the Royal Society Interface*, 18(178):20201000, 2021.

M. Barthélémy and L. A. N. Amaral. Small-world networks: Evidence for a crossover picture. *Physical review letters*, 82(15):3180–3183, 1999.

M. Barthélemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani. Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. *Physical review letters*, 92(17):178701, 2004.

D. S. Bassett and O. Sporns. Network neuroscience. *Nature neuroscience*, 20(3): 353–364, 2017.

D. S. Bassett, M. Yang, N. F. Wymbs, and S. T. Grafton. Learning-induced autonomy of sensorimotor systems. *Nature neuroscience*, 18(5):744–751, 2015.

L. E. Baum and J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(3):360–363, 1967.

L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.

L. E. Baum et al. An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3(1):1–8, 1972.

A. Bazzani, B. Giorgini, S. Rambaldi, R. Gallotti, and L. Giovannini. Statistical laws in urban mobility from microscopic gps data in the area of florence. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(05):P05001, 2010.

P. Bedi and C. Sharma. Community detection in social networks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 6(3):115–135, 2016.

U. S. Bhalla and R. Iyengar. Emergent properties of networks of biological signaling pathways. *Science*, 283(5400):381–387, 1999.

G. Bianconi and A.-L. Barabási. Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)*, 54(4):436, 2001.

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

J. Blake, W. Austin, M. Cannon, A. Lisus, and A. Vaughan. The relationship between memory span and measures of imitative and spontaneous language complexity in preschool children. *International Journal of Behavioral Development*, 17 (1):91–107, 1994. doi: 10.1177/016502549401700106.

P. M. Blau and J. E. Schwartz. *Crosscutting social circles: Testing a macrostructural theory of intergroup relations*. Routledge, 2018.

J. Blumenstock, G. Cadamuro, and R. On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

E. P. Bowers and M. Vasilyeva. The relation between teacher input and lexical growth of preschoolers. *Applied Psycholinguistics*, 32(01):221–241, Jan. 2011. ISSN 0142-7164, 1469-1817. doi: 10.1017/S0142716410000354. URL `http://www.journals.cambridge.org/abstract_S0142716410000354`.

J. Budd, B. S. Miller, E. M. Manning, V. Lampos, M. Zhuang, M. Edelstein, G. Rees, V. C. Emery, M. M. Stevens, N. Keegan, et al. Digital technologies in the public-health response to covid-19. *Nature medicine*, 26(8):1183–1192, 2020.

E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3):186–198, 2009.

L. Cameron and D. Larsen-Freeman. Complex systems and applied linguistics. *International journal of applied linguistics*, 17(2):226–239, 2007.

J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of physics A: mathematical and theoretical*, 41(22):224015, 2008.

G. Carra, I. Mulalic, M. Fosgerau, and M. Barthelemy. Modelling the relation between income and commuting distance. *Journal of the Royal Society Interface*, 13 (119):20160306, 2016.

C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, and A. Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PloS one*, 5(7):e11596, 2010.

G. Cencetti, G. Santin, A. Longa, E. Pigani, A. Barrat, C. Cattuto, S. Lehmann, M. Salathe, and B. Lepri. Digital proximity tracing on empirical contact networks for pandemic control. *Nature communications*, 12(1):1–12, 2021.

B. Chowdhury, M. U. Chowdhury, and N. Sultana. Real-time early infectious outbreak detection systems using emerging technologies. In *2009 International Conference on Advances in Recent Technologies in Communication and Computing*, pages 506–508. IEEE, 2009.

P. Chunaev. Community detection in node-attributed social networks: a survey. *Computer Science Review*, 37:100286, 2020.

F. Chung and L. Lu. The diameter of sparse random graphs. *Advances in Applied Mathematics*, 26(4):257–279, 2001.

J. E. Cohen, F. Briand, and C. M. Newman. *Community food webs: data and theory*, volume 20. Springer Science & Business Media, 2012.

R. Cohen and S. Havlin. Scale-free networks are ultrasmall. *Physical review letters*, 90(5):058701, 2003.

J. S. Coleman. Relational analysis: The study of social organizations with survey methods. *Human organization*, 17(4):28–36, 1958.

V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences*, 103(7):2015–2020, 2006.

G. Conti-Ramsden and K. Durkin. Language Development and Assessment in the Preschool Period. *Neuropsychology Review*, 22(4):384–401, Dec. 2012. ISSN 1573-6660. doi: 10.1007/s11065-012-9208-z. URL `https://doi.org/10.1007/s11065-012-9208-z`.

D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

S. Currarini, M. O. Jackson, and P. Pin. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045, 2009.

S. Currarini, M. O. Jackson, and P. Pin. Identifying the roles of choice and chance in network formation: Racial biases in high school friendships. *Proceedings of the National Academy of Sciences*, 107:4857–4861, 2010.

S. Dai, H. Bouchet, A. Nardy, E. Fleury, J.-P. Chevrot, and M. Karsai. Temporal social network reconstruction using wireless proximity sensors: model selection and consequences. *EPJ Data Science*, 9(1):19, 2020.

S. Dai, H. Bouchet, M. Karsai, J.-P. Chevrot, and A. Nardy. Longitudinal data collection to follow social network and language development dynamics at preschool, 2022. submitted.

J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. Complex networks in climate dynamics. *The European Physical Journal Special Topics*, 174(1):157–179, 2009.

S. Duncan and D. W. Fiske. *Face-to-face interaction: Research, methods, and theory*. Routledge, World wide, 2015.

A. Duval, T. Obadia, L. Martinet, P.-Y. Boëlle, E. Fleury, D. Guillemot, L. Opatowski, and L. Temime. Measuring dynamic social contacts in a rehabilitation hospital: effect of wards, patient and staff characteristics. *Scientific Reports*, 8(1): 1–11, 2018.

H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Physical review E*, 66(3):035103, 2002.

S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 467–474, 2015.

J.-P. Eckmann, E. Moses, and D. Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences*, 101(40): 14333–14337, 2004.

W. J. Edmunds, C. O'callaghan, and D. Nokes. Who mixes with whom? a method to determine the contact patterns of adults that may lead to the spread of airborne infections. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 264(1384):949–957, 1997.

T. Elmer, K. Chaitanya, P. Purwar, and C. Stadtfeld. The validity of rfid badges measuring face-to-face interactions. *Behavior Research Methods*, page 1–19, 2018.

T. Elmer, K. Chaitanya, P. Purwar, and C. Stadtfeld. The validity of rfid badges measuring face-to-face interactions. *Behavior research methods*, 51(5):2120–2138, 2019.

P. Erdos, A. Rényi, et al. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

J. Ernst and M. Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–216, 2012.

D. Felmlee, S. Sprecher, and E. Bassin. The dissolution of intimate relationships: A hazard model. *Social Psychology Quarterly*, pages 13–30, 1990.

L. Festinger. *A theory of cognitive dissonance*, volume 2. Stanford university press, 1957.

L. R. Ford and D. R. Fulkerson. Maximal flow through a network. *Canadian journal of Mathematics*, 8:399–404, 1956.

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

J. Fournet and A. Barrat. Contact patterns among high school students. *PloS one*, 9 (9):e107878, 2014.

A. Gautreau, A. Barrat, and M. Barthelemy. Global disease spread: statistics and estimation of arrival times. *Journal of theoretical biology*, 251(3):509–522, 2008.

L. Gauvin, M. Génois, M. Karsai, M. Kivelä, T. Takaguchi, E. Valdano, and C. L. Vestergaard. Randomized reference models for temporal networks. *arXiv preprint arXiv:1806.04032*, 2018.

H. Giles and A. C. Billings. Assessing Language Attitudes: Speaker Evaluation Studies. In A. Davies and C. Elder, editors, *The Handbook of Applied Linguistics*, pages 187–209. Blackwell Publishing Ltd, 2004. ISBN 978-0-470-75700-0. URL `http://onlinelibrary.wiley.com/doi/10.1002/9780470757000.ch7/summary`.

A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(3):14–es, 2007.

M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

E. Goffman. *Interaction ritual: Essays in face-to-face behavior*. Routledge, World wide, 2017.

K.-I. Goh and A.-L. Barabási. Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 81(4):48002, 2008.

M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779–782, 2008.

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

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

K. H. Grantz, H. R. Meredith, D. A. Cummings, C. J. E. Metcalf, B. T. Grenfell, J. R. Giles, S. Mehta, S. Solomon, A. Labrique, N. Kishore, et al. The use of mobile phone data to inform analysis of covid-19 pandemic epidemiology. *Nature communications*, 11(1):1–8, 2020.

A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013.

F. A. Haight. Handbook of the poisson distribution. Technical report, 1967.

S. Hanhijärvi, G. C. Garriga, and K. Puolamäki. Randomization techniques for graphs. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 780–791. SIAM, 2009.

D. Helbing. Traffic and related self-driven many-particle systems. *Reviews of modern physics*, 73(4):1067, 2001.

G. T. Henry and D. K. Rickman. Do peers influence children's skill development in preschool? *Economics of Education Review*, 26(1):100–112, Feb. 2007. ISSN 02727757. doi: 10.1016/j.econedurev.2005.09.006. URL `http://linkinghub.elsevier.com/retrieve/pii/S0272775706000227`.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9 (8):1735–1780, 1997.

P. Holme and J. Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.

J. Horn. Human research and complexity theory. *Educational philosophy and theory*, 40(1):130–143, 2008.

J. Huttenlocher, M. Vasilyeva, H. R. Waterfall, J. L. Vevea, and L. V. Hedges. The varieties of speech to young children. *Developmental Psychology*, 43(5):1062–1083, 2007. ISSN 1939-0599(Electronic),0012-1649(Print). doi: 10.1037/0012-1649.43.5.1062.

L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck. What's in a crowd? analysis of face-to-face behavioral networks. *Journal of theoretical biology*, 271(1):166–180, 2011.

L. M. Justice, Y. Petscher, C. Schatschneider, and A. Mashburn. Peer Effects in Preschool Classrooms: Is Children's Language Growth Associated With Their Classmates' Skills? *Child Development*, 82(6):1768–1777, Nov. 2011. ISSN 00093920. doi: 10.1111/j.1467-8624.2011.01665.x. URL `http://doi.wiley.com/10.1111/j.1467-8624.2011.01665.x`.

A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

M. Karsai, M. Kivelä, R. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, 83(2):025102, 2011.

M. Karsai, N. Perra, and A. Vespignani. Time varying networks and the weakness of strong ties. *Scientific reports*, 4(1):1–7, 2014.

M. Karsai, H.-H. Jo, K. Kaski, et al. *Bursty human dynamics*. Springer, 2018.

M. Kibanov, M. Atzmueller, C. Scholz, and G. Stumme. Temporal evolution of contacts and communities in networks of face-to-face human interactions. *Science China Information Sciences*, 57(3):1–17, 2014.

M. Kivelä, J. Cambe, J. Saramäki, and M. Karsai. Mapping temporal-network percolation to weighted, static event graphs. *Scientific reports*, 8(1):12357, 2018.

K. Klemm and V. M. Eguiluz. Growing scale-free networks with small-world behavior. *Physical Review E*, 65(5):057102, 2002.

A. Kornai. *Mathematical linguistics*. Springer Science & Business Media, 2007.

G. Kossinets and D. J. Watts. Origins of homophily in an evolving social network. *American journal of sociology*, 115(2):405–450, 2009.

D. Krackhardt and R. N. Stern. Informal networks and organizational crises: An experimental simulation. *Social psychology quarterly*, pages 123–140, 1988.

P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Physical review letters*, 85(21):4629, 2000.

A. Krogh, B. Larsson, G. Von Heijne, and E. L. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of molecular biology*, 305(3):567–580, 2001.

W. Labov. *Principles of linguistic change Vol. 2 : Social Factors*. Blackwell, Oxford, 2001. ISBN 0-631-17915-1 978-0-631-17915-3 0-631-17916-X 978-0-631-17916-0.

J. Ladyman, J. Lambert, and K. Wiesner. What is a complex system? *European Journal for Philosophy of Science*, 3(1):33–67, 2013.

V. Lampos, N. Aletras, J. K. Geyti, B. Zou, and I. J. Cox. Inferring the socioeconomic status of social media users based on behaviour and language. In *European conference on information retrieval*, pages 689–695. Springer, 2016.

M. Latapy, T. Viard, and C. Magnien. Stream graphs and link streams for the modeling of interactions over time. *Social Network Analysis and Mining*, 8(1):1–29, 2018.

D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009a.

D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009b.

M. Le Normand, C. Parisse, and H. Cohen. Lexical diversity and productivity in French preschoolers: developmental, gender and sociocultural factors. *Clinical Linguistics & Phonetics*, 22(1):47–58, Jan. 2008. ISSN 0269-9206. doi: 10.1080/02699200701669945. URL `https://doi.org/`

10.1080/02699200701669945. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/02699200701669945.

O. Lederman, D. Calacci, A. MacMullen, D. C. Fehder, F. E. Murray, and A. Pentland. Open badges: A low-cost toolkit for measuring team communication and dynamics. *arXiv preprint arXiv:1710.01842*, 2017.

J. Lee, C. D. Campbell, and S. Chen. China multi-generational panel dataset, liaoning (cmgpd-ln), 1749-1909. *Data Sharing for Demographic Research (DSDR)*, 2010.

J. Z. Lee and C. D. Campbell. China multi-generational panel dataset, liaoning (cmgpd-ln), 1749-1909 (icpsr 27063). Dataset at DSDR https://www.icpsr.umich.edu/web/DSDR/studies/27063, 2016. URL https://doi.org/10.3886/ICPSR27063.v10.

J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. In *Proceedings of the 2007 SIAM international conference on data mining*, pages 551–556. SIAM, 2007.

A. Li, S. P. Cornelius, Y.-Y. Liu, L. Wang, and A.-L. Barabási. The fundamental advantages of temporal networks. *Science*, 358(6366):1042–1046, 2017.

Y. Liu, S. Dai, C. Wang, Z. Zhou, and H. Qu. Genealogyvis: A system for visual analysis of multidimensional genealogical data. *IEEE Transactions on Human-Machine Systems*, 47(6):873–885, 2017.

N. V. Loayza. The economics of the informal sector: a simple model and some empirical evidence from latin america. In *Carnegie-Rochester conference series on public policy*, volume 45, pages 129–162. Elsevier, 1996.

L. Lotero, R. G. Hurtado, L. M. Floría, and J. Gómez-Gardeñes. Rich do not rise early: spatio-temporal patterns in the mobility networks of different socioeconomic classes. *Royal Society open science*, 3(10):150654, 2016.

S. Luo, F. Morone, C. Sarraute, M. Travizano, and H. A. Makse. Inferring personal economic status from social network location. *Nature Communications*, 8(1):1–7, 2017.

C. Lynn Martin, R. A. Fabes, L. D. Hanish, and T. Hollenstein. Social dynamics in the preschool. *Developmental Review*, 25(3):299–327, Sept. 2005. ISSN 0273-2297. doi: 10.1016/j.dr.2005.10.001. URL https://www.sciencedirect.com/science/article/pii/S0273229705000249.

N. Madar, T. Kalisky, R. Cohen, D. Ben-avraham, and S. Havlin. Immunization and epidemic dynamics in complex networks. *The European Physical Journal B*, 38(2):269–276, 2004.

S. K. Maity, T. V. Manoj, and A. Mukherjee. Opinion formation in time-varying social networks: The case of the naming game. *Physical Review E*, 86(3):036110, 2012.

R. N. Mantegna and H. E. Stanley. *Introduction to econophysics: correlations and complexity in finance*. Cambridge university press, 1999.

R. Marion and M. Uhl-Bien. Leadership in complex organizations. *The Leadership Quarterly*, 12(4):389–418, 2001.

L. Martinet, C. Crespelle, E. Fleury, P.-Y. Boëlle, and D. Guillemot. The link stream of contacts in a whole hospital. *Social Network Analysis and Mining*, 8(1):59, 2018.

A. J. Mashburn, L. M. Justice, J. T. Downer, and R. C. Pianta. Peer effects on children's language achievement during pre-kindergarten. *Child development*, 80(3): 686–702, 2009. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8624.2009.01291.x/full.

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

J. M. McPherson and L. Smith-Lovin. Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American sociological review*, pages 370–379, 1987.

M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

K. A. Mollica, B. Gray, and L. K. Trevino. Racial homophily and its persistence in newcomers' social networks. *Organization Science*, 14(2):123–136, 2003.

J. Moody. Race, school integration, and friendship segregation in america. *American journal of Sociology*, 107(3):679–716, 2001.

S. R. Musse and D. Thalmann. A model of human crowd behavior: Group interrelationship and collision detection analysis. In *Computer Animation and Simulation'97*, pages 39–51. Springer, 1997.

A. Nardy. Dylnet - espace families. `https://dylnet.univ-grenoble-alpes.fr/fr/espace-familles`, 2016. [in French].

A. Nardy, J.-P. Chevrot, E. Fleury, M. Karsai, H. Bouchet, and S. Dai. Dylnet - language dynamics, linguistic learning, and sociability at preschool. Dataset and code at synapse.org `https://www.synapse.org/#!Synapse:syn26560886`, 2022. URL `http://doi.org/10.7303/syn26560886`.

M. Newman. *Networks*. Oxford university press, 2018.

M. E. Newman. Complex systems: A survey. *arXiv preprint arXiv:1112.1440*, 2011.

M. E. Newman, S. Forrest, and J. Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101, 2002.

V. Nicosia, J. Tang, C. Mascolo, M. Musolesi, G. Russo, and V. Latora. Graph metrics for temporal networks. In *Temporal networks*, pages 15–40. Springer, 2013.

A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: universal patterns in human urban mobility. *PloS one*, 7(5):e37027, 2012.

J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the national academy of sciences*, 104(18):7332–7336, 2007.

R. K. Pan and J. Saramäki. Path lengths, correlations, and centrality in temporal networks. *Physical Review E*, 84(1):016105, 2011.

R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925, 2015.

A. Portes and J. Sensenbrenner. Embeddedness and immigration: Notes on the social determinants of economic action. *American journal of sociology*, 98(6):1320–1350, 1993.

D. Preoţiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, and N. Aletras. Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9):e0138717, 2015.

J. Ray, A. Pinar, and C. Seshadhri. Are we there yet? when to stop a markov chain while generating random graphs. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 153–164. Springer, 2012.

M. L. Rowe. A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development*, 83(5): 1762–1774, Oct. 2012. ISSN 1467-8624. doi: 10.1111/j.1467-8624.2012.01805.x.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

M. Salathe, L. Bengtsson, T. J. Bodnar, D. D. Brewer, J. S. Brownstein, C. Buckee, E. M. Campbell, C. Cattuto, S. Khandelwal, P. L. Mabry, et al. Digital epidemiology. 2012.

A. J. Santos, B. E. Vaughn, and K. K. Bost. Specifying social structures in preschool classrooms: Descriptive and functional distinctions between affiliative subgroups. *Acta Ethologica*, 11(2):101–113, 2008.

D. R. Schaefer, J. M. Light, R. A. Fabes, L. D. Hanish, and C. L. Martin. Fundamental principles of network formation among preschool children. *Social Networks*, 32(1):61–71, Jan. 2010. ISSN 03788733. doi: 10.1016/j.socnet.2009.04.003. URL `http://linkinghub.elsevier.com/retrieve/pii/S0378873309000197`.

C. Schechter and B. Bye. Preliminary evidence for the impact of mixed-income preschools on low-income children's language growth. *Early Childhood Research Quarterly*, 22(1):137–146, Jan. 2007. ISSN 08852006. doi: 10.1016/j.ecresq.2006.11.005. URL `http://linkinghub.elsevier.com/retrieve/pii/S0885200606000846`.

B. Schuller, G. Rigoll, and M. Lang. Hidden markov model-based speech emotion recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 2, pages II–1. Ieee, 2003.

M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

J. Shin, L. Jian, K. Driscoll, and F. Bar. The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, 83:278–287, 2018.

W. Shrum, N. H. Cheek Jr, and S. MacD. Friendship in school: Gender and racial homophily. *Sociology of Education*, pages 227–239, 1988.

K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1): 22–36, 2017.

K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188, 2020.

C. E. Snow. Academic Language and the Challenge of Reading for Learning About Science. *Science*, 328(5977):450–452, Apr. 2010. doi: 10.1126/science.1182597. URL `https://www.science.org/doi/10.1126/science.1182597`. Publisher: American Association for the Advancement of Science.

T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In *Motion-based recognition*, pages 227–243. Springer, 1997.

J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Régis, J.-F. Pinton, N. Khanafer, W. Van den Broeck, et al. Simulation of an seir infectious disease model on the dynamic contact network of conference attendees. *BMC medicine*, 9(1):1–15, 2011a.

J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PloS one*, 6(8):e23176, 2011b.

T. Takaguchi, M. Nakamura, N. Sato, K. Yano, and N. Masuda. Predictability of conversation partners. *Physical Review X*, 1(1):011008, 2011.

D. Tang, B. Qin, and T. Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432, 2015.

A. Varga and R. K. Moore. Hidden markov model decomposition of speech and noise. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 845–848. IEEE, 1990.

A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):036127, 2006.

A. Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939): 425–428, 2009.

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.

N. Voirin, C. Payet, A. Barrat, C. Cattuto, N. Khanafer, C. Régis, B.-a. Kim, B. Comte, J.-S. Casalegno, B. Lina, et al. Combining high-resolution contact data with virological data to investigate influenza transmission in a tertiary care hospital. *Infection Control & Hospital Epidemiology*, 36(3):254–260, 2015.

A. J. Walhout, R. Sordella, X. Lu, J. L. Hartley, G. F. Temple, M. A. Brasch, N. Thierry-Mieg, and M. Vidal. Protein interaction mapping in c. elegans using proteins involved in vulval development. *Science*, 287(5450):116–122, 2000.

J. Wallinga, W. J. Edmunds, and M. Kretzschmar. Perspective: human contact patterns and the spread of airborne infectious diseases. *TRENDS in Microbiology*, 7(9):372–377, 1999.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

C. Werner and P. Parmelee. Similarity of activity preferences among friends: Those who play together stay together. *Social Psychology Quarterly*, pages 62–66, 1979.

M. J. Williams and M. Musolesi. Spatio-temporal networks: reachability, centrality and robustness. *Royal Society open science*, 3(6):160196, 2016.

L. Wu, F. Morstatter, K. M. Carley, and H. Liu. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, 21(2):80–90, 2019.

S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 705–714, 2011.

J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *CVPR*, volume 92, pages 379–385, 1992.

W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.

K. Zhao, J. Stehlé, G. Bianconi, and A. Barrat. Social network dynamics of face-to-face interactions. *Physical review E*, 83(5):056109, 2011.

Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*, pages 791–800, 2009.

Y. Zheng, X. Xie, W.-Y. Ma, et al. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.