Imperial College of Science, Technology and Medicine
Department of Physics, Centre for Complexity Science

Yi-Er-San Topics in Network Science: Centrality, Bicycle, Triplet

Bingsheng Chen

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Physics of Imperial College London, Mar 2022.

Abstract

Network science studies interactions between different entities. This thesis covers three different topics in network science: higher-order network structures, human mobility and network centralities. Higher-order network is an emerging field in recent years and combinatorial models use multi-body interactions to describe the structures beyond pairwise. In chapter two, I focus on studying the evolution of the links in temporal networks using three nodes motif– triplets. In specific, I develop a method that use a transition matrix to describe the evolution of triplets in temporal networks. To identify the importance of higher-order interactions in the evolution of networks, we compare both artificial and real-world network data to a model based on pairwise interactions only. The differences between the transition matrix and the calculated matrix from the fitted parameters demonstrate that non-pairwise interactions exist for various real-world systems in space and time. Furthermore, this also reveals that different patterns of higher-order interaction are involved in different real-world situations. To test my approach, I used the transition matrix to design a link prediction method– Triplet Transition score. I investigate the performance of the methods on four temporal networks, comparing my approach against ten other link prediction methods. My results show that higher-order interactions in both space and time play a crucial role in the evolution of networks as I find Triplet Transaction method, along with two other methods based on non-local interactions, gives the best overall performance. The results also confirm the concept that the higher-order interaction patterns, i.e., triplet dynamics, can help us understand and predict the evolution of different real-world systems.

In chapter three, I investigate the behaviours of human mobility in different cities using gravity models. Due to previous technical challenges in collecting data on riding behaviours, there have been only a few studies focusing on patterns and regularities of biking traffic. To extend the research, I use the data from mobike and apply the gravity model to study the mobility of dockless bicycles. I validate the effectiveness of the general gravity model on predicting biking traffic at fine spatial resolutions of locations. I then further study the impacts of spatial scale on the gravity model, and reveal that the distance-related parameter grows in a similar way as population-related parameters when the spatial scale of locations increases. The result reveals the emergence of the scaling can be explained by the gravity models.

Measuring the importance of nodes in networks via centrality measures is an important task

in many network systems. There are many centrality measures available and it is speculated that many encode similar information but the reason behind them is rarely studied. In chapter four, I give an explicit non-linear relationship between two of the most popular measures of node centrality: degree and closeness. Based on a shortest-path tree approximation, I give an analytic derivation that shows the inverse of closeness is linearly dependent on the logarithm of degree. I show that the hypothesis works well for a range of networks produced from stochastic network models and for networks derived from many real-world data sets. I connect our results with previous results for other network distance scales such as average distance. My results imply that measuring closeness is broadly redundant unless our relationship is used to remove the dependence on degree from closeness. The success of our relationship suggests that most networks can be approximated by shortest-path spanning trees which are all statistically similar two or more steps away from their root nodes.

Acknowledgements

I would never think I would pursue a Ph.D. during my first two years at Imperial, since I almost failed my QM exam. Puwen and Yichen reminded me that the purpose of taking a physics course was curiosity instead of grades. With the help of Prof. Vvedensky, I learned how to write notes and the proper way of learning physics. William Zhong, Luca Coconi, Andreas, Giovanni, and Adam saved physics life by tutoring my physics and demonstrating to me a different way of thinking. Thanks to my previous lab and project partners, Dily, Zhengyu, and Tony. Without working with you I do not think I can survive undergraduate.

I would appreciate my first supervisor, Prof. Kim Christensen. When I received quite negative feedback from editors or referees, he always encouraged me to treat these comments on a positive side, and think about how to improve the original manuscripts, though they may be quite irrelevant. In addition, Kim taught me a lot about what is good science and how to think independently and be critical, the result is not everything but the explanation behind them is more important. These qualities are important, not only for academics but also for the life. He provided a lot of help with my thesis and I believe he knows my thesis better than I do. Thank you, Kim, you are a great supervisor and I hope I will become an independent researcher in a foreseeable future.

I would appreciate my second supervisor, Dr. Tim Evans. Thank you for trusting me even when I was in a very difficult situation during my third year of undergrad. I cherished the time discussing with him and also coffee. He is always very supportive and ready for discussing new ideas on the blackboard. Though most ideas from me were rubbish, he would listen and try to develop these ideas with me. He also taught me to stand out if everyone in the collaboration team is being a little bit silent. He makes me feel that Imperial is my academic home.

I would also thank my colleague and some friends who meet at conferences. Hardik, Vaiva, and Andrew, the discussion inspired me that people working in a very similar field should not only be competitors but also have the chance of being a collaborator. Max, thank you for letting me know my life is not contradicted with my studies. Qing and Nanxing, thank you for being the most supportive, without you, I think my life in the UK would be very difficult. Chester provides a lot of valuable help to me, which helps me to find a postdoc. Henry and Henrik, thank you for helping me to learn how to collaborate with people from other fields. Junming

and Fernando helped a lot in formulating the Triplet paper and others too numerous to list.

The last 6 months of my Ph.D. studies is my most difficult time. I have nowhere to stay but thank you for the help from Hongzheng, Ruiqi, and Jingfang. Ruiqi was my first collaborator outside Imperial and formulating many ideas with him from scratch is one of my happiest times. Gezhi and Runze encourage me a lot not to give up. They also tell me many possible choices for my future I have never thought of and make me feel more positive. Thanks to Yudi, who provides me with places to stay and finish this thesis. I would appreciate my parents and my girlfriend since they are always supporting me, though they do not necessarily know what I am doing.

In the end, I would appreciate readers who are interested in reading this thesis. That is the only time I feel I am contributing a tiny little piece of work to the world.

Declaration of originality

I declare that the work presented in this thesis is entirely my own work unless stated otherwise. I have acknowledged contributions from collaborators and others according to the standard referencing practices. Further, I have sought permission to reproduce any third-party copyrighted material in this thesis from the copyright holders and collaborators. All permissions are included at the end of this thesis.

Bingsheng Chen, Mar 2022

Copyright declaration

Publications

The following publications form the basis of this thesis:

Li, R., Gao, S., Yao, Q., Chen, B., Luo, A., Shang, F., Jiang, R. & Stanley, H.E. *Gravity model in dockless bike-sharing systems within cities. Physical Review E* **103**(1), 012312 (2021).

Chen, B. , Yao, Q. , Christensen, K. & Evans, T.S. *Higher-Order Temporal Network Effects through Triplet Evolution. Scientific Report*, **11**-15419 (2021).

Evans, T.S. & Chen, B. *Linking network centrality measures: degree and closeness.* accepted by *Communication Physics* (2022).

Publications completed during my PhD which are not included in this thesis:

Chen, B., Lin, Z., Evans, T. S. *Analysis of the Wikipedia Network of Mathematicians.* arXiv:1902.07622 (2018).

Contents

List of Tables

List of Figures

Symbols

Symbols & acronyms for Chapter 1

$\mathcal{G}$	a graph
v	a node (vertex) in a graph
e	an edge (link) in a graph
$\mathcal{V}$	set of all vertices in a graph
N	number of nodes in a graph
$\mathcal{E}$	set of all edges in a graph
$\mathbf{A}$	adjacency matrix of a graph
k_i	degree of node i
C_i	clustering coefficient of node i
$d(i,j)$	geodesic distance between nodes i and j
c_i	clustering coefficient of node i
b_i	betweenness of node i
$\sigma(i,j)$	number of shortest paths available from vertex i to j
$\sigma(i,j\|v)$	number of shortest paths from i to j which pass through vertex v
$\mathbf{T}$	diffusion matrix of a graph
α	probability of hyperjump
G	PageRank matrix of a graph
ER	Erdős-Renyi
BA	Barabási-Albert

Δt	time resolutions of a temporal network

Symbols & acronyms for Chapter 2

$\mathcal{M}$	sets of triplet states
m_i	states of triplet
$\mathsf{T}(s)$	probability of triplet transition at time s
$\widehat{\mathsf{T}}(s)$	empirical probability of triplet transition at time s
p	probability of an edge is created when no edge before
$\hat{p}$	empirical probability of an edge is created when no edge before
q	probability of an edge is destroyed given there is an edge before
$\hat{q}$	empirical probability of an edge is destroyed given there is an edge before
$\mathbf{Z}$	Z-score matrix obtained from compare network with a pair-wise model
$\psi(i,j;s)$	states vector of node pair i,j at time s
β	indication of an edge
$L_\beta(i,j;s)$	probability that a pair of nodes u,v has a link/does not have a link at time s
$s(i,j)$	similarity between node i and j
J	an objective function of clustering
AUC	area under curve

Symbols & acronyms for Chapter 3

$\rho(r)$	average density of riding activities at a distance r to the city center

d_{ij}	distance between location i and j
$\langle d_i \rangle$	average distance travelled from location i
$\mathbf{T}$	biking flux matrix between locations
w_{ij}	the ratio between traffic volume from location i to j.
P_i	population of the location i

Symbols & acronyms for chapter 4

SPT	shorest path tree
$d(u, v)$	geodesic distance between nodes u and v
$c(v)$	closeness of node v
$n_v(\ell)$	the number of nodes from the root node v at distance ℓ
ℓ	distance from root node v
$\bar{z}$	growth factor of the shortest path tree
$\mathcal{T}(v)$	shortest path tree from root node v
$n_\ell(v)$	number of nodes at distance ℓ from the root v
L	cut-off distance from root node v
β	intercept term in the relation between closeness and degree
$\bar{z}^{(\text{fit})}$	growth factor of the shortest path tree estimated from the data
$\beta^{(\text{fit})}$	intercept term in the relation between closeness and degree estimated from the data
$\langle \ell \rangle$	the average distance between all nodes pairs

Chapter 1

Introduction

1.1 Motivation and objectives

Complexity science is a study of the systems and one of the goals is to understand how collective behaviour emerges from microscopic level interactions in different types of systems. Those collective behaviours can be highly dynamical and non-linear [1, 2]. Traditional mathematical approaches rely on homogeneous assumptions that may not working well on some systems. One example would be modelling epidemiology spreading over population. The traditional approaches, such as compartmental models, assumes that individuals in the whole population are identical and have equal probability to infect each other or being recovered. However, the assumption is not realistic since individuals are confined in their own social groups and more difficult to interact with individual in other social groups. To provide a more accurate way to model the complex system, network can be used to describe the structure in the system. If any interactions occur between a pair of entities, then an edge connects these two nodes [3].

Network methods are widely used in many fields. In computational social science, network methods can be used to infer topical structures from the collection of documents [4] and to understand the emergence of a new field of science [5]. Network methods also help to better understand biological science, from forecasting protein-protein interactions, predicting drug combinations to drug re-purposing [6, 7, 8]. In addition, network methods can be used to

understand human mobility and spatial scaling laws behind cities [9, 10].

There are many dynamical processes taking place on networks, for instance, disease spreading [11] and synchronisation process in voting dynamics [12]. To predict outcomes of these systems, determining the most influential nodes is an important goal. The node importance is also known as node centrality and have wide applications in many fields, from social network analysis [13] to search engines [14]. There are a vast number of centrality indices available as visualised nicely by Schoch [15, 16]. However, many different centrality measures encode similar information, which leads to redundancy. A variety of research found the strong correlations between centrality indices [17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27]. In particular, Pearson correlation coefficients are invariably used which are most sensitive to linear correlations between centrality measures[1]. However, there seems to be no clear consensus from these studies other than there are often strong relationships between centrality measures but these vary from network to network. The reason behind the strong correlations is not clear. There may exists some intrinsic structures behind observed topology, for instance, in a shortest path tree structures, the inverse of closeness will highly correlated with its degree.

In addition, the structures in networks are hidden and can be misleading in network modelling. The edge is often used to represent pairwise interactions. However, in systems represented by networks, there are often processes, known as higher-order interactions, where groups of more than two participants interact. In these cases, the set of pairwise interactions does not capture the whole picture [28]. Research has revealed that many empirical systems display higher-order interactions, for example social systems [29, 30, 31], neuroscience [32, 33, 34], biology [35], and ecology [36]. Examples of network structure beyond pairwise edges would be network motif and [37]. Network motif refers to recurring and significant interconnection patterns appears in the networks comparing to corresponding random graph null model. Network motif are also known as sub-graphs(i.e. collection of the nodes and only connection patterns within these nodes). It is found that network motifs act as functional units in different systems[37], thus, they can be thought as a single unit. At the same time, graphlet focused on induced subgraphs. An

[1]Linear correlation between two variables means if one variable increases, another variable also grows in a constant ratio. Otherwise, the correlation is non-linear, for instance, another variable grows exponentially faster than its correlated variable.

induced sub-graph means that once the nodes was selected in the large network, all the edges between them to form the sub-graph, which helps to predict disease protein in protein-protein interaction networks [38].

In addition, the network is not static and the edges contain temporal information, for instance, while people sending emails, the replies will arrive later, so in network, they should not appear at the same time. More formally, a temporal network is a network whose structure changes over time. There are many examples of temporal networks in social systems, including conference meeting [39], shareholder co-investing [40] and co-operative behaviours [41]. More examples in different fields can be found in Holmes's summary [42]. However, apart from triadic closure [30, 31], the evolution's of temporal network is less well explained by higher-order interactions. Thus, we looked into details about the gap between these two topics.

The title of this thesis refers to the three topics discussed. "Yi", or 一 is Chinese for the number one and this refers to chapter 4 on centrality where the focus is on the importance of each individual node. The goal of chapter 4 is to understand the relations behind centrality measures. "Er" or 二 is the Chinese for two and here this represents chapter 3 on the spatial network of bicycle hire used to understand human mobility patterns. "San" or 三 is the Chinese for three which exemplifies chapter 2 on node triplets used to understand the edge dynamics in temporal networks.

1.2 Background of network theory

Network theory is developed from graph theory and many concepts are used in both studies, since both fields concerned with pairwise relations between objects[43, 44]. However, these two fields are commonly taken as two separate directions with little overlaps. Graph theory is closer to pure mathematics and has focused on providing rigorous proofs for graph properties[43], whereas, network science focused more on providing explanations to observations of real-world data sets. Real world data sets are very complex, which includes heterogeneous interactions[45], temporal structure[46] and thus, analytically intractable, while graph theory are more interests

on working with analytically tractable problems, such as random graph[43], thus, this thesis belongs to network science instead of graph theory. In addition, since network science is a very interdisciplinary study. Therefore, multiple terminologies are used in different fields to describe the same objects, for instance, computer science communities prefer 'node' and 'link' to describe the object and their binary relations, but this would be identical to 'vertex' and 'edge', which is more commonly used in mathematics communities. If any of them appears in the thesis, they mean the same things. In the current chapter, we will introduce common definitions and terminologies for network theory first. For those specific terminologies used in each chapter, they are introduced in each chapter for simplicity. We begin with a general mathematical frameworks for representing graph theory and then some classes of random models, including Erdős-Renyi networks, Barabási-Albert networks and configuration models. We will later show how we can measure structural importance of nodes (centrality) in networks. The chapter is a summary of terminologies about the network, readers can find more details in [3, 46, 47, 48].

1.3 Networks, representation and structure

In complex systems, many units are interacting with one another. The interactions can be studied via a network approach. For instance, if the interactions only occur between every pair of units, then a *dyadic*[2] network is used to capture the pairwise interaction. A similar logic can be extended to take into consideration more units, for instance, triad interactions. More details can be found in [50, 28]. In this chapter, we focus on the pairwise networks. Dyadic networks are also called graphs $\mathcal{G}$, which is a collection of sets of *vertices* $\mathcal{V}$ and *edges*, $\mathcal{E}$ as:

$$\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}. \tag{1.1}$$

Furthermore, an edge is defined by the pair of nodes connected, for example, if i is connected to j, then we denote this edge by $e = (i, j) \in \mathcal{E}$. Meanwhile, the order of the vertices in an edge corresponds to the direction of the edge, for instance, (i, j) means i is the source node

[2]Dyadic means a group is consisting of two parts

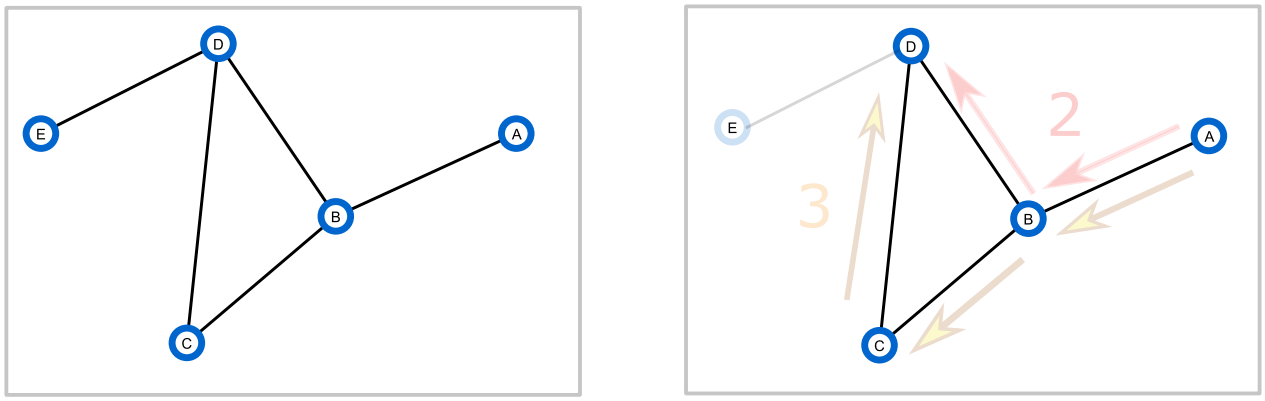


Figure 1.1: An illustration of network and path. The left figure is the topology described via the adjacency matrix in Eq.(1.4) and the right figure demonstrates how a path is defined in the network. The red path is the shortest path of length 2 started from A to D and the yellow path is the longest path of length 3 from A to D. The number corresponds to the path length. The node E is denoted in lighter blue since it is not an immediate node in any direct path origin from A to D.

and j is the target node. In reality, more information can be included to describe features of networks, for instance, attributes of nodes (node type) can be used to describe the characters of nodes, edge weight (edge importance) can be used to describe the strength or frequency of links, etc. Neglecting these features can cause the resultant network to have different structures, for example, treating a graph as unweighted, while many edges in the network are significantly weaker than other links, in terms of edge weight, can be very misleading. However, while constructing the network from the data, in many cases, the information is not fully accessible and assumptions need to be made. There are several different types of networks based on the characters of the links in networks. For example, a network is called a simple network if there is no self-loop, i.e. $\forall(i,j) \in \mathcal{E} : i \neq j$. For an undirected graph, the links are symmetrical which means, $\forall(i,j) \in \mathcal{E}, \exists(j,i) \in \mathcal{E}$ and reversely, for a directed network, $\exists(i,j) \in \mathcal{E}, s.t.(j,i) \notin \mathcal{E}$. A path is a finite or infinite sequence of edges that joins a sequence of vertices, which simulate a walker travelling from a source node i to a target node j. The length of a path is defined as the number of links traversed by the walker. The path from one node to another node is not unique, for instance, there exists two paths from node A to node D (see the yellow and red paths in Figure 1.1). The longest path of length 3 between node A to D is indicated by yellow arrows and the shortest path of length 2 is indicated by red arrows. The shortest path length is also known as the geodesic distance between two nodes, which plays an important role in

many network processes, such as , in a traffic networks, people will choose the shortest path route to minimise travelling time.

There are two major data structures to represent network: one is the edge-list representation and another one is the adjacency matrix representation. The edge list representation stores all connected pairs of nodes (tuples of two nodes) in an array, for instance:

$$\mathcal{E} = \{(i, j)\}, \forall i, j \in \mathcal{V}. \tag{1.2}$$

For example shown in Figure 1.1, the corresponding edge list representations are stated as the following:

$$\mathcal{E} = \{(A, B), (B, C), (B, D), (C, D), (D, E)\}. \tag{1.3}$$

Depending on the task, it is necessary to choose an appropriate representation of a network. Each representation of network is amenable to mathematical operations that are more optimal for solving certain tasks. The advantage of edge-list representation is that only connected edges in the graph are stored as a list or vector and therefore, memories are used more efficiently, especially for sparse networks. Further advantages include, link randomisation and numerical simulations in sparse networks can be computed more easily [46]. The edge-list representation is capable of storing all sorts of graphs, either dense or sparse, since the upper-bound of the memory is N^2, where $N = |\mathcal{V}|$ is the number of nodes in the graph $\mathcal{G}$.

In cases that involves determining the paths in the network, the edge list representation is not the most ideal way, since it needs to repeatedly loop through the list each time, depending on the path length l and the complexity will scale $\mathcal{O}(N^l)$. Alternatively, the adjacency matrix $\mathbf{A}$ can be used to represent a graph, where each binary entry A_{ij} indicates whether an edge exists between nodes i and j.

The corresponding adjacency matrix for the graph shown in Figure 1.1 is

$$\mathbf{A} = \begin{array}{c} \\ \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \\ \mathrm{D} \\ \mathrm{E} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{D} & \mathrm{E} \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}. \tag{1.4}$$

The path in a network can be studied through the power of the associated adjacency matrix, for instance, if we want to understand all possibilities of where the random walker can land within path length of two, then we have $\mathbf{A}^2$ the following:

$$\mathbf{A}^2 = \begin{array}{c} \\ \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \\ \mathrm{D} \\ \mathrm{E} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{D} & \mathrm{E} \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 3 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 3 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{pmatrix} \end{array}. \tag{1.5}$$

Each entry A^2_{ij} counts the number of length 2 paths between node i and j. Therefore, the adjacency matrix method is very efficient to determine all paths with a certain length and helps to solve triadic closure problems. Notice, the diagonal term A^2_{ii} counts the number of paths of length 2, which go back to the original node i via node j, for instance, $A \to B \to A$. Therefore, the diagonal term A^2_{ii} is the same as the degree of node j. The adjacency matrix method is very powerful in studying dynamical processes and theoretical structure in networks, since it provides ways to apply linear algebra tools, such as eigenvectors and eigenvalue spectra [46]. The memory cost of the adjacency approach is $\mathcal{O}(N^2)$. In many real data sets, most networks are sparse and therefore, a majority of entries in the network will be zeros and the memory is not used efficiently. To tackle this problem, sparse matrices can generally be used to provide more efficient usages of memory and faster computations, however, the numerical

instability need to be checked [52] [3]. We can use the package *Scipy 1.3* to operate the sparse matrix and *Networkx 1.1* to store graphs and compute the corresponding quantities in networks.

1.4 Structures and centrality

There are many graph properties that can be used to describe the structures of different networks. Social Network Analysis (SNA) is one of the methods that can be used to analyse the structure of the network in terms of nodes and links (not only in sociology but also in many other fields, including epidemiology). In this section, we will introduce several centrality measures that can be used to quantify the local or global importance of nodes.

1.4.1 Centrality measures

There are several centrality measures that can be used to characterise the structure importance of nodes in networks (see more details in[15, 16]). Degree is the most simple and widely used centrality measure, it measures the number of nodes a given node is connected to. The degree of node i is defined as:

$$k_i = \sum_{j \in \mathcal{V}} A_{ij}. \tag{1.6}$$

Degree is a local measure, since the measure only reflects the information about the node itself and the nearest neighbours. The clustering coefficient is a measure of how likely the nodes in a graph tend to cluster together. Clustering coefficient, C_i, can be used to measure the fraction of complete triangles centred on vertex i with the nearest neighbours:

$$C_i = \frac{2}{k_i(k_i - 1)} \sum_{j<k} A_{ij} A_{jk} A_{ik}. \tag{1.7}$$

If $C_i = 1$, all neighbours of node i will connect to each other. If $C_i = 0$, node i have a local tree structure. In addition, the network clustering coefficient, C of a network is simply the average

[3] During the matrix decomposition, the value is approximated by product and thus, sometimes not-exact and have some round-off errors.

of all clustering coefficients:

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i. \tag{1.8}$$

Clustering coefficient can also be used to distinguish some specific structures, for instance,

- A fully connected graph has a cluster coefficient, $C = 1$.
- Square lattices have zero clustering coefficient, so do trees.

Many measures are able to detect a global structure, for instance, closeness measures average geodesic distance to other nodes. Closeness c_v for a vertex v is defined to be [3, 53, 54]:

$$c_v = \frac{|\mathcal{C}_v| - 1}{\sum_{u \in \mathcal{C}_v \setminus v} d(u, v)}, \tag{1.9}$$

where $\mathcal{C}_u$ is the connected components where node u belongs to, $d(u, v)$ is the length of the shortest path between vertex u and some distinct vertex v which is in the same component, $\mathcal{C}_u$. Note that we use a standard normalisation using $|\mathcal{C}_v| = N - 1$, the number of remaining vertices apart from the node v in the connected components, but this is irrelevant after rescaling.

Betweenness measures the bridgeness of the node, that is, the fraction of the shortest paths pass through a node v. Betweenness has many important real world applications, including finding the target customer for advertisement and stopping the hyper-spreader in epidemiological breakout. Betweenness b_v of a vertex v is [3, 54, 55, 56]

$$b_v = \sum_{s,t \in \mathcal{C}_v} \frac{\sigma(s, t|v)}{\sigma(s, t)}. \tag{1.10}$$

$\mathcal{C}_v$ is the set of vertices of the component containing vertex v, $\sigma(s, t)$ is the number of shortest paths available from vertex s to t, and $\sigma(s, t|v)$ is the number of shortest paths from s to t which pass through vertex v. The numerator $\sigma(s, t)$ takes account of cases where there are more than one shortest paths between a pair of nodes s and t, which makes the nodes more important if less available paths exist.

Eigenvalue centrality measures the influence score by a broadcasting process. Broadcasting process refers to the spread of a piece of information to all the neighbours and the neighbours replicate information again. The Eigenvector centrality for a vertex i is simply the i-th entry of the eigenvector of $\mathbf{A}$ associated with the largest eigenvalue [3, 54].

PageRank simulates a random walker on network together with a hyper-jump process [14]. The process mimics the behaviour of humans browsing websites by clicking the hyperlinks on websites, with an additional probability of suddenly moving to another set of pages. PageRank is defined in terms of a transfer matrix, $\mathbf{T}$ where each entry, T_{ji} represents the probability of a random walker at vertex i moving to vertex j at the next time step. We have that

$$T_{ji} = \frac{1}{s_i^{(\text{out})}} A_{ji}\,, \quad \text{where} \quad s_i^{(\text{out})} = \sum_j A_{ji}\,. \tag{1.11}$$

An additional stochastic process also occurs. At each step, with probability α, the random walker follows a link chosen at random as given by the transfer matrix $\mathbf{T}$ but with probability $(1-\alpha)$ the current walk is deemed to end, or equivalently, we follow a new user or a new walk by starting at a randomly chosen vertex. The Markovian matrix $\mathbf{G}$ which describes this process is given by:

$$G_{ij} = \alpha T_{ij} + (1-\alpha)\frac{1}{N} \tag{1.12}$$

where N corresponds to the total number of vertices and α is the damping factor. The probability that a random walker is at vertex i in the long-time limit is proportional to the PageRank for that vertex and this is given by the i-th entry of the eigenvector associated with the largest eigenvalue of $\mathbf{G}$.

In addition, node centrality can be used to predict the existence of a link. There is much literature and many applications based on node centrality (see a good survey in [57]). In chapter 2, we will discuss how centrality can be used to do the task of link prediction.

1.4.2 Community structure and detection

Many networks are heterogeneous and vertices of a graph can be partitioned into different clusters that have the propensity of the links [58, 59]. Within the groups, the density of links is high, but there are fewer connections between different groups, see Figure 1.2. The communities

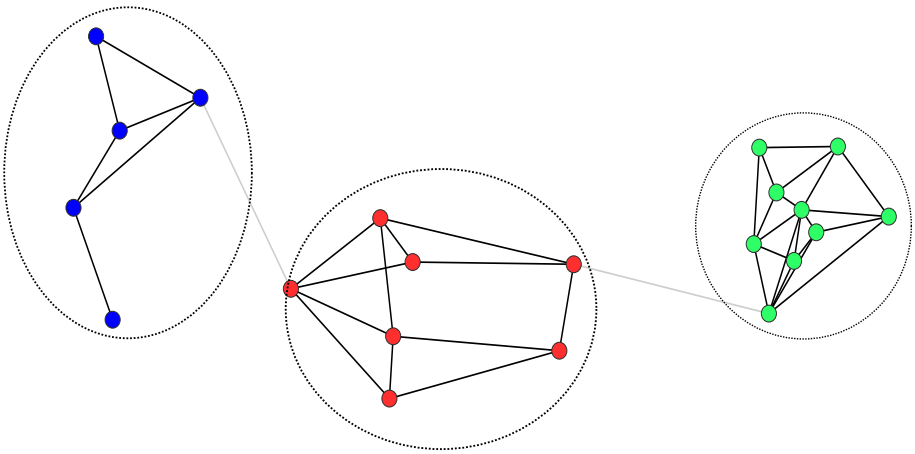
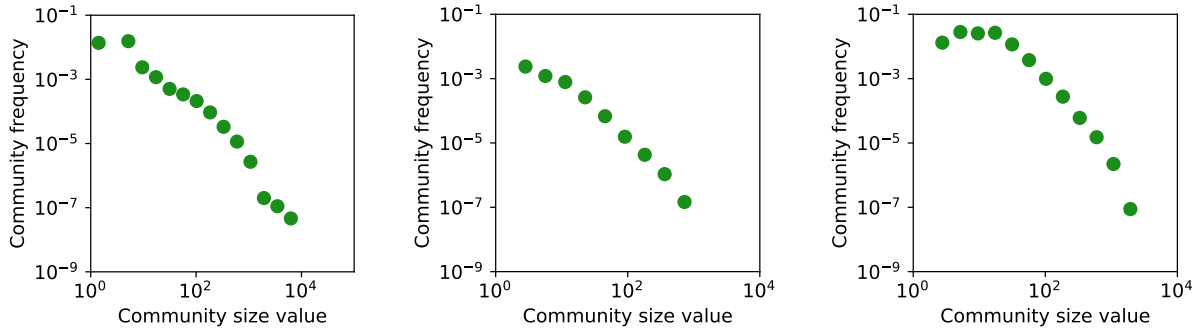


Figure 1.2: An example of the non-overlapping community structure in a network. There are three communities in this network, which are represented by dashed ellipses. We can see denser intra-community links (black links) within the group but less denser inter-community links between communities (grey links).

may also overlap, that is, nodes are shared between various communities [60]. There are many applications of the community structure in network analysis, for instance, discovering unknown functional modules such as topics in information networks, cyber-communities in social networks or functional units in the biochemical network [61, 62].

There are many community detection algorithms in the literature, for instance, divisive algorithms detect inter-community links (links between communities) according to their edge betweenness and remove them from the network [63], merging similar nodes/communities using random walks recursively [64] and maximising an objective function, such as modularity [65]. Since networks are generated from a variety of processes and formed from very complicate mechanisms, there is no single algorithm that can succeed in all detection goals [60, 66]. Furthermore, there are no clear guidelines on how to assess and compare the performance of different community detection algorithms [67], which leaves the question of the best algorithm detection open.

We also see the emergence of the community structures in real networks, several papers also suggest a scaling behaviour of the community-size distribution in networks[4], including networks of shareholder in Turkey and Netherlands [40], information networks, social networks, biological networks and internet networks, see Figure 1.3 [69].



(a) Communication-Email (b) Internet-AS-caida (c) Information-Amazon

Figure 1.3: Community size distributions in three different networks. All data illustrate similar scaling behaviours in log-log plots and the distribution of community sizes detected by Infomap from was quite broad [69].

1.5 Network models

Real networks have many interesting characteristics, for instance, power-law degree distributions [70] and small world (six-degree separation) phenomena [71]. Many networks will not simply be the structure of a lattice, fully connected or branching like a tree. One of the key goals in network science is to understand how a specific property is generated from a known mechanism. In this section, we look at three cases, the simplest Erdős-Renyi random network, Barabási-Albert networks and configuration models. There are many other processes such as copying model, block type model and Kronecker graph but these are not discussed in this thesis (see details in [72, 73, 74]).

[4]Partition result of a community detection algorithm is dependent on the choice of the algorithm, thus, the detected scaling behaviours of communities vary so stability need to be checked.

1.5.1 Gilbert and Erdős-Renyi random network

Gilbert random network and Erdős-Renyi random networks are essential since they can set basic benchmarks for real networks that help us to understand the uncertainty in networks. Originally, these simple random networks were formulated mathematically by Edgar Gilbert[75]. Two parameters are used to define the random graph, one is the number of the nodes N and the other is the probability, p, of any two nodes are connected. The model assumes that nodes may connect to any other node in the network with probability p. Therefore, the probability of a node do have a degree of k follows a binomial distribution:

$$p(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}. \tag{1.13}$$

In addition, it can be shown that if $k \ll N$, then the degree distribution can be approximated as a Poisson distribution (see details in [70]):

$$p(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}. \tag{1.14}$$

Paul Erdős and Alfred Reyni [76] proposed another version of the random graph. The only difference is instead of using probability of an edge will exist, ER model assumes that the total number of edges in the network is fixed and these links are assigned equally likely to all nodes. In the asymptotic limit $N \to \infty$, two model becomes equivalent.

Since the random graph has a low clustering coefficient, it is locally tree like.

1.5.2 Barabási-Albert model

The Barabási-Albert (BA) model uses preferential attachment mechanisms to generate scale-free networks. Scale-free networks have degree distributions $p(k) \sim k^{-\gamma}, k \gg 1$ and γ can vary depending on different type of networks, see more details in [77]. Preferential attachment is a process where nodes in a network gain edges using a cumulative advantages principle, where the richer get richer. The algorithm can be summarised as follows:

1. Create an initial graph $\mathcal{G}_t$.
2. Set $t \to t+1$.
3. Add a new vertex to the graph.
4. Add m edges at one end that connect to the new vertex.
5. Add the other ends of these m edges to the nodes in the existing graph with a probability $\Pi(k,t) = k/2N_E$, where N_E is the total number of edges in the existing network.
6. Check whether self-loop or multi-edges exists. If so, remove the edges and repeat step 5.
7. Repeat from step 2 until $|\mathcal{G}_t| + N$ nodes have been added to network.

As a consequence, the final degree distribution after a long time can be derived as (see detailed derivation at [70]):

$$p(k) = \frac{2m(m+1)}{k(k+1)(k+2)} \qquad k \geq m. \tag{1.15}$$

If $k \gg 2$, then $p(k) \sim k^{-3}$. There are some debates about the extension of scale-free networks in the real world and many other mechanisms should be studied to explain networks that are not scale-free [78].

1.5.3 Configuration model

The ER network can capture the property of variation of real networks but this is not enough to make it the best reference model. Real networks are not regular but nor are they random. The configuration model allows users to generate a random network with a given degree distribution, which provides a arbitrary reference model without higher-order structures, such as degree-degree correlations. Denoting $N_E = |\mathcal{E}|$ as the total number of edges in network, then the probability of two nodes i and j being connected purely depend on the degree product as followings:

$$p(A_{ij} = 1) = \frac{k_i k_j}{2N_E}. \tag{1.16}$$

Then a configuration model can be generated using a random number generator. Alternatively, the configuration model can also be made via an edge swap process. The edge swap process breaks two existing edges into two stubs. For high-degree nodes, the probability of self-loops and multi-edges formed will be more often than for low degree nodes.

1.6 Organisation of the thesis

In this thesis, chapter 2-4 can be read separately since each chapter covers one specific topic and does not overlap with other topics. We organise the thesis into the following chapters. In chapter 2, we discuss triplets. The goal of chapter 2 is to study higher-order network evolution mechanism in temporal networks. We treat three-nodes (triplet is 三,"San", in the title) as a whole unit and use triplets to explain the observed edge dynamics. In chapter 3, we discuss data on bicycles. Here, two is 二, "Er" in the title. We use the gravity models to demonstrate how scaling of flows can emerge from locations. In addition, we also investigate how the gravity model performs in different spatial resolutions. In chapter 4, we discuss the centrality of single node (一, 'Yi' in the title). Correlations between centrality measures always exist in many real data set but the reasons behind them are rarely studied. In chapter 4, we derive an approximate relationship between the inverse of closeness and logarithm of degree based on the shortest-path tree approximation. At the end, we use our results to explain the relation between average shortest path length and average of logarithm of degree in many real-world networks.

Chapter 2

Higher-order temporal network effect: Triplet evolution

The following chapter is based on:

Q., Yao, B., Chen, T., Evans and K., Christensen

Higher-order temporal network effects through triplet evolution.

Scientific Reports ***11****, 15419 (2021) .*

B. Chen contributed to all aspects of this paper.

2.1 Introduction

Higher-order structures in networks refer to non-dyadic interactions in networks. Dyadic networks use links to describe binary/pairwise relations between nodes. However, many phenomena in network are governed by local processes which involve more than two nodes. Detecting higher-order interactions in the network is not a simple task, since most of the data only reveals pairwise interactions making the simple graph the appropriate representation. For instance, data on phone calls are intrinsically pairwise connections, whereas, through face-to-face meetings, the pattern of calls can reveal the existence of higher-order connections. Therefore, inferring higher-order structures hidden in the data becomes an increasingly important question

to answer.

There are three common models used to understand higher-order networks, multi-layer network, combinatorial higher-order model and transitive path-based model [28]. The first type of the models are Multi-layer networks [45] can capture different types of interactions, examples include multi-layer financial networks [99] and so on.

The second type of models are combinatorial models, which are used to represent multi-body interaction in networks. There are many higher-order structures, for instance, motif, graphlet, hyper-edges and cliques. The simplest method to model local group interactions in a network is to use small sub-graphs, motifs [29, 37] and graphlets [205, 206]. Motifs usually quantify whether a sub-graph is under/over-represented compared to configuration networks. A example of a group of three-node motifs can be seen in Figure 2.1. Graphlets are different from motifs, a given set of graphlets are non-isomorphic maximal induced subgraphs which contain all the edges between nodes, whereas, the motif is the partial subgraph which may only contain some of the edges between the nodes. Higher-order analysis in terms of paths is important in several contexts [100, 101] but the use of cliques are more commonly used in higher-order networks. Cliques are fully connected subgraphs. Cliques have long played an important role in social science, for instance, see [102], and can be used in many contexts such as community detection [103, 104]. Cliques in networks are the basis for analysis in terms of simplicial complexes as used in algebraic topology. In network analysis, simplicial complexes [105, 106, 107] have been used to analyse network geometry [108], to model structure in temporal networks [109], investigate synchronization phenomenon [110, 111, 112, 113], social contagion [114], epidemic spreading [115], and neuroscience [33, 34]. In addition, hypergraphs [116, 117, 118] naturally encode higher-order interactions in terms of fundamental units, their hyperedges represent the co-incidence of many nodes.

The third type of model is a non-markovian type of model, which relies on using high-resolution time series to forecast higher-order patterns in temporal networks, applications include temporal patterns in trade relations [119, 120],

In the current chapter, we design a combinatorial model which uses three-node interactions, the

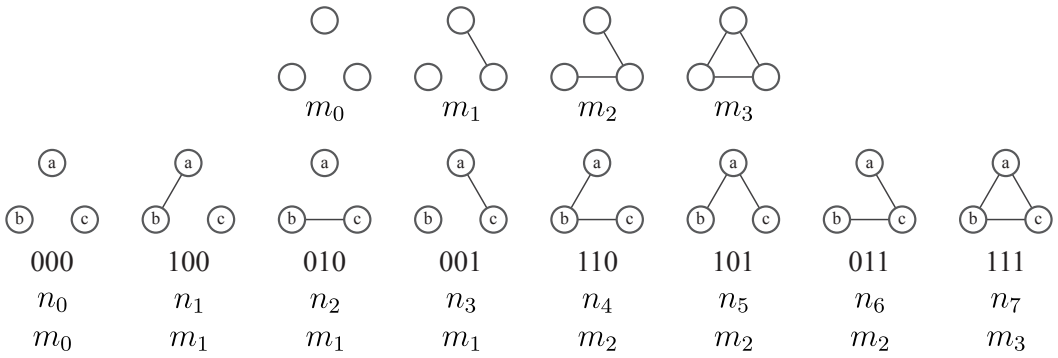


Figure 2.1: In the top row are the four unlabelled triplet states of $\mathcal{M}_4$, used only for illustration in this Chapter. The four states $m_i \in \mathcal{M}_3$ ($i = 0, 1, 2, 3$) are characterised by the number of links i in the given three-node graph. In the bottom row, there are the eight distinct states of triplets when nodes are distinct (labelled). Below each labelled diagram is the binary representation of the edge set, the label $n \in \mathcal{M}_8$ ($i = 0, 1 \ldots 7$), and the corresponding m_i triads if labels are ignored.

triplet. The triplet is the simplest unit of an interaction mechanism beyond pairwise. Triplets are configurations of three nodes in a network along with any edges between those nodes as shown in Figure 2.1. One area where network analysis is less well developed is the temporal evolution of networks. Many complex systems are not static so an important application of networks is to analyse their behaviour over time [121, 122, 123, 124, 125, 126, 46, 42]. Using higher-order interactions in an evolving network context has been considered in a few contexts [101, 114, 127, 128]. Most of the research on motifs in temporal networks focuses on how the number of each type of motif changes as the network evolves. On the other hand, the study of simplicial complexes is interested in how the fully connected triangles affect the structures of networks. So there is a gap between understanding how three-node combinations will evolve and how that evolution will affect the evolution of the whole network. This gap motivates us to design a method to investigate whether any non-pairwise interactions will be observed by measuring three-node dynamics. We will study higher-order processes in temporal networks through the evolution of the node triplet.

This chapter is organised in the following way: We start in section 2.2 by introducing the transition matrix that describes the Markovian evolution of the triplet. In section 2.4, we propose a null model that can demonstrate if higher-order interactions exist in networks. We use artificial networks with known higher-order interaction mechanisms to verify our approach is valid and use real networks to show the existence of higher-order interactions. In section 2.5,

based on the triplet evolution, we create an algorithm that can predict the existence of the links in the temporal networks and compare the performance of our method with other methods in section 2.5 & 2.6.

2.2 Quantifying non-pairwise interaction

2.2.1 Capturing dynamics of the temporal network via transition matrix

We start with temporal graphs $\mathcal{G}(s)$, a sequence of graphs with one node set $\mathcal{V}$ but with variable edges sets $\mathcal{E}(s)$, where s is a discrete time variable. The variable $\mathsf{M}_s(u,v,w)$, often abbreviated to M_s, records the state of a triplet (u,v,w) at time s. The states are the subgraphs equivalent to the three nodes and all the edges between them at time s. That is $\mathsf{M}_s \equiv \mathsf{M}_s(u,v,w)$ is a map from a node triplet (u,v,w), where $u,v,w \in \mathcal{V}$, to the induced graphlet, the maximal subgraph in $\mathcal{G}(s)$ containing the nodes u, v and w. We will use $\mathcal{M}$ to denote the set of all the possible three-node graphs and $m_i \in \mathcal{M}$ represents one of the possible graphs in $\mathcal{M}$. So formally

$$\begin{aligned} \mathsf{M}_s\colon\ \mathcal{V}\times\mathcal{V}\times\mathcal{V} &\to \mathcal{M}, \\ (u,v,w) &\mapsto \mathsf{M}_s(u,v,w) = m_i. \end{aligned} \tag{2.1}$$

The choice of $\mathcal{M}$ is not unique and it depends on the characteristics of the nodes and links used to distinguish configurations. If we characterise the states by the number of links among the three nodes, that is, use unlabelled graphs, there are just four distinct states in $\mathcal{M}$ which we can name as m_0, m_1, m_2 and m_3 as shown in the top row of Figure 2.1. So $\mathcal{M}_4$ is the state set characterised by the number of links and here m_i is the unlabelled graph of three nodes and i edges. We will use $\mathcal{M}_4$ for visualisation and illustration only in our work here.

By contrast, we want to consider the labels (identities) of the nodes in our work. So for our results we work with eight states in $\mathcal{M}$ since each link between the three pairs of nodes can

either be present or absent. That means the links between different pairs of nodes are distinct. This state set is represented by $\mathcal{M}_8$ as illustrated in the bottom row of Figure 2.1.

Typically one uses graphlets by counting the frequency of each graphlet $m \in \mathcal{M}$ in each graph $\mathcal{G}(s)$ in the temporal graph sequence. So a simple way to look at the evolution is to see how these counts change. We wish to go beyond this and look at the way local structure controls the local evolution. In order to do this we define a transition matrix T which describes the likelihood that a given triplet is to be transformed into another triplet in one time step. That is $\mathsf{T}_{ij}(s)$ gives the probability that a triplet of nodes in the state m_i in $\mathcal{G}(s-1)$ at time $s-1$ becomes the triplet m_j in $\mathcal{G}(s)$ at the next time step, s. That is

$$\mathsf{T}_{ij}(s) = \mathbb{P}(\mathsf{M}_s = m_j | \mathsf{M}_{s-1} = m_i)\,. \tag{2.2}$$

By definition, all entries of T are non-negative, $\mathsf{T}_{ij}(s) \geq 0$, and each row in the transition matrix satisfies a normalisation condition

$$\sum_j \mathsf{T}_{ij}(s) = \sum_{m_j \in \mathcal{M}} \mathbb{P}(\mathsf{M}_s = m_j | \mathsf{M}_{s-1} = m_i) = 1,\ \forall\ m_i \in \mathcal{M}\,. \tag{2.3}$$

In practice we have to use an estimate $\widehat{\mathsf{T}}(s)$ for the transition matrix $\mathsf{T}(s)$. We do this by using a subset $\mathcal{T}_{s-1}$ of all possible distinct node triplets so $\mathcal{T}_{s-1} \subseteq \mathcal{V}^3$. From this subset of node triplets, we then count how often the associated graphlet transforms from m_i to m_j. More formally we define

$$\widehat{\mathsf{T}}_{ij}(s) = \frac{1}{k_i} \sum_{(u,v,w)\in\mathcal{T}_{s-1}} \delta(\mathsf{M}_s(u,v,w), m_j)\ \delta(\mathsf{M}_{s-1}(u,v,w), m_i)\,, \tag{2.4}$$

and the normalisation constant is computed from:

$$k_i = \sum_j \sum_{(u,v,w)\in\mathcal{T}_{s-1}} \delta(\mathsf{M}_s(u,v,w), m_j)\ \delta(\mathsf{M}_{s-1}(u,v,w), m_i), \tag{2.5}$$

where $\delta(\mathcal{G}_1, \mathcal{G}_2) = 1$ (0) if graphlets $\mathcal{G}_1$ and $\mathcal{G}_2$ are isomorphic (not isomorphic). The best estimate $\widehat{\mathsf{T}}_{ij}(s)$ of $\mathsf{T}_{ij}(s)$ is produced if $\mathcal{T}_{s-1}$ is the set of all possible distinct node triplets. However, for a large graph with N nodes there are $\binom{N}{3} = N\cdot(N-1)\cdot(N-2)/6$ three node combinations

making it computationally inefficient to use all triplets. For example, for $N = 100,000$, $\binom{N}{3} = 1.7 \cdot 10^{14}$, Therefore, we will use random sampling of triplets of a sufficient amount to produce our estimates for $\widehat{\mathsf{T}}(s)$.

To illustrate the construction of a triplet transition matrix, we use the simpler $\mathcal{M}_4 = \{m_0, m_1, m_2, m_3\}$. Note that the subscripts of T, $i, j = 0, 1, 2, 3$ correspond to subscripts of states m_i in $\mathcal{M}_4$ shown in Figure 2.1. An example of the evolution of a network of 5 nodes and the triplet transition matrix is shown in Figure 2.2. Note we will use labelled subgraphs and $\mathcal{M}_8$ for our analysis.

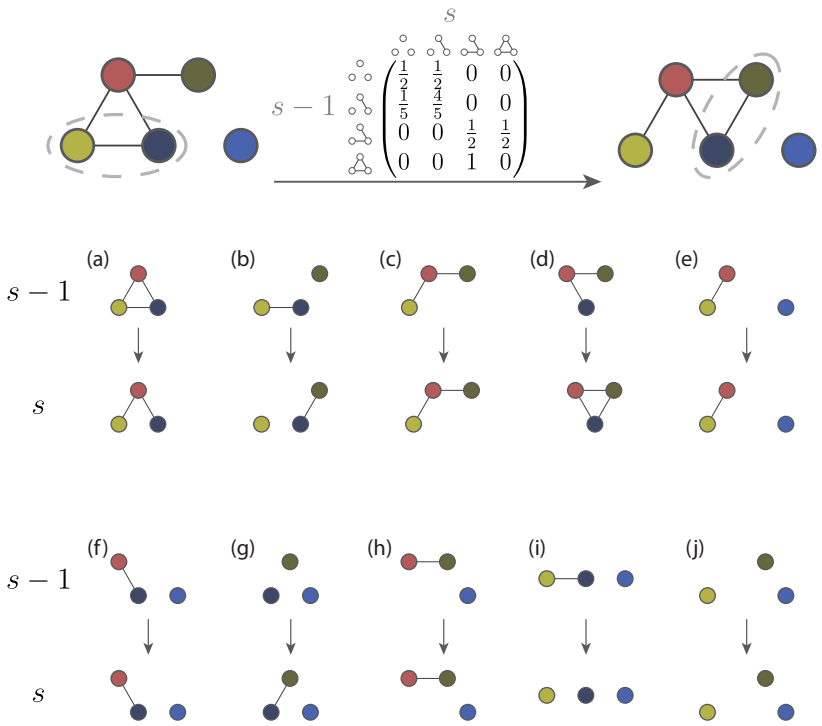


Figure 2.2: The empirical transition matrix on top centre can be computed from evolution of the network $\mathcal{G}(s-1)$ shown top left, with links at time $s-1$, to $\mathcal{G}(s)$ shown top right, with links at time s. When the states are only characterised by the number of links of the three-node combinations, $\mathcal{M}_4$, for illustrative purposes, there are four possible states being denoted by m_i ($i = 0, 1, 2, 3$) of Figure 2.1. The subscript i represents the number of the links i in the graphlet m_i, so m_0 is the triplet with zero links, m_1 is a triplet with one link and so on. There are $\binom{5}{3} = 10$ ways of choosing 3 nodes for a set of 5 nodes, and the evolution of all ten triplets are shown in the remaining rows in the figure above. For instance, the subgraph induced by a triplet (a) is the triplet m_3. This triplet loses an link in the next graphlet, so the same node triplet is now associated with the two link triplet m_2. This change, therefore, contributes to the $\widehat{\mathsf{T}}(s)_{32}$ entry. As this is the only induced subgraph (graphlet) in the earlier $\mathcal{G}(s-1)$ graph which is isomorphic to the m_3 triad, this means $\widehat{\mathsf{T}}(s)_{32} = 1$ while $\widehat{\mathsf{T}}(s)_{3j} = 0$ for $j = 0, 1, 3$. On the other hand, the double link m_2 triad appears twice as an induced subgraph of node triplets in $\mathcal{G}(s-1)$, namely (c) and (d). These triplets have two different triads in $\mathcal{G}(s)$ leading to two non zero entries in the $\widehat{\mathsf{T}}(s)_{2j}$ row, $\widehat{\mathsf{T}}(s)_{22} = \widehat{\mathsf{T}}(s)_{23} = 1/2$. Note we use $\mathcal{M}_8$ of Figure 2.1 in our analysis.

The memory needed for the calculations in our Triplet Transition (TT) method scales with the number of combination of triplet in network, that is $\binom{N}{3} = N(N-1)(N-2)/6 \sim O(N^3)$ and this can be seen in Figure 2.3.

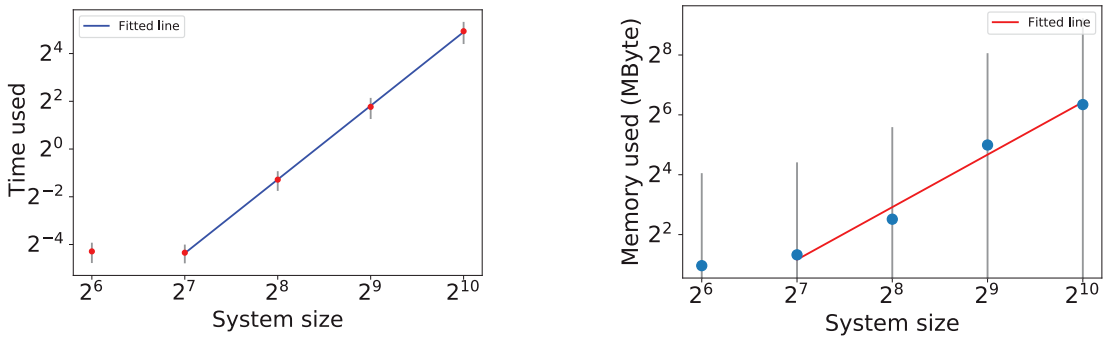


Figure 2.3: Time and memory needed for generating all three node graphlet combinations against a different number of nodes $N = 2^n$ $n = 6, 7, 8, 9, 10$. For each n, the time and memory for each run are measured and the error bars are corresponding to standard deviation. The time used scales as expected order (the slope of the fitted line is 3.09 ± 0.02) of the system size. The slope of the fitted line for the memory used is 1.75 ± 0.02.

To ensure the number of triplets is sufficiently large to estimate the transition matrix $\widehat{\mathsf{T}}(s)$, we use all three-node combinations if the number of nodes in the system is smaller than 10^3; otherwise, we sample 10^5 triplets chosen uniformly at random from the set of all three-node combinations. If the calculated $\widehat{\mathsf{T}}(s)$ is stable, we choose this sample number for the following analysis. Otherwise, we continue to sample a further 10^5 triplets until the $\widehat{\mathsf{T}}(s)$ is stable. Stable means that the maximum difference between two estimated transition matrix $\widehat{\mathsf{T}}(s)$ is less than 0.01.

2.3 Data sets description

Before we look at the results, we briefly discuss the datasets first. We use several different datasets to produce temporal networks with different resolutions. The resolution of a network is the time interval used to create each snapshot $\mathcal{G}(s)$ of our network. This can be done in two ways, depending on the context. In either case, the resolution can be 'seconds', 'hours', 'days' or 'months' and should be chosen to suit the context.

In the first type of temporal data, the data capture interactions between pairs of nodes which occur briefly on the time scale of the network resolution. A list of the times of phone calls between members of a social network would be an example of this type of data. In this case, each edge in a single snapshot indicates that an event linking the two nodes occurred during the time interval.

The second approach is where the pairwise interactions recorded in the data typically last for much longer than the resolution, but they do change slowly over time. An example of this would be hyperlinks between webpages. In this second case, the snapshots are the network at one instant in time, and the interval is now the time between these snapshots.

We use five different data sets which are as follows.

- `Turkish Shareholder Network` (`Shareholder`).

  The nodes are shareholders in Turkish companies. The shareholders are linked if they both hold shares in the same company during the time interval associated with the snapshot [40].

- `Wikipedia Mathematician` (`WikiMath`). Each biographical Wikipedia page of an individual mathematician forms a node. If a hyperlink links two biographies (in either or both directions), a link is present in the network. The edges are edited by users and are both added and removed over time. Each snapshot represents the state of these webpages at one moment in time. The data is taken at one point in three different years, 2013, 2017, and 2018, so the intervals are not constant in this case. See [79] for further details on this dataset.

- `College Message` (`CollegeMsg`). The nodes are students, and an edge in a snapshot indicates that the students exchanged a message within the interval associated with that snapshot. The data was collected over a seven month period in 2004, see [80, 81, 82] for more details.

- `Email` (`Email`). This is derived from the emails at a large European research institution sent between October 2003 and May 2005 (18 months). Each node corresponds to an

email address. An edge in a given snapshot indicates that an email was sent between the nodes in the time interval corresponding to that snapshot [82, 83, 84].

- `Hypertext` (`Hypertext`). This is the network of in-person face-to-face contacts of the attendees of the Association of Computing Machinery (ACM) Hypertext 2009 conference. In the network, a node represents a conference visitor, and an edge represents a face-to-face contact that was active for at least 20 seconds [83, 39].

  For temporal networks based on real data, the actual time interval between consecutive graphs in our temporal network, between $\mathcal{G}(s-1)$ and $\mathcal{G}(s)$, is given in real units as Δt. In some data sets we are able to look at the same data with different real time intervals between data sets. We provide a summary of the graph statistics in Table 2.1:

Dataset	(Abbreviation)	Nodes (N)	Edges (E)	Time Period (T)	Resolutions (Δt)	Source
Turkish Shareholder	(`Shareholder`)	39901	68017	2010,2012,2014,2016	2 years	[40]
Wikipedia Mathematician	(`WikiMath`)	6049	36315	2013,2017,2018	1 and 4 years	[79]
College Message	(`CollegeMsg`)	1899	20296	6 months, 2004	7 days, 1 month	[80, 81, 82]
Institution Email	(`Email`)	986	24929	1 year, 1970-1971	8 hours, 7 days, 1 month	[82, 83, 84]
Hypertext	(`Hypertext`)	113	5246	3 days in 2009	40, 60 min	[83, 39]

Table 2.1: The detailed information of graph statistics. Nodes and edges columns correspond to number of nodes or edges in networks, where the edge number is the number of the events occurred. The number can be different from edge numbers in a networks, since an edge (event) may occur multiple times.

2.4 Evidence for higher-order interactions

To investigate the effectiveness of three-node interactions, we start by considering a "pairwise model" whose dynamics is driven only by pair-wise relationships. This will act as null models when analysing the transition matrix for artificial and real-world networks.

2.4.1 A benchmark pairwise model

The pairwise model is a stochastic graph model for dynamic networks [208] in which the evolution of the edge is based on comparison of the evolution of the edge between pairs of nodes, so a two-node graphlet model. In the pairwise model, moving from one network, $\mathcal{G}(s-1)$ to the next network in the sequence, $\mathcal{G}(s)$, any pair of nodes with no existing link gains a link with probability p otherwise with probability $(1-p)$ the pair of nodes remains unconnected. Similarly, every existing link $e \in \mathcal{E}(s-1)$ is removed with probability q otherwise with probability $(1-q)$ the link remains. The number of links is not preserved in this model. This null model is a lower order description of the local interactions than our full analysis in terms of triplets. It is straightforward to write down the form of the transition matrix in our triplet based analysis when assuming this pairwise model gives a precise description, giving us a two-parameter transition matrix denoted as $\mathsf{T}^{(\mathrm{pw})}(s)$. If we assume that graph evolution follows this pairwise mechanism, we can estimate values $\hat{p}(s)$ and $\hat{q}(s)$ for the parameters p and q respectively by looking at how links changed over one time step, i.e., from the edge set $\mathcal{E}(s-1)$ in $\mathcal{G}(s-1)$ to edge set $\mathcal{E}(s)$ of $\mathcal{G}(s)$. Formally we have that

$$\hat{q}(s) = 1 - \frac{|\mathcal{E}(s-1) \cap \mathcal{E}(s)|}{|\mathcal{E}(s-1)|}, \tag{2.6}$$

$$\hat{p}(s) = \frac{|\mathcal{E}(s) \setminus (\mathcal{E}(s-1) \cap \mathcal{E}(s))|}{N(N-1)/2 - |\mathcal{E}(s-1)|}. \tag{2.7}$$

This gives us our pairwise model prediction for the triplet transition matrix $\widehat{\mathsf{T}}^{(\mathrm{pw})}(s)$, where we substitute $\hat{p}(s)$ from Eq.(2.7) and $\hat{q}(s)$ from Eq.(2.6) for p and q in $\mathsf{T}^{(\mathrm{pw})}$. That is $\widehat{\mathsf{T}}^{(\mathrm{pw})}(s) = \mathsf{T}^{(\mathrm{pw})}(\hat{p}(s), \hat{q}(s))$ where

$$\mathsf{T}^{(\mathrm{pw})}(p,q) = \begin{pmatrix} (1-p)^3 & 3p(1-p)^2 & 3p^2(1-p) & p^3 \\ q(1-p)^2 & (1-q)(1-p)^2 + 2qp(1-p) & 2p(1-p)(1-q) + qp^2 & (1-q)p^2 \\ q^2(1-p) & 2(1-q)q(1-p) + q^2p & (1-q)^2(1-p) + 2qp(1-q) & (1-q)^2p \\ q^3 & 3(1-q)q^2 & 3q(1-q)^2 & (1-q)^3 \end{pmatrix}. \tag{2.8}$$

2.4.2 Artificial networks

Apart from the pairwise model we introduced in the last section. The second stochastic model is "edge swap model", we swap the ends of a pair of edges so edges (u, v) and (w, x) in snapshot $s-1$ are removed and are replaced by edges (u, x) and (w, v) in the next snapshot. This preserves the degree of every node. For each update from $\mathcal{G}(s-1)$ to $\mathcal{G}(s)$ we update 20% edges.

The final third model is "random walk model". It is an edge rewiring model but it is based on higher-order structures as we use random walks to select the new edges. The model starts with an Erdős-Rényi graph. The initial node for a random walker, say u, is chosen uniformly at random from the set of nodes. Then one of the edges from u, say the edge to a node y is chosen uniformly from the set of neighbours. Finally, three non-backtracking steps are made on the network starting from u and ending at a node x, in which edges are always chosen uniformly from those available excluding any edge used in the previous step of the random walk. The existing edge (u, y) is removed and replaced by a new edge (u, x). The final graph is then created through a projection where nodes are connected if they share a common neighbour, that is if directed edges from u to v and w to v exists, then the projected graph has an undirected link between u and w. This rewiring and projection procedure maintains the number of edges and nodes in the original graph but not in the projected graph.

To produce the time evolution, we rewire 20% of the edges using this rewiring procedure and then use this new network as the next snapshot. In our context, our random walk model is used to produce test networks with local correlations between nodes. While motivated by real-world examples and building on existing experience with the model [40], here it is used as a toy model to illustrate the approach.

We use these three simple models to generate artificial temporal networks to test our approach. The results in terms of the transition matrix $\widehat{\mathsf{T}}$ derived from the artificial networks are shown in Figure 2.4. To show the deviation from the our pairwise interaction null model, we look at the difference $\Delta\widehat{\mathsf{T}}(s)$ between actual results for the average of $\widehat{\mathsf{T}}(s)$ and those predicted in the

null model $\widehat{\mathsf{T}}^{(\mathrm{pw})}(s)$, so

$$\Delta\widehat{\mathsf{T}}(s) = \widehat{\mathsf{T}}(s) - \widehat{\mathsf{T}}^{(\mathrm{pw})}(s)\,. \tag{2.9}$$

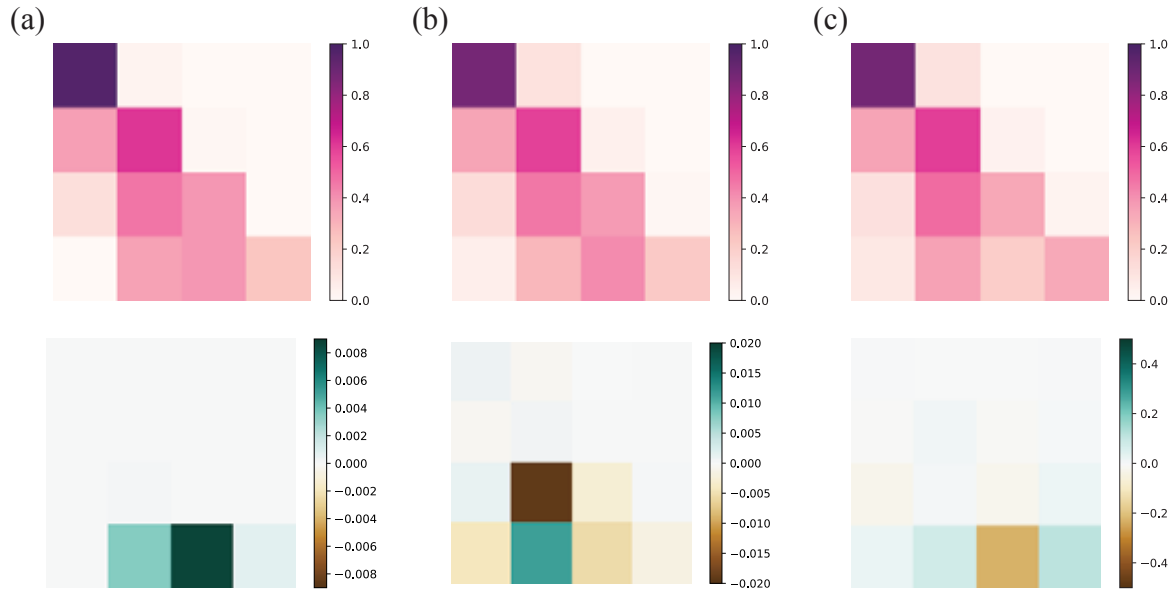


Figure 2.4: The triplet transition matrix $\widehat{\mathsf{T}}$ estimated from the artificial data. The three columns of heat maps show results for the three different artificial temporal networks created numerically using the stochastic models: (a) "pairwise model", (b) "edge swap model", and (c) "random walk model". Each of the six four-by-four heat maps shows values where the rows represent one initial triplet graphlet m_i and the columns representing the final graphlet m_j using our unlabelled graphlet $\mathcal{M}_4$ set of triplets to make the visualisations manageable. There are multiple snapshots of networks in each data-set, therefore, we use average to avoid statistical insignificant event. The top three heat maps show the average value of entries $\langle\widehat{\mathsf{T}}_{ij}\rangle$, which are calculated from all snapshots of temporal network. The bottom row of three heat maps give the values of the difference matrix $\langle\Delta\widehat{\mathsf{T}}\rangle = \langle\widehat{\mathsf{T}}^{(\mathrm{pw})}\rangle - \langle\widehat{\mathsf{T}}\rangle$, which is average of Eq.(2.9) from multiple realisations in which the numerical data is compared to the analytical form predicted from the simple pairwise model. The scales for the colours of the first two heat maps on the bottom row are much smaller (a factor of twenty or more) than for the results for $\langle\Delta\widehat{\mathsf{T}}\rangle = \langle\widehat{\mathsf{T}}^{(\mathrm{pw})}\rangle - \langle\widehat{\mathsf{T}}\rangle$ in the random walk model shown in the bottom right corner. Any large entries in these lower rows of heatmaps indicate higher-order effects not present in our simple pairwise model of $\widehat{\mathsf{T}}^{(\mathrm{pw})}$. Only the networks formed with random walks show significant higher-order effects.

The behaviour of our simple "pairwise model" should be completely captured by the reference transition matrix $\widehat{\mathsf{T}}^{(\mathrm{pw})}(s)$ and, as expected, the numerical results shown in Figure 2.4(a) show no significant difference between numerical data $\widehat{\mathsf{T}}(s)$ and theoretical $\mathsf{T}^{(\mathrm{pw})}$, that is, $\Delta\widehat{\mathsf{T}}(s)$ is small.

For Figure 2.4(b), we use the artificial networks generated by the "edge swap model". While this involves two pairs of edges, so in principle is a higher-order model, in a sparse graph, the

four nodes selected by a pair of randomly selected edges are unlikely to be linked by other edges. So in practice, in terms of the triplet graphlets, this model behaves much like the pairwise model and shows little difference from that model.

It is only with the networks generated using our three-step random walk that we see significant differences between the data and the pairwise model. This is to be expected as higher-order processes were used to create the numerical networks, see Figure 2.4(c).

2.4.3 Quantifying non pairwise interactions

We then apply this framework to analyse some real data sets. For clarity we show results for our triplet transition matrix Eq.(2.9) when working with $\mathcal{M}_4$ and these are shown in Figure 2.5. These results for real-world systems have significant non-zero entries in our measure $\Delta\widehat{\mathsf{T}}(s)$. These results for $\Delta\widehat{\mathsf{T}}(s)$ are also very different from each other, reflecting distinct mechanisms behind the evolution of these systems.

When we look for higher-order interactions, we find clear differences between the triplet transition matrix $\widehat{\mathsf{T}}$ and the simple pairwise reference model of $\widehat{\mathsf{T}}^{(\mathrm{pw})}$, especially in the Turkish Shareholder network Figure 2.5(a). For example, compared to the corresponding probability in the pairwise model, the real probability of any triplet state becoming disconnected in the next snapshot (i.e., moving from $m_i \to m_0$, $i \in \{0,1,2,3\}$, the first column of the transition matrix) is much less, showing that this subgraph is very stable compared with the pairwise case in the Turkish Shareholder network. Additionally, the state is much more likely to evolve to m_1 at $s+1$ (the second column of the transition matrix), which demonstrates that in many real networks, the interactions are beyond the pairwise interaction. In the Turkish shareholder network, most shareholders generally only exist in the market for only one snapshot of network. And many of them share similar investing behaviours with some financial institutions. Since financial institutions invest in many companies and their shareholding behaviours are quite long time, thus, the links between institutions and survivors are much more likely.

In our analysis of real data, we use Δt to denote the physical time difference between snapshots

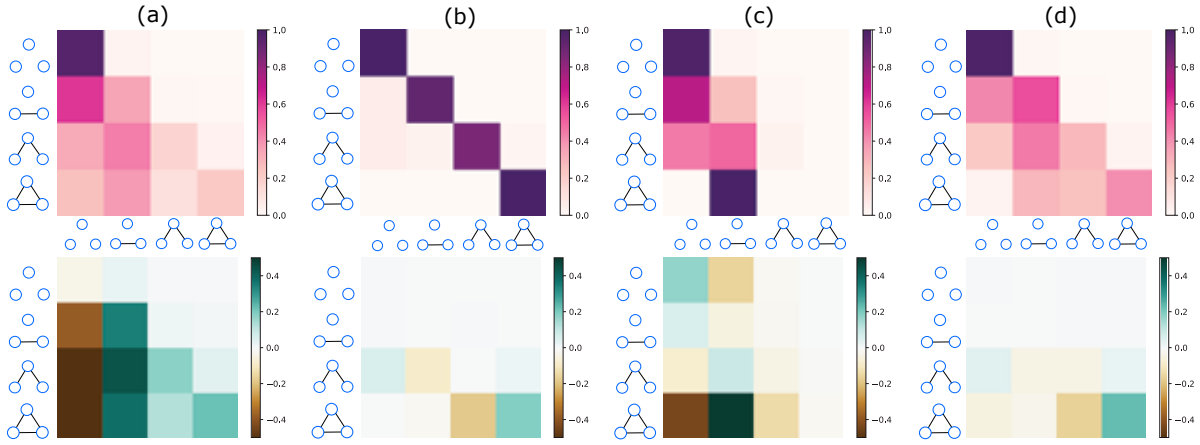


Figure 2.5: The Triplet transition matrix $\widehat{\mathsf{T}}$ evaluated for real world networks. Each column of heat maps is for a different data set; from left to right we have: (a) Turkish Shareholder network (Δt = 2yr), $p = 3.49\times10^{-2}$, $q = 0.638$, (b) Wikipedia Mathematician network (Δt = 1yr), $p = 3.19\times10^{-4}$, $q = 0.068$, (c) College Message network (Δt = 1mo), $p = 0.067$, $q = 0.510$, and (d) Email network (Δt = 1mo), $p = 9.43\times10^{-3}$, $q = 0.432$. The actual time difference between snapshots $\mathcal{G}(s-1)$ and $\mathcal{G}(s)$ is Δt. The average values $\langle\widehat{\mathsf{T}}\rangle$ from Eq.(2.4) are shown on the top row while the average of differences from the simple pairwise model, $\langle\Delta\widehat{\mathsf{T}}\rangle$ of Eq. (2.9), are shown on the bottom row. Each of the four-by-four heat map grids is organised in the same way as the $\widehat{\mathsf{T}}$ matrix shown in Figure 2.2. That is, the rows indicate the initial triplet state (m_0 to m_3 from top to bottom as indicated) and the columns indicate the final triplet state (m_0 to m_3 from left to right as indicated). The values of $\langle\widehat{\mathsf{T}}\rangle$ on the top row run from 0 (white) to 1.0 (dark red). For $\langle\Delta\widehat{\mathsf{T}}(s)\rangle$ on the bottom row, values run from dark brown (-0.5) through white (0.0) to dark blue (+0.5).

$\mathcal{G}(s)$ and $\mathcal{G}(s+1)$. This time interval can be varied when working with the College message data of Figure 2.5(c) and the Email network of Figure 2.5(d). In these cases we try several values of Δt before choosing an appropriate value the the illustrations in Figure 2.5. We are looking for a value that is not too short (nothing much happens) and not too long (averaging over uncorrelated changes).

Unsurprisingly, all four real world networks show considerable higher-order effects but there are interesting differences that reveal the processes behind the evolution are likely to have some significant differences. The $\widehat{\mathsf{T}}$ results are shown in the top row of Figure 2.5. The result for the Turkish Shareholder network in Figure 2.5(a) and the Email network in Figure 2.5(d) look very similar. However, when we compare them to our reference model, the pairwise model, we see large differences, showing that these are very different types of temporal network and using raw values of $\widehat{\mathsf{T}}$ can be misleading.

In fact, the networks which are most similar, once simple pairwise processes are subtracted from original process. In the Wikipedia Mathematician network Figure 2.5(b) and the Email network Figure 2.5(d), both processes where an edge is added (upper triangle in the heat map) there is little difference from the pairwise model. This might be expected for the cases where there is at most one edge in the triplet. Only the $\Delta\widehat{\mathsf{T}}_{23}$ entry for the $m_2 \to m_3$ transition shows a slight increase over what would be predicted based on the rate of edge addition in these models (the p parameter of the pairwise model) which suggests triadic closure ($m_2 \to m_3$) [30, 31, 129, 130] does play a role here but it is very slight by these measures. On the other hand, there is a strong sign of "triadic stability", that is the complete graphlet m_3 is much more stable than we would expect given the rate of edge loss in the pairwise model (the q parameter). It is natural to think that processes responsible for triadic closure would also slow the rate at which such triangles break up ($m_3 \to m_2$) but we do not see this in our result. One interpretation of our results is that the social processes normally invoked for triadic closure [30, 31, 129, 130] can, in some cases, be more important in preventing the breakup of triangles than in the creation of triangles.

In some ways the similarity between the Wikipedia Mathematician network Figure 2.5(b) and

the Email network Figure 2.5(d) is surprising as we might have expected the greatest similarity between the two communication networks, the College Message network Figure 2.5(c) and the Email network Figure 2.5(d) but that is not what we see in our measures. In particular, the stability of the complete triangle in the college message is exactly as we would expect based on pairwise measures suggesting different properties in these two communication networks, i.e., that many of the college messages are between actors who do not have strong ties.

However, the main message in Figure 2.5 is that in almost any data, we find the evolution has clear signals of higher-order interactions playing an important role.

2.4.4 Significance test of the pairwise interactions

The significance of the transition of the triplet transition can be quantified by using the z-score (standard score). The z-score has been used as a qualitative measure of statistical significance of different motifs [37] or temporal motifs [211]. We apply similar procedures to compute z-scores for different triplet transitions and for a transition from graphlet m_i to graphlet m_j we define

$$Z_{ij} = \frac{\langle \widehat{\mathsf{T}}_{ij}(s) \rangle - \langle \widehat{\mathsf{T}}_{ij}^{(\mathrm{pw})}(s) \rangle}{\sigma_{ij}^{(\mathrm{pw})}} . \tag{2.10}$$

Here $\sigma_{ij}^{(\mathrm{pw})}$ is the standard deviation in the ij-th entry of the transition matrices $\widehat{\mathsf{T}}^{(\mathrm{pw})}(s)$ obtained from the simple pairwise model.

To calculate the z-score, we generate an ensemble of $R = 1000$ realisations of the null model, the pairwise null model. The number of simulations R needed was found as follows. We start from $R = 100$ realisations, increase to 200, 300 and so on. Each time we increase R we compare the difference in the results to those found with the $(R - 100)$ realisations; if the results of the transition counting do not change about 0.1% from $(R - 100)$ to R, we assume R is sufficient, and we stop increasing R. Otherwise, we continue to increase the number of realisations until the results do not change. The values of p and q used in the calculation of $\widehat{\mathsf{T}}_{ij}^{(\mathrm{pw})}(s)$ are those inferred from real world networks as described in Figure 2.5. A z-score Z_{ij} with absolute value much bigger than one shows us that the data has behaviour not accounted for in our simple

pairwise model. So it quantifies how likely a specific transition from state i to j is derived from higher-order processes not captured by simple pairwise interactions. The results for Z_{ij} for some of our networks are shown in Figure 2.6.

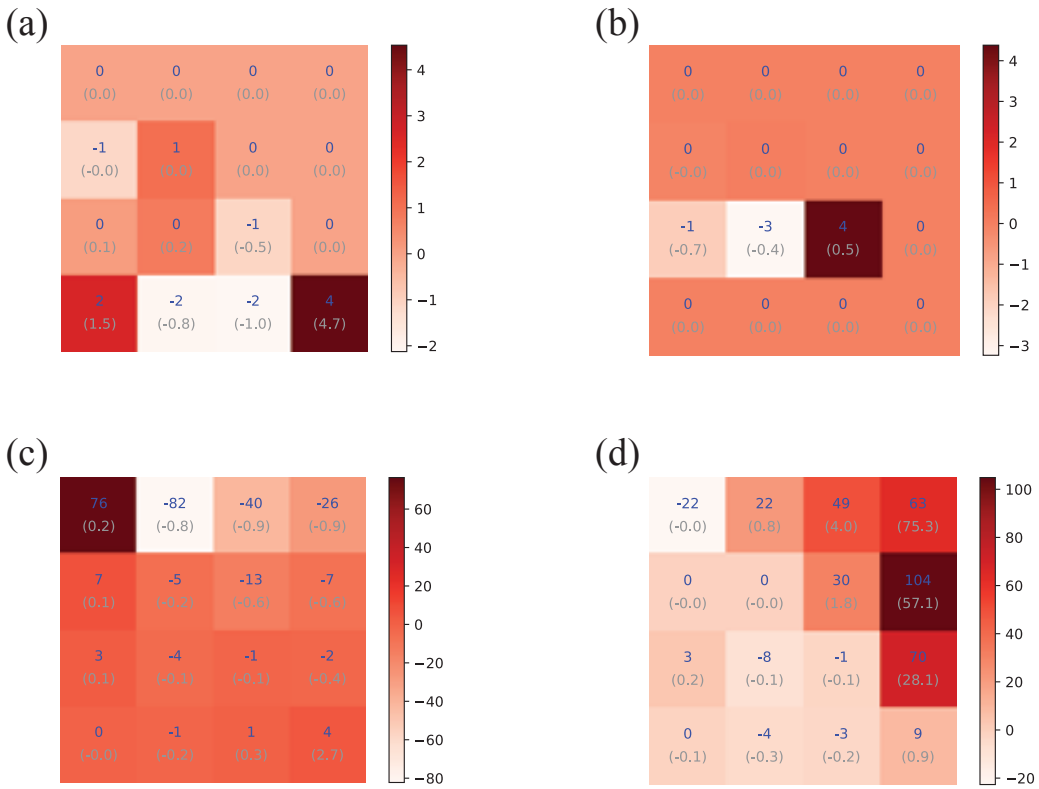


Figure 2.6: Each of the large four squares corresponds to a Z matrix from Eq.(2.10) for a different data source, labelled as follows: (a) Turkish Shareholder network (Δt = 2yr), (b) Wikipedia Mathematician (Δt = 1yr), (c) College Message data (Δt = 1mo), and (d) Email network data (Δt = 1mo). Each large square represents a four by four grid of smaller squares. The rows (columns) are the triplets at the earlier (later) arranged in order of size from smallest m_0 to largest m_3 going top to bottom (left to right). In each, the colour represents Z_{ij} on the scale given on the right of each large square, the top blue digits represent Z-scores, and the bottom grey digits in brackets gives $(N^{\mathrm{r}}(m) - N^{\mathrm{pw}})/N^{\mathrm{pw}}$, where N^{r} gives the number of triplets in the real data and N^{pw} gives the number predicted in the simple pairwise model for the same sized sample.

2.5 Link prediction

The analysis of our transition matrix of triplets reveals that non-pairwise interaction exists in network evolution. We now ask if these higher-order interaction patterns are essential for network evolution. We investigate this question by performing link predictions for dynamic

networks based on the higher-order information stored in our triplet transition matrices.

2.5.1 Triplet transition score

The idea behind our algorithm is that the formation or removal of links between a node pair is encoded in our triplet interactions. The likelihood of a link appearing (or disappearing) between a node pair in a snapshot can be obtained by looking at the triplets containing this node pair and using the triplet transition matrix to see what that suggests about the evolution of any edge between our chosen node pair.

To keep our analysis simple we will assume that the interval between snapshots Δt is about the same size as the appropriate time scale for changes in the network. If we look at snapshots covering an extremely short period of time, say one that covers the typical difference in time between events, then the transition matrix contains too little information. In such a case one could then look at larger times scales by using $\widehat{\mathsf{T}}(s), \widehat{\mathsf{T}}(s-1), \cdots, \widehat{\mathsf{T}}(s-n)$ (for some n) to predict links in the next network snapshot $\mathcal{G}(s+1)$. However, we will use the simpler assumption that working with one snapshot by choosing Δt appropriately (where we have that choice) captures the essential information. This implies we have avoided the opposite problem, where Δt is too large so the information in the transition matrix is averaging over mostly uncorrelated and unrelated events which means we have very little signal encoded in our transition matrices.

Our second assumption is that the transition depends only on the state of the system at the last time step, which means we assume the process is Markovian. In some ways this need not be completely true. Provided the snapshots $\mathcal{G}(s-1)$ and $\mathcal{G}(s)$ are in a similar state to the one we are trying to predict, $\mathcal{G}(s+1)$, then because we are constructing our transition matrices based on the similar states, the history of the evolution may well be encoded in this. For instance, patterns of activity in an email network of a large institution could well change if there is a major reorganisation with many people changing roles or locations in which case the history of the system has an impact on the evolution. Put another way, we assume that any non-Markovian behaviour is happening on much larger time scale than we are studying and so we can use a Markovian approximation.

Our link prediction algorithm assigns a score to each node pair. By looking at the distribution of these scores we can separate them into node pairs with low scores, predicting no link in the next snapshot, or a high score meaning this node pair will be connected. For a given snapshot $\mathcal{G}(s)$, for each node pair, say (u, v), we look at all $(N-2)$ triplet pairs and count the number of each triplet falling in each state $m_i \in \mathcal{M}$. Normalising this gives us our node-pair state vector $\psi_i(u, v; s)$ as following:

$$\psi_i(u, v; s) = \frac{1}{N-2} \sum_{w \in \mathcal{V} \setminus \{u,v\}} \delta(\mathsf{M}_s(u, v, w), m_i)\,. \tag{2.11}$$

Our estimate for the transition matrix $\widehat{\mathsf{T}}(s)$, based on $\mathcal{G}(s-1)$ to $\mathcal{G}(s)$ evolution, is to be used to tell us about the evolution of this triplet state distribution $\psi_i(u, v; s)$ in Eq. (2.11). For instance we estimate that the probability L_β that a pair of nodes u, v has an link, $\beta = 1$, or no link, $\beta = 0$, in the graph $\mathcal{G}(s+1)$ is given by the projection from predicted $\psi_i(u, v; s)$:

$$L_\beta(u, v; s+1) = \sum_{i,j} \psi_i(u, v; s)\, \widehat{\mathsf{T}}_{ij}(s)\, P^{(\mathrm{out})}_{j\beta}, \tag{2.12}$$

where $P^{(\mathrm{out})}_{j\beta}$ is an entry of projection vector for the triplet evolution and $\sum_{\beta=0}^{1} P^{(\mathrm{out})}_{j\beta} = 1$. With the other definition $\sum_j \widehat{\mathsf{T}}_{ij} = 1$, and $\sum_i \psi_i(u, v; s) = 1$, then we have that L_β is properly normalised $\sum_{\beta=0}^{1} L_\beta(u, v) = 1$. It is this L_β in Eq.(2.12) that we use as a score for link prediction.

Take the case of $\mathcal{M}_4$ as an example, the projection from triplet distribution onto links uses

$$\mathsf{P}^{(\mathrm{out})} = \begin{pmatrix} 1 & 0 \\ 2/3 & 1/3 \\ 1/3 & 2/3 \\ 0 & 1 \end{pmatrix}. \tag{2.13}$$

The factors here arise for the state set $\mathcal{M}_4$ since the unlabelled graphlets do not distinguish which links are occupied in the one-link triplet m_1 or two-link triplet m_2. In m_3 states, link will appears between all three nodes, thus, the projection vector is always equal to one. For the state set $\mathcal{M}_8$ which we use in most of our work, this projection matrix is much simpler

with entries either 0 or 1. If we choose (u, v) to be the first link, so it is the only link in the triplet we call n_1 in Figure 2.1, then we can use bitwise logical operators to represent $\mathsf{P}^{(\mathrm{out})}_{j\beta}$ as (j AND 1) XOR β.

2.5.2 Node similarity

We are using link prediction as a way to test that our triplet transition matrices $\widehat{\mathsf{T}}$ capture important higher-order interactions in the evolution of temporal networks. In order to see how effective our approach is, we need to compare against other methods of link prediction. All the methods we use are listed in Table 2.2. All these methods can be discussed in terms of a node similarity score and in this section we will start our examination of these methods by considering the node similarity scores used in each method. This will allow us to examine the relationships between these various methods and ask if they capture higher-order interactions to any extent. The next stage is to turn these similarity scores into a link prediction and we will look at this step in section 2.5.3.

Abbreviation	*Method*[Reference]	Length Scale	`Code`
AAI	*Adar-Academic Index*[131]	2	`nx`
CN	*Common Neighbour*[131]	2	`nx`
EE	*Edge Existence*	1	`-`
JC	*Jaccard Coefficient*[131]	2	`nx`
Katz	*Katz*[132]	∞	`own`
LLHN	*Local Leicht-Holme-Newman*[139]	2	`own`
LPI	*Local Path Index*[138]	3	`own`
MFI	*Matrix Forest Index*[140]	∞	`own`
PA	*Preferential Attachment*[131]	2	`nx`
RAI	*Resource Allocating Index*[138]	2	`nx`
TT	*Triplet Transition*	∞	`own`

Table 2.2: The length scale given indicates the longest path length involved in the method or equivalently the largest power of the adjacent matrix involved in the method. Under `code`, `nx` indicates that a `NetworkX` [209] library routine was used, while `own` indicates the authors' own code was used. The Edge Existence (EE) approach was not investigated numerically but was included for the sake of comparison. All these methods, except for the Triplet Transition method, have a temporal path length of 0. That is they are derived solely from the current snapshot, $\mathcal{G}(s)$, when making a prediction for links in the next snapshot $\mathcal{G}(s+1)$. The Triplet Transition method has a temporal path length of 1 as it uses $\mathcal{G}(s-1)$ and $\mathcal{G}(s)$ to predict $\mathcal{G}(s+1)$.

The various link prediction methods can be categorised based on the type of information used to make a prediction about each node pair. Those methods which use only local interactions probing a fixed distance from the node pair, and those global methods which use nodes arbitrarily far from the node pair of interest. This is indicated by the length scale of methods in Table 2.2 and will be clear from the power of the adjacency matrix A appearing in the similarity scores defined in this section.

One other point can be made. In terms of the temporal network, none of these existing similarity scores, none of the these link prediction methods from the literature, use more than one temporal snapshot $\mathcal{G}(s)$. So one standout feature of our method is the use of two temporal snapshots, $\mathcal{G}(s-1)$ and $\mathcal{G}(s)$. The evolution of a triplet from one snapshot to another records, in an indirect way, the effects of other nodes beyond the triplet of interest as illustrated in Figure 2.7. This is why our method is listed as probing long length scales in Table 2.2.

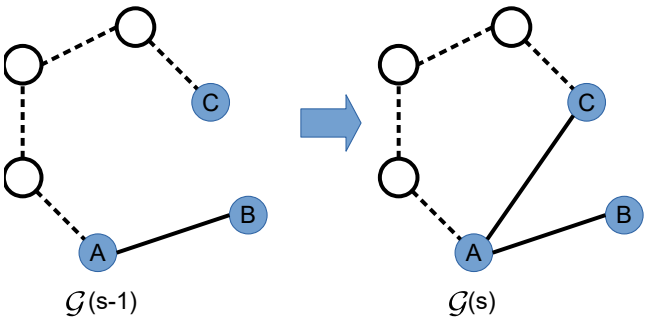


Figure 2.7: An illustration of how information on long range paths between nodes plays a role in our Triplet Transition method. This shows the evolution of one m_1 triplet in snapshot $\mathcal{G}(s-1)$ to become an m_2 triplet in the next snapshot. The triplet nodes are shown as solid blue circles connected by solid black edges. The network outside the triplet is shown with nodes as open circles connected by edges shown as dashed lines. It is the network outside the triplet which provides a long distance connection between triplet nodes A and C. Within the original ABC triplet in snapshot $\mathcal{G}(s-1)$, C appears disconnected. By using two snapshots, our transition matrix includes the effects of such long-range links and so this approach can explain why the edge (A, C) might appear more often than otherwise expected from the triplet subgraph alone.

In the following $s(u, v)$ is a (similarity) score assigned to a pair of vertices $u, v \in \mathcal{V}$ where $\mathcal{N}(u)$ is the set of neighbours of vertex u, i.e. $\mathcal{N}(u) = \{w|(u, w) \in \mathcal{E}\}$. The similarity scores are then used to decide if an edge should exist between u and v and that will be used in turn to make the link prediction for the next snapshot in time according to the rank of scores. We will assume a simple graph in these discussions.

The simplest node similarity measure is the *Edge Existence* (EE) index $s_{\mathrm{EE}}(u,v)$. After all, if there is already an edge between two vertices u and v this is a good indication of a close relationship between these nodes, and vice versa. We define Edge Existence index to be

$$s_{\mathrm{EE}}(u,v) = \sum_{e\in\mathcal{E}} \delta_{e,(u,v)} = A_{uv}\,. \tag{2.14}$$

Every edge between u and v adds one to this index so, for a simple graph, this is the corresponding entry of the adjacency matrix, A_{uv}. This is the simplest edge prediction method in that it predicts no changes at all so it is not very useful and we do not use it in our work. However, we will see this score contributes to some of the other, more sophisticated methods, so it is useful to define this score. The Edge Existence score records no higher-order effects, there is a path of length at most one between the nodes.

While not very useful, this *Edge Existence* (EE) index could produce some very good statistics. For instance, if very few edges are changing in each time interval, then this method would predict the behaviour of the vast majority of edges as most remain unchanged. It would only do badly if we used statistics that specifically measured the success of node pairs changing their connectedness from one time step to the next. A good example of when the Edge Existence method would appear to be successful is when we break up the data into small steps in time where few edges change. However we would learn little of interest in this case.

The *Common Neighbours* (CN) method [131, 132, 133] simply scores the relationship between two nodes based on the the number of neighbours they have in common

$$s_{\mathrm{CN}}(u,v) = \sum_{w\in\mathcal{V}} A_{uw}A_{vw} = |\mathcal{N}(u)\cap\mathcal{N}(v)|\,. \tag{2.15}$$

This will tend to give large scores if u and/or v have high degrees.

One way to compensate for the expected dependence of s_{CN} on the degree of the nodes is to normalise by the total number of unique neighbours. This gives us the *Jaccard Coefficient* (JC) [131, 133] method based on the well known similarity measure [134] in which the likelihood that two nodes are linked is equal to the number of neighbours they have in common relative

to the total number of unique neighbours.

$$s_{\mathrm{JC}}(u,v) = \frac{|\mathcal{N}(u) \cap \mathcal{N}(v)|}{|\mathcal{N}(u) \cup \mathcal{N}(v)|}\,. \tag{2.16}$$

The *Preferential Attachment* (PA) method for link prediction [131, 133] is based on the idea that the probability of a link between two vertices is related to the product of their degrees of the two vertices

$$s_{\mathrm{PA}}(u,v) = |\mathcal{N}(u)| \cdot |\mathcal{N}(v)| = k_u \cdot k_v\,. \tag{2.17}$$

This is proportional to the number of common neighbours expected in the Configuration model [135] as has been used in the context of collaboration (bipartite) graphs [131, 136, 137].

The *Resource Allocating Index* (RAI) [138] method and the *Adamic-Adar Index* (AAI) [131, 133] method are both based on the idea that if two vertices u and v share some "features" f that is very common in the whole network then that common feature is *not* a strong indicator that the two vertices should be linked. The converse is true if the common feature is rare, that is a good indicator that the vertices u and v should be linked. So in general if $W(x)$ is a monotonically decreasing function of x we can use this on the frequency $n(f)$ of the occurrence of feature f to give a generic similarity function of the form $s_{\mathrm{W}}(u,v) = \sum_{f \in u,v} W(n(f))$. In our case, we will not assume any meta-data exists, but we will look for methods that use features which are based purely upon the topology of the network.

In the *Resource Allocating Index* [138], the inverse of the degree of the neighbour is used as the weighting function, so $W(n(f)) \equiv 1/|\mathcal{N}(w)|$ giving

$$s_{\mathrm{RAI}}(u,v) = \sum_{w \in \mathcal{N}(u) \cap \mathcal{N}(v)} \frac{1}{|\mathcal{N}(w)|} = \sum_{w} \frac{1}{k_w}\,. \tag{2.18}$$

Note that this means that the contribution from any one node w to the total of all scores $S_{\mathrm{RAI}}(w) = \sum_{u,v} s_{\mathrm{RAI}}(u,v)$ is half of the degree of w minus one, $(|\mathcal{N}(w)| - 1)/2$. Thus the Resource Allocating Index still gives high degree nodes more weight.

On the other hand, the *Adamic-Adar Index* [131, 133] uses the inverse logarithm of the degree

to weight the contribution of each common neighbour to the score, $W(n(f)) \equiv 1/\ln(|\mathcal{N}(w)|)$. That is

$$s_{\mathrm{AAI}}(u,v) = \sum_{w \in \mathcal{N}(u) \cap \mathcal{N}(v)} \frac{1}{\ln(|\mathcal{N}(w)|)} \,. \tag{2.19}$$

All the indices mentioned above have been based either on the degree of the two nodes of interest s_{EE} or on the properties of u, v and their common nearest neighbours $w \in \mathcal{N}(u) \cap \mathcal{N}(v)$. This involves paths between the two nodes u and v of length two or less. The next logical step is to include paths of length three and the *Local Path Index* (LPI) s_{LPI} [132, 138] is an example of this where

$$\begin{aligned} s_{\mathrm{LPI}}(u,v) &= [\mathsf{A}^2 + \beta \mathsf{A}^3]_{uv} \\ &= s_{\mathrm{CN}}(u,v) + \beta [\mathsf{A}^3]_{uv} \,. \end{aligned} \tag{2.20}$$

Here β is a real parameter where $\beta = 0$ reproduces the Common Neighbours score of Eq.(2.15). The $\beta \mathsf{A}^3$ term is counting the number of walks of length three that start at u and end at v. If there is already an edge between u and v then $[\mathsf{A}^3]_{uv}$ includes backtracking paths such as the sequence u, w, u, v. This means the second term also includes a term equal to $A_{uv}(|\mathcal{N}(u)| + |\mathcal{N}(v)|)$, i.e. there is a contribution from the Edge Existence similarity $s_{\mathrm{EE}}(u,v)$ Eq.(2.14) in this method.

The *Katz Index* [131, 132, 133] (Katz) counts the number of paths between each pair of vertices, where each path of length ℓ contributes a factor of β^ℓ to the score. The score is simply the appropriate entry of the matrix $[\mathsf{I} - \beta \mathsf{A}]^{-1}$,

$$s_{\mathrm{Katz}}(u,v) = ([\mathsf{I} - \beta \mathsf{A}]^{-1})_{uv} \,, \tag{2.21}$$

where β is positive but must be less than the largest eigenvalue of the adjacency A. Note for low β and for a simple graph we have that

$$\begin{aligned} s_{\mathrm{Katz}}(u,v) &= \beta A_{uv} + \beta^2 \sum_w A_{uw} A_{wv} + \beta^3 \sum_{w,x} A_{uw} A_{wx} A_{xv} + O(\beta^4) \\ &= \beta s_{\mathrm{EE}}(u,v) + \beta^2 s_{\mathrm{LPI}}(u,v) + O(\beta^4). \end{aligned} \tag{2.22}$$

The *Local Leicht-Holme-Newman Index* [57] (LLHN) is based on the vertex similarity index of [139], but while this gives a specific motivation for the form, it is in the end just a specific rescaling of the Katz index (2.21), namely

$$\begin{aligned} s_{\mathrm{LLHN}}(u,v) &= \frac{s_{\mathrm{Katz}}(u,v)}{s_{\mathrm{PA}}(u,v)} \\ &= \frac{s_{\mathrm{Katz}}(u,v)}{|\mathcal{N}(u)|\,|\mathcal{N}(v)|} \\ &= \mathsf{D}^{-1}\left(\mathsf{I}-\beta\mathsf{A}\right)^{-1}\mathsf{D}^{-1}. \end{aligned} \tag{2.23}$$

where D is a diagonal matrix whose entries are equal to the degrees of the nodes, $D_{uv} = \delta_{u,v}|\mathcal{N}(u)|$. The motivation for using this normalisation is that $|\mathcal{N}(u)| \cdot |\mathcal{N}(v)|$ is proportional to the number of neighbours expected in the configuration model. So the Local Leicht-Holme-Newman Index is the Katz score relative to the Katz score expected for the same pair of nodes in the configuration model.

The *Matrix Forest Index* [140] (MFI) is defined as:

$$s_{\mathrm{MFI}}(u,v) = \left[(\mathsf{I}+\mathsf{L})^{-1}\right]_{uv}, \tag{2.24}$$

where $\mathsf{L} = \mathsf{D}-\mathsf{A}$ is the Laplacian. One way to understand the Matrix Forest Index is to consider the diffusion process described by a Laplacian. If we were to demand that at time $t+1$ we only had particles at one site u, then $s_{\mathrm{MFI}}(u,v)$ tells us how many of those particles were at vertex v at the previous time step t.

2.5.3 From node similarity to link prediction

All the link prediction methods used in this Chapter, see Table 2.2, assign a similarity score between pairs of nodes. To turn this into a link prediction, the basic conjecture is that the higher the similarity between a pair of nodes, the more likely we are to find a link between these two nodes.

All methods, therefore, require a precise method to turn the scores into predictions, essentially

to define what is meant by a "high" or a "low" score by introducing a *classification threshold*. Often this is done very simply by ranking the scores and using a fixed number of the most highly ranked node pairs to predict a link. This method is typically used only for *link addition* in which one is only trying to predict when an unlinked node pair gains a link, that is, $A_{uv}(s) = 0$ to $A_{uv}(s+1) = 1$ process.

We are interested in the most general predictions, looking at all four possible changes for node pairs from one snapshot in time to the next, that is, all the four possible $A_{uv}(s) = 0, 1$ to $A_{uv}(s+1) = 0, 1$ processes. This is *link evolution* rather than link addition. In order to make these more general predictions for any method, we use k-means clustering methods to separate the prediction scores produced by each method into two classes: A high score group and a low score group of node pairs. Any node pair with a score in the high scoring group will be predicted to have a link in the next snapshot; node pairs in the low scoring group will be predicted to have no link.

To show how this works, we give some examples of how the score from our triplet transition method produces a natural split into low and high scores which is easily discovered by an automated clustering method such as k-means, as seen in the first-row of Figure 2.8. In our final results below for our link prediction algorithm, we use $\mathcal{M}_8$ and the clear separation of node similarity scores into a low and high group is shown in the second row of Figure 2.8. For comparison we also show the results for our method using the less sensitive $\mathcal{M}_4$. In practice, both seem to separate low and high scoring node pairs well but we get a slightly clearer separation in some cases when using $\mathcal{M}_8$.

We can look at how the other nine link prediction methods perform in terms of identifying two clear groups of low- and high-scoring node pairs. What is surprising is that most of the algorithms fail to do this well, or in some cases at all. An example of a poor separation, the Jaccard Coefficient method, is shown in Figure 2.9. The one exception is the Katz method, also shown in Figure 2.9, which works well in three of the four data sets. Given this is one of the methods probing large distances in the network, this would seem to support the idea that higher-order interactions are important in link prediction. Since our triplet transition method

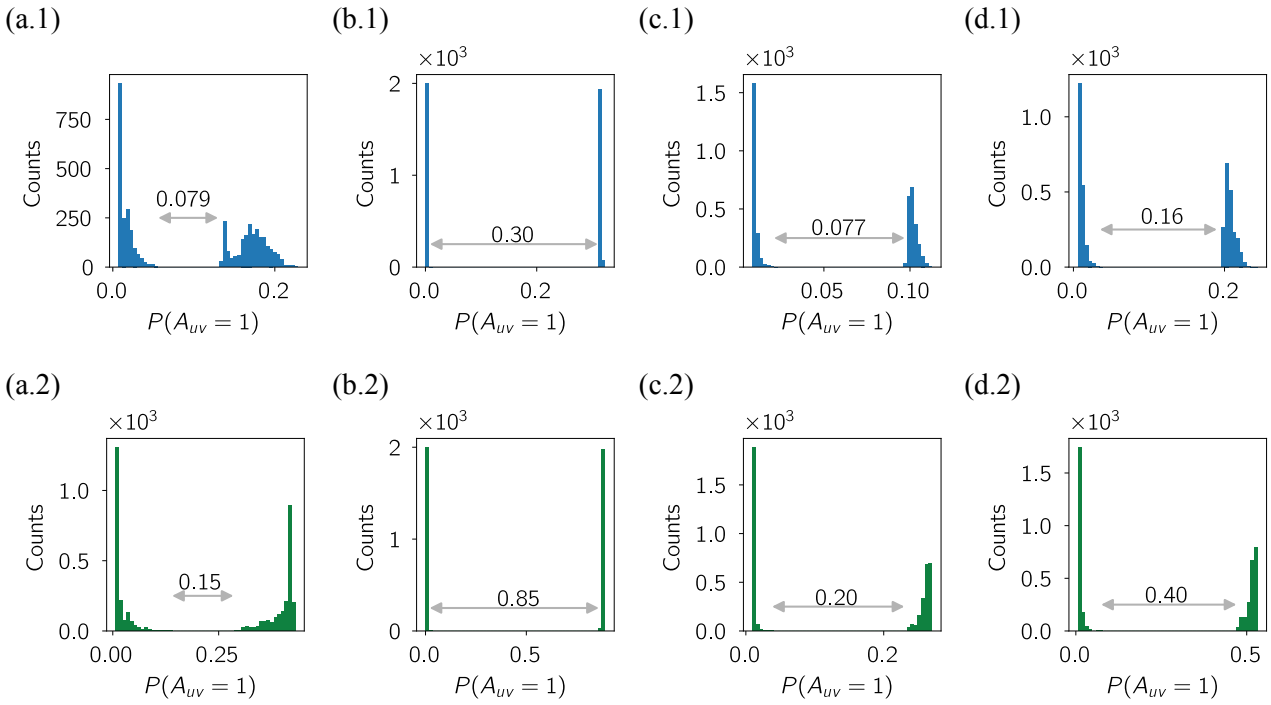


Figure 2.8: The histogram of node similarity scores in our triplet transition (TT) method for (a) Turkish Shareholder network, (b) Wikipedia Mathematician network, (c) College Message network, and (d) Email network. The precise values are not important as the important feature is the clear separation of the node similarity scores into two groups. One cluster c_0 appears to have a 'low' probability that a link will exist in the next snapshot and the other cluster has 'high' probability. We therefore predict that if the node pair has a score in the lower score cluster c_0, then a link will not exist in the next snapshot. Conversely, if a node pair exists in the higher score cluster c_1, then we predict a link will exist between this node pair in the next snapshot. The first row, where the histogram is plotted in blue is the clustering results for $\mathcal{M}_4$ while the second row, where the histogram is plotted in green is the clustering results for $\mathcal{M}_8$. It shows that $\mathcal{M}_8$ has higher separation between two clusters than $\mathcal{M}_4$. The results for each network are based on samples which contain at least 2000 node-pairs with an initial link and at least 2000 more node-pairs which started without a link.

splits the node similarity scores so well, any unsupervised clustering method should be able to split the node pairs into a low scoring cluster and a high scoring cluster. As the problem is in one dimension a version of k-means clustering is sufficient and this will always assign our node pairs to either a low or to a high score even if a method does not have a clear threshold visually. The objective function of our clustering is J where

$$J(s_{\text{th}}) = \sum_{(u,v)} \Big[\theta\big(s_{\text{th}} - s(u,v)\big) \cdot \big|s(u,v) - \mu_-\big| + \theta\big(s(u,v) - s_{\text{th}}\big) \cdot \big|s(u,v) - \mu_+\big| \Big], \tag{2.25}$$

where θ is the Heaviside step function. The sum over (u,v) is over all distinct nodes pairs (so

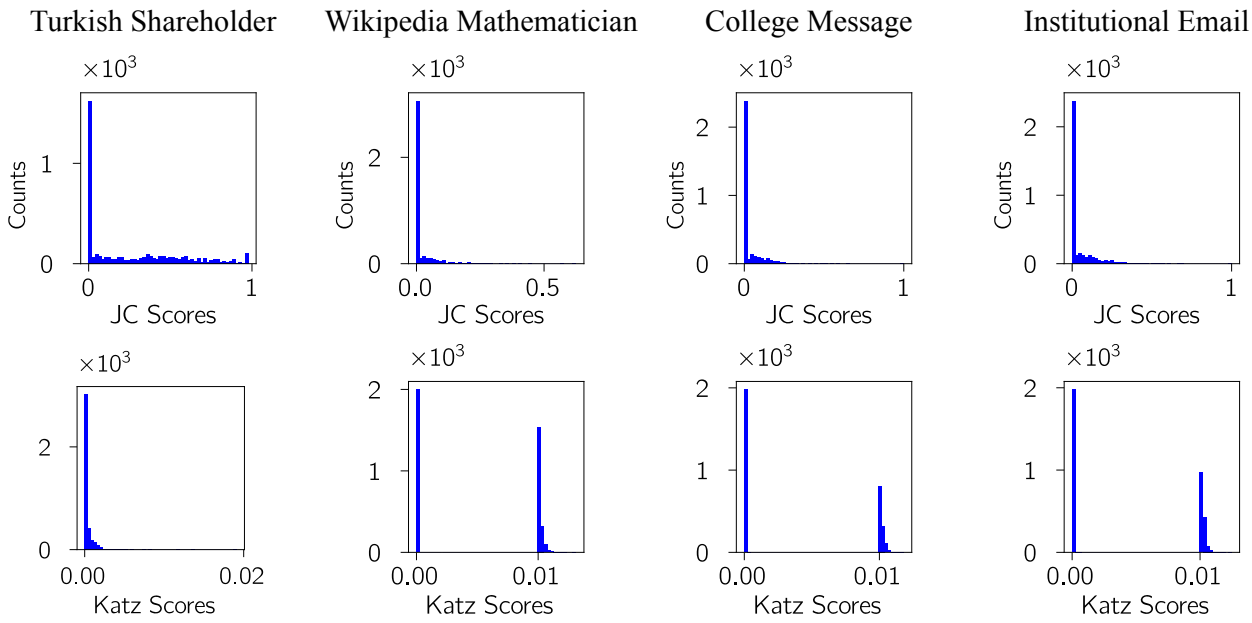


Figure 2.9: The histogram of node similarity scores using Jaccard Coefficient (JC) & Katz methods on four real data sets. Each column is for one data set which are, from left to right: Turkish Shareholder network, Wikipedia Mathematician network, College Message network, and Email network. The first row is the similarity score used in Jaccard Coefficient method and the second is for the Katz method. The precise values are not important here as the key feature is the success or failure to identify two clear groups of low- and high-scoring node pairs. We under-sampled 1000 linked node pairs and 1000 non-connected node pairs [141]. Only the Katz method (bottom row) shows the necessary clustering of scores needed for link prediction, except for Turkish shareholder network.

$(u, v) \in \mathcal{V}^2 | u \neq v)$. The μ_+ (μ_-) is the average similarity score of node pairs in the high score (low score) cluster. The problem is reduced to finding the threshold value s_{th} that minimises J.

2.5.4 Evaluation metrics

To evaluate the performance of our Triplet Transition and the other link prediction methods, we use a variety of standard metrics.The simplest metric we use is the fraction of edges that are removed from snapshot s to the next snapshot $s+1$. We denote this by $f_{\mathcal{E}}(s)$,

$$f_{\mathcal{E}}(s) = 1 - \frac{|\mathcal{E}(s+1) \cap \mathcal{E}(s)|}{|\mathcal{E}(s)|}, \tag{2.26}$$

where $\mathcal{E}(s)$ is the link set of the network in snapshot s.

We also use more traditional metrics such as precision and area under the curve for which we need a binary classification. Thinking of snapshot $\mathcal{G}(s)$ as the current state while we are trying to predict the state in the next snapshot $\mathcal{G}(s+1)$, we consider four transitions $(A_{uv}(s) \to A_{uv}(s+1))$ of node pairs: $0 \to 0$, $0 \to 1$, $1 \to 0$, and $1 \to 1$. We map these states onto a binary set of states, namely $A_{uv}(s+1)$, so what we call a positive result (negative result) is where a link is present (is not present) in snapshot $s+1$ regardless of the state that pair started in. We then consider whether the prediction for the state of the node pair in snapshot $s+1$ was true or false. So a false positive is where we predict a link for a node pair when that pair did end up with no link, while a true negative is where we correctly predict a pair of nodes will not be connected in the next snapshot. The table of predicted classes and actual classes is shown in Table 2.3. With this traditional binary classification, we can then evaluate the performance

		Actual Classes	
		$A_{uv}(s+1) = 1$	$A_{uv}(s+1) = 0$
Predicted Classes	$A_{uv}(s+1) = 1$	True positive ($01+, 11+$)	False positive ($01-, 11-$)
	$A_{uv}(s+1) = 0$	False negative ($00-, 10-$)	True negative ($00+, 10+$)

Table 2.3: The Confusion matrix for the prediction of edges between nodes u and v in the next snapshot, $\mathcal{G}(s+1)$. The adjacency matrix at snapshot s in the temporal network is $\mathsf{A}(s)$. Each of the four outcomes comes from two situations as this confusion matrix does not consider the state of the edge in snapshot $\mathcal{G}(s)$. The two states are shown for each of outcome in brackets with the notation that $\alpha\beta\pm$ shows the value of $A_{uv}(s)$ as α, the predicted $A_{uv}(s+1)$ as β, and a + symbol (a − symbol) indicates that the prediction made was correct (incorrect).

of our Triplet Transition method using standard metrics — area under curve and precision. To express these, we define $N_{\alpha\beta\pm}$ to be the number of node pairs satisfying the following criteria. The node pair starts in snapshot $\mathcal{G}(s)$ in state α equal to 1 if the node pair is connected by an edge, 0 otherwise. This edge pair is then in state β in snapshot $\mathcal{G}(s+1)$ with β is 1 if the node pair is connected by an edge in snapshot $\mathcal{G}(s+1)$, and is 0 otherwise. Finally the sign indicates if the prediction made for that node pair in $\mathcal{G}(s+1)$ was correct (true, +) or incorrect (false, −).

We can evaluate the methods independent of the classification threshold chosen through k-means by using the *area under the curve* (AUC) where the curve is the receiver operating

characteristic curve. For any binary classifier of the results of an algorithm, such as defined here in Table 2.3, the curve plotted is the fraction of positive results (TPR, true positive rate) which are correct against the fraction of negative results which are incorrect (FPR, false positive rate) as the threshold $s_{\rm th}$ is varied:

$$\mathsf{TPR} = \frac{N_{11+} + N_{01+}}{N_{11+} + N_{01+} + N_{00-} + N_{10-}}, \quad \mathsf{FPR} = \frac{N_{01-} + N_{11-}}{N_{01-} + N_{11-} + N_{00+} + N_{10+}}. \tag{2.27}$$

Once the threshold $s_{\rm th}$ has been fixed, in our case using k-means in Eq.(2.25), the *precision* score $S_{\rm prec}$ is defined as the number of times that we predict a link exists between node pairs in the later snapshot correctly (a true positive, $N_{11+} + N_{01+}$) divided by the number of times we predict a link, correctly (true positive) or incorrectly (false positive, $N_{11-} + N_{01-}$):

$$S_{\rm prec} = \frac{N_{11+} + N_{01+}}{N_{11+} + N_{01+} + N_{11-} + N_{01-}}. \tag{2.28}$$

A high precision score means we can trust that links predicted by the algorithm will exist.

We can set a baseline value for precision using a simple model which predicts a link in snapshot $\mathcal{G}(s+1)$ exists for any given node pair with a probability given by the fraction of node pairs which have an link in snapshot $\mathcal{G}(s)$, that is the density $\rho(s)$. In this case, the precision (see Eq.(2.28)) is simply equal to the density in snapshot $\mathcal{G}(s)$, $S_{\rm prec,base} = \rho(s)$.

A very low precision score for baseline method indicates that the links existing between node pairs are not random. Some of the local measurement algorithms give lower scores than the baseline, suggesting that those types of local information are less likely to drive the evolution of the networks.

2.5.5 Edge sampling

We use the whole network to evaluate the performance, which is different from the sampling method used in Figure 2.8 to take account of the full nature of the data. In the following analysis, we predict the node pairs for all the possible node pairs of a whole network. We

expect predictions can capture evolving characteristics of networks and the Mathematician network change tiny fractions which are not sufficient enough to evaluate the predictions. For example, the algorithm can predict edges stay in the next snapshot of the network will attain at least 99% accuracy. Therefore, we replace the Mathematician network with Hypertext network in the following prediction analysis. Time data is sampled over different time intervals Δt. For Turkish Shareholder network, the smallest time separation is 2 years, and we consider four or more years too long for the evolution, since the shareholders on average quit market after 2 years; for the Hypertext data, which records the short communications, we choose Δt as 40 and 60 minutes; for the College Message and Email network, we consider that 8 hour to 1 month communication frequency is reasonable.

2.6 Results

Figure 2.10 shows a comparison of AUC for the ten algorithms when we apply them to node pairs sampled uniformly at random from all possible node pairs.

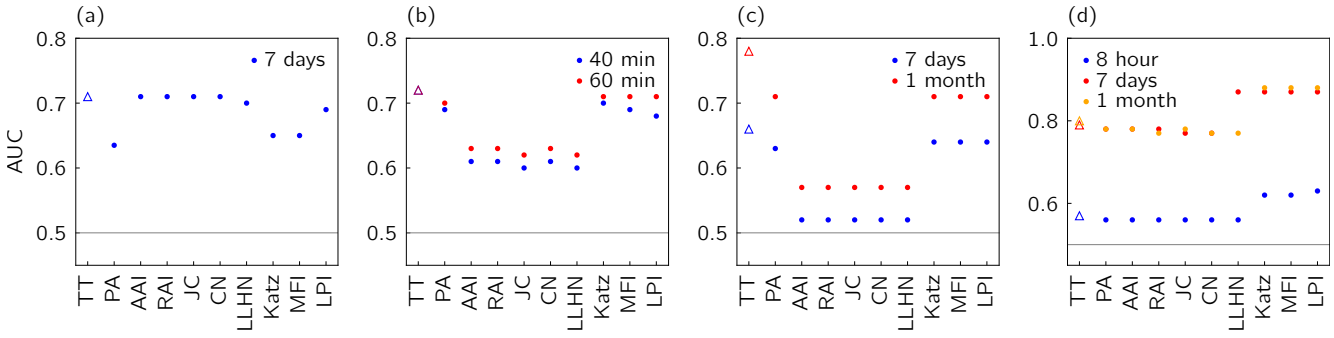


Figure 2.10: The area under the curve (AUC) results for four temporal networks constructed from real data sets: (a) Turkish Shareholder network, (b) Hypertext network, (c) College Message network, (d) Email network. The results compare ten different methods of Table 2.2 including our TT algorithm denoted by the 'triangle' symbol. For networks (b),(c) and (d), we show the results for different time scales (windows). See Table 2.2 for abbreviations used in indicate link prediction methods.

The Triplet Transition (TT) approach defined here is the left-most triangular point in each figure. In the Hypertext network (b) and the College Message network (c), the Triplet Transition method has the highest AUC though the PA, Katz, MFI and LPI algorithms perform almost

as well (see Table 2.2 for abbreviations). For the Turkish Shareholder network (a) the AUC of our Triplet Transition method is again the highest though with a similar AUC value for various algorithms. Note that PA, Katz, and MFI are now weaker. Only for the Email network (d) is our Triplet Transition outperformed, in this case by the Katz, MFI and LPI algorithms.

Type	Algorithm (Time scale) $[\overline{f_\varepsilon(s)}]$	Shareholder (2 years) [0.63]		Hypertext (1 h) [0.52]		CollegeMsg (1 mon) [0.93]		Email (1 mon) [0.50]		Avg. Rank
Global	*Triplet Transition*	**0.71**	1	**0.72 (7)**	1	**0.77 (6)**	1	0.80 (2)	4	1.8
	Katz ($\beta = 0.01$)	0.65	8	**0.71 (10)**	1	0.69 (8)	2	**0.88 (2)**	1	3.0
	Matrix Forest Index	0.65	8	**0.71 (9)**	1	0.69 (8)	2	**0.88 (2)**	1	3.0
	Local Path Index	0.69	7	0.69 (9)	5	0.69 (6)	2	**0.88 (2)**	1	3.8
Local	*Resource Allocating Index*	**0.71**	1	0.63 (8)	6	0.58 (3)	6	0.78 (2)	5	4.5
	Adar-Academic Index	**0.71**	1	0.63 (8)	6	0.58 (3)	6	0.78 (2)	5	4.5
	Common Neighbour	**0.71**	1	0.63 (8)	6	0.58 (3)	6	0.77 (2)	8	5.3
	Jaccard Coefficient	**0.71**	1	0.62 (8)	9	0.58 (2)	6	0.77 (2)	8	6.0
	Preferential Attachment	0.63	10	0.70 (7)	4	0.69 (6)	2	0.78 (1)	5	5.3
	Local Leicht-Holme-Newman	0.70	6	0.62 (7)	9	0.57 (2)	10	0.77 (2)	8	8.3
	Baseline	0.5	11	0.5	11	0.5	11	0.5	11	11.0

Table 2.4: Summary of AUC-ROC performance of different algorithms. Δt is taken by selecting the highest AUC-ROC. The AUC scores for each network are in the left-hand column, the rank of the score in the right-hand column. The errors quoted on the AUC scores are standard deviations from all the computed network snapshots of the selected time scale except for Turkish Shareholder network where only one prediction is made. We highlight the highest result and any whose result is within the error quoted on the largest result. The baseline method for AUC is the random clustering of node pairs into two groups, and the AUC is 0.5 which means it is like 'tossing a coin'. The Type column indicates if the link prediction method probes only local scales (path length two) or global scales (infinte path lengths)

Figure 2.11 shows that there are some clear patterns in our results for precision. First, each link prediction method has a similar performance relative to the other methods on three of the networks: the Hypertext network (b), the College Message network (c), and the Email network (d). On these three data sets, three algorithms are consistently better than the others though with similar performances relative to one another: our Triplet Transition (TT) method, the Katz method [131, 132, 133], and the Matrix Forest Index [140] (MFI) method. All of these are probing non-local information in the networks, suggesting this is necessary to understand the time evolution of these networks.

Overall we see some similarity in these AUC results, summarised in Table 2.4 as we saw for precision in Table 2.5. The same three algorithms, the Triplet Transition (TT) methods, the Katz method, and the MFI method, have a similar high performance on the same three

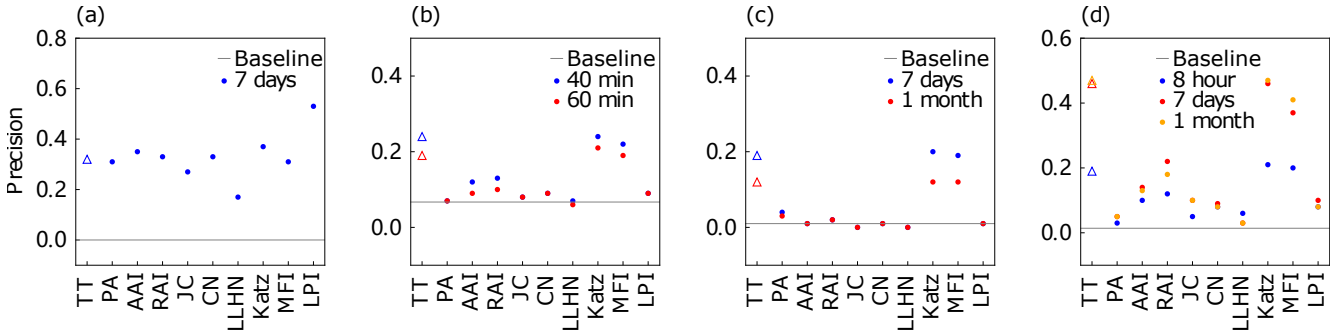


Figure 2.11: Precision scores for the link prediction results for ten algorithms applied to four temporal networks constructed from real data sets: (a) Turkish Shareholder network, (b) Hypertext network, (c) College Message network, (d) Email network. Results for the Triad Transition algorithm based on transition matrix denoted by the triangle symbol △. For the networks in (b), (c) and (d), we also show the results for different time scales (window). See Table 2.2 for abbreviations used in indicate link prediction methods.

Type	Algorithm (Time scale) $[f_\mathcal{E}(s)]$	Shareholder (2 years) [0.63]		Hypertext (1h) [0.52]		CollegeMsg (1 mon) [0.93]		Email (1 mon) [0.50]		Avg. Rank
Global	*Triplet Transition*	0.27	5	0.19 (9)	2	0.11 (4)	2	**0.47 (6)**	1	2.5
	Katz ($\beta = 0.01$)	0.25	8	**0.22 (8)**	1	0.11 (4)	2	**0.47 (6)**	1	3.0
	Matrix Forest Index	0.19	9	0.19 (9)	2	**0.12 (3)**	1	0.4 (5)	3	3.8
	Local Path Index	**0.36**	1	0.09 (4)	5	0.009 (2)	7	0.08 (3)	7	5.0
Local	*Resource Allocating Index*	0.28	4	0.10 (6)	4	0.019 (4)	5.0	0.18 (7)	4	4.3
	Adar-Academic Index	0.33	2	0.09 (5)	5	0.008 (2)	8	0.13 (5)	5	5.0
	Common Neighbour	0.31	3	0.09 (5)	5	0.008 (2)	8	0.08 (1)	7	5.8
	Jaccard Coefficient	0.25	7	0.08 (6)	8	0.0010 (9)	10	0.10 (4)	6.0	7.8
	Preferential Attachment	0.27	6	0.07 (6)	9	0.03 (1)	4	0.059 (6)	9	7.0
	Local Leicht-Holme-Newman	0.09	10	0.06 (5)	11	0.0001 (2)	11	0.03 (1)	10	10.5
	Baseline	0.0001	11	0.067	10	0.010	6	0.014	11	9.5

Table 2.5: Summary of average precision scores of different algorithms. Δt is taken by selecting the highest precision based on k-means clustering. The precision score for each network are in the left-hand column, the rank of the score in the right-hand column. The errors quoted are standard deviations from all the computed network snapshots of the selected time scale, except for Turkish Shareholder network which have one prediction for three snapshots.

networks, the Hypertext network, the College Message network and the Email network. For these three networks, we might also pick out the Preferential Attachment (PA) algorithm. However, now the AUC values for the Turkish Shareholder network show that our Triplet Transition method continues to perform well, unlike for the precision measurements. The Katz, Matrix Forest Index and Preferential Attachment methods are, though, in a weaker group of algorithms as measured by their AUC performance on Turkish Shareholder network.

For three of these networks, we also measure AUC over snapshots of these networks defined over different time scales. The comparison among methods rarely changes. However, the

performance of most algorithms is altered by the size of the time window chosen in some cases, reflecting inherent timescales in the different systems. The most noticeable is that for the Email network, there is a large difference between windows of eight hours and one week but not so much change between a week and one month. The gaps may suggest that if a person is going to produce an email, perhaps following up an email request of bringing a third person into the conversation, that new email is done often on the scale of a few days not always on the scale of a few hours.

The Triplet Transition proposed here is used to predict a link between two nodes by considering these alongside a third node which can be in any position in the network. In this way, this third node not only captures the higher-order interactions but also enables the method to encode both local and non-local information about the link of interest. We find that our Triplet Transition method and the two other global methods used here, Katz Index method [131, 132, 133] and the Matrix Forest Index [140] are generally the best for most of networks we studied, in particular for the Hypertext network, the College Message network and the Email network. As the most successful methods here perform better than other approaches based on local measures, this shows that in most systems the pattern of connections depends on the broader structure of interactions. This dependence of the behaviour of systems on structure beyond nearest neighbours is the crucial motivation for using the language of networks rather than just looking at the statistics of pairs [210]. For instance, the Katz method counts the number of paths between each pair of nodes, probing all paths though giving less weight to longer paths, so nearest neighbours contribute the most.

The results for theTurkish Shareholder network were a little different. In this case, predictions based on local measurements (paths of length two), the semi-local Local Path Index method (LPI), as well as our non-local triplet transition method outperformed the other global methods in terms of AUC, see Table 2.4. The global methods also perform poorly on precision, see Table 2.5. An algorithm with low precision and high AUC, such as Jaccard Coefficient (JC), is predicting the disconnected pairs well.

2.7 Discussion and conclusion

In this chapter, we demonstrated that higher-order interactions are needed to understand the evolution of networks. For example, in the Email network, derived from the emails within an EU Institution, if three nodes are connected in the previous temporal snapshot, they are less likely to be disconnected in the following snapshot.

To take into consideration of the interactions of neighbours, we designed a link prediction algorithm, Triplet Transition (TT) method, based on the transition matrix T of Eq. (2.2). From a range of different temporal networks, we found Triplet Transition method was as good as two methods based on non-local (global) information in the network, namely, the Katz Index method [131, 132, 133] and the Matrix Forest Index (MFI) method [140]. While not always the best on every network or every measure, these three global methods were usually better than the other methods we studied, all of which used information on paths of length two in the network. Intriguingly, the one other method that used paths of length three, Local Path Index [132, 138] (LPI), often performed well too though rarely as well as the top three global methods.

Since the most successful methods in our tests were those that access non-local information, it seems that such information is essential in the evolution of most networks and therefore, it is important to include this in network measures. However, including information from the whole network is numerically intensive and, for any reasonably sized network, the evolution of a link is unlikely to depend directly on what is happening a long way from that link. One reason why our Transition Triplet method works well is that the triplet transition mechanism appears in the vast majority of networks. Most of the information in the transition matrix network is based on neighbours of one or other of the link of interest. For large networks, we use sampling to add the necessary global information into the transition matrices. The Katz index method does include information from all scales but suppresses contributions from more distant parts of the network. The success of the transition triplet approach suggests that there is no need to access all of the global information of a network in order to know what is going on locally. Notably, the overall better performance of the Triplet Transition method reveals that

the information of different higher-order interaction patterns can help understand and predict different dynamics of networks.

There is also another reason why our Triplet Transition method may work better than the local methods we look at, and that is because our method is also probing a longer time scale as well as a longer spatial scale. That is we use *two* snapshots, $\mathcal{G}(s-1)$ and $\mathcal{G}(s)$ in order to create the transition matrix T which in turn we use for the predictions in snapshot $\mathcal{G}(s+1)$. All the other methods used here are based on information from one snapshot $\mathcal{G}(s)$ only. So again, the success of the Triplet Transition method points to correlations over short but non-trivial time scales as being important in understanding network evolution. In our case, the dependence of results on the time intervals can be seen in the effect of Δt used to define the transition matrices are important. For different data sets, we found different Δt gave optimal results which of course reflects inherently different timescales in the processes encoded by our different data sets. So another conclusion is that higher order effects in terms of both time and space are needed to understand the evolution of temporal networks and to make effective predictions for links. The inclusion of higher-order order effects in terms of time remains undeveloped relative to the work on higher-order spatial network features.

In our approach we have focused on changes in the edges and ignored changes in the set of nodes. Our method can include nodes which are not connected in some or even all our snapshots provided these nodes are known. However, in many cases the data sets may only record interactions, our edges, and our set of nodes is only inferred from that. As a result we often have no information about such (totally isolated) nodes. Suppose we are given a list of emails, our edges, sent at a given time between email accounts, our nodes. We cannot distinguish accounts which are dormant from those that are deleted or added to the system without additional information. Should we have such data on node birth and death, then in some cases it might be an important process to include, but it is not one we address in our approach.

Currently constructing optimal higher order models is a timely-topic [28]. Our method can be generalised to include even higher-order interactions, such as quadruplet and so on. The method

can be used to indicate higher-order evolutionary mechanisms in a network and suggest what is the most likely order of interaction. Our work shows that studying the evolution of small graphs over short time periods can reveal important information and predictions regarding network evolution.

2.8 Summary

In this chapter, we can summarize the content as follows:

- designed a transition matrix that measures Markovian evolution of triplets (three-nodes motifs) based on two snapshots of networks
- designed a benchmark pairwise model to detect the higher-order interactions and illustrated existence of higher-order interactions in both artificial networks with known mechanism and real world networks
- calculated a score based on Markovian triplet transition that can be used to predict link evolution in temporal networks
- compared our methods with nine methods in four real-world data sets. Our method is always one of the best algorithms in terms of AUC-ROC and precision, which shows a strong evidence that the evolution of the temporal network is indeed based on three nodes configurations.
- discussed the advantage and disadvantage of our method, with potential future work

Chapter 3

Modelling the spatial dynamics of dockless bicycles using gravity model

The following chapter is based on:

R.,Li, S., Gao, A. Luo, Q., Yao, B., Chen, F., Shang, R., Jiang. & H.E., Stanley

Gravity model in dockless bike-sharing systems within cities.

Physical Review E ***103****(1), 012312 (2021) .*

B. Chen contributed to write all parts of the manuscript, numerical and statistical analysis in 3.2 and 3.3.

3.1 Introduction

Along with rapid urbanisation, we encounter many challenges including environmental degradation, traffic congestion, air pollution and greenhouse gas emission. Developing a greener and more sustainable transportation system is critical for future urban development to mitigate the aforementioned urban issues induced by rapid urbanisation, among which bike-sharing systems would play an important role. In recent years, with a booming of the sharing economy [142, 143], and development of "Internet of Things"(IoT) and mobile payment technology, dockless bike-sharing systems (including brands, such as Mobike and OFO) have been quite

popular, especially in China [144]. Compared to docked station-based sharing bikes (i.e., a bike has to be returned to a dock at certain fixed established sites), dockless ones are free from such restrictions and give better accessibility and more flexibility to users. Basically, you can park and "return" a dockless bike anywhere suitable for a bike near your destination. Yet there are many challenges on maintaining the system efficient and controlling the cost, which strongly depends on a better understanding of patterns of human mobility on biking behaviours.

The dockless bike-sharing platform would record the departure and arrival location as well as the start and end timestamp of each order for billing purposes. In addition, different from the station-based bike-sharing system where origins and destinations of trips are fixed, and the number of available docks for parking bikes is also fixed and usually quite limited, dockless ones are free from such restrictions and thus provide us a unique opportunity to investigate a more spatially and quantitatively accurate and realistic biking patterns within cities. The origin and destination locations can be quite accurate, and the flow between locations suffers much less from under-supply and can be much closer to the real demand, since the number of deployed dockless bikes are usually quite abundant or even excessive, especially in large Chinese cities [145].

As for predicting various types of human movements, the gravity model has been a widely applied framework with a simple form. It dates back to the early 1780s [146], and was first suggested for use in human interaction systems by Carey in 1860s [147]. Its contemporary form, which was proportional to populations of the origin and destination locations and the inverse of the distance between them, is introduced by Zipf in 1940s [149]. Yet to our best knowledge, there's no work testing whether the gravity model on biking traffic holds or not within cities[148], let alone newly emerged dockless sharing bikes. Most of the previous works are about human movements on highway [149, 150], railway [149], airline [149, 151], or telecommunication flows [152, 153, 154], cargo shipping [155], migration [156, 157, 158], bilateral trade [159], even scientific citation and collaboration [160], etc. Generally speaking, validating the gravity model is important for predicting human mobility and making better urban planning and design [161, 162, 163, 164], which are crucial for traffic engineering [165, 166, 167, 168], predicting epidemic spreading [151, 169, 170, 171, 172], emergency management [173, 174]. In

addition, in conventional studies on the gravity model, a location usually refers to a large region, such as a nation, a county with a fixed boundary or a city i.e., the spatial scale in most works are at international [151, 155, 159, 160], inter-county level [152, 153, 151, 169, 176] or inter-city [149, 150, 151, 175]; in comparison, fewer works are at intra-city level, such as at inter-ward level in UK [177], inter-*dong* level in the subway system in Seoul [178], or even at a finer spatial resolution [179, 180, 181]. Moreover, the effects of spatial scale of locations on gravity model is not systematically studied.

In this chapter, we first investigate the spatial distribution of riding activities, and find their cumulative distribution from the city centre to suburbs are in line with the spatial scaling theory [182] in both Beijing and Shanghai. The volume of biking traffic between locations (at a fine spatial resolution, e.g., a location might refer to a region of 500 m×500 m) manifest a power-law with a similar magnitude of the power-law exponent in Beijing and Shanghai. We observe that general gravity models with two population related parameters (no matter with distance in spatial domain or dimensionless distance in topological space) on biking traffic holds across different spatial resolutions, ranging from 500 m × 500 m to 5 km × 5 km, and discover that with the increase of spatial scale, the accuracy of the gravity model improves as indicated by a higher coefficient of determinant, R^2 (see more details in Eq.(3.4)). Our findings indicate that the gravity model can be applied to model biking behaviour at a quite fine spatial resolution. Besides, we also reveal that the distance related parameter increases in a similar way as population related parameters. This indicates that there might be a deeper relationship between the population and distance, which would require further studies. In comparison, conventional gravity models, where the population related parameters are all assumed to be 1, don't hold well especially at finer spatial scales in our project, which further confirms that the effects of population and spatial scale are nontrivial. Furthermore, we investigate the features and patterns of some special locations (sources and sinks) that can not be fully explained by the gravity model but are crucial for better operation of biking systems and enhancing the user experience.

3.2 Data set description

The datasets on the dockless bike-sharing system are obtained from Mobike, Inc., which was the first and largest dockless bike-sharing platform in China. The company was registered in Beijing in 2015, and its service was first launched on April 22^{nd}, 2016 in Shanghai. We have two separate datasets for Shanghai and Beijing, respectively. The dataset for Shanghai has 1.02 million bill records of more than 17 thousand "users" (also referred as "riders" thereafter) and 0.3 million dockless shared bikes over a whole month (from Aug. 1^{st} to Sep. 1^{st}, 2016 – roughly 3 months after its service first got online). In Beijing, there are more than 3.2 million records, detailing the riding behaviours of 0.35 million users and 0.485 million dockless shared bikes over two weeks (from May 10^{th} to 24^{th}, 2017).

Each record has an order ID, a user ID, a bike ID, departure and arrival locations and their corresponding timestamps. The start and arrival locations of Beijing's data set is geohashed which means that each unique location has a unique string (e.g., "wx4snhx"), and we convert geohashed strings to longitudes and latitudes.

We filter the raw data in the following ways. We firstly discard the records not located within the boundary of Beijing and Shanghai, and also discard the ones with a riding duration longer than one day (which might be some bikes left unlocked after its initial order) or less than one minute (which might be due to unsatisfied tryouts) or a Euclidean distance between origin and destination longer than 100 km (which exceeds the diameter of the urban area in both Shanghai and Beijing) or the average riding speed faster than 0.6 km/min (which is around twice of the normal cycling speed).

To avoid possible selection bias, we didn't pose any further filtering criterion. As a result of these criteria, around 200 million (0.12 million) noisy records are discarded in Shanghai's (Beijing's) dataset.

3.3 Statistics on the biking patterns

3.3.1 Spatial distribution of riding activities

We first investigate the spatial distribution of departure activities, which are relatively concentrated at the central urban area (see Fig. 3.1(a) for Shanghai and Fig. 3.1(c) for Beijing). To make quantitative comparison between the two cities, we exploit the idea of spatial buffering – drawing monocentric circles with increasing radius from the city centre (indicated as a purple dot in Fig. 3.1) to compute the activity density within each ring with a distance r to the city centre (see Fig. 3.2(a)). The activity density is formulated as the following:

$$\rho(r) \propto r^{-\eta} \left(R^{1+\eta} - r^{1+\eta} \right), \tag{3.1}$$

where r is the distance to the city centre, the power-law exponent η describes a characteristic velocity of density decay and can vary across different cities, R is the total radius of the city with riding activities, and $\rho(r)$ is the average density of riding activities at a distance r to the city centre. Notice, here a ride is defined as if a path exists.

The power-law exponent η for both arrival and departure activities were obtained by minimising the fitting error between the data and Eq. (3.1). Equivalently, we can get the spatial scaling relation between the cumulative quantity $V(r)$ and the distance to the city centre r by a simple double integration:

$$\begin{aligned} V(r) &= \int_{\theta=0}^{2\pi} \int_{x=0}^{r} \rho(x,\theta) x dx d\theta \\ &\propto r^{2-\eta} \left(\frac{R^{1+\eta}}{2-\eta} - \frac{r^{1+\eta}}{3} \right) \sim r^{2-\eta}, \end{aligned} \tag{3.2}$$

Notice, the proportionality only hold at the condition $r \ll R$. Otherwise, the approximation is no longer valid and scaling relation deviates.

We observe that the riding activity density distributions are in line with the theory of spatial scaling [182], which suggests that riding activities follow a power-law distribution inside the

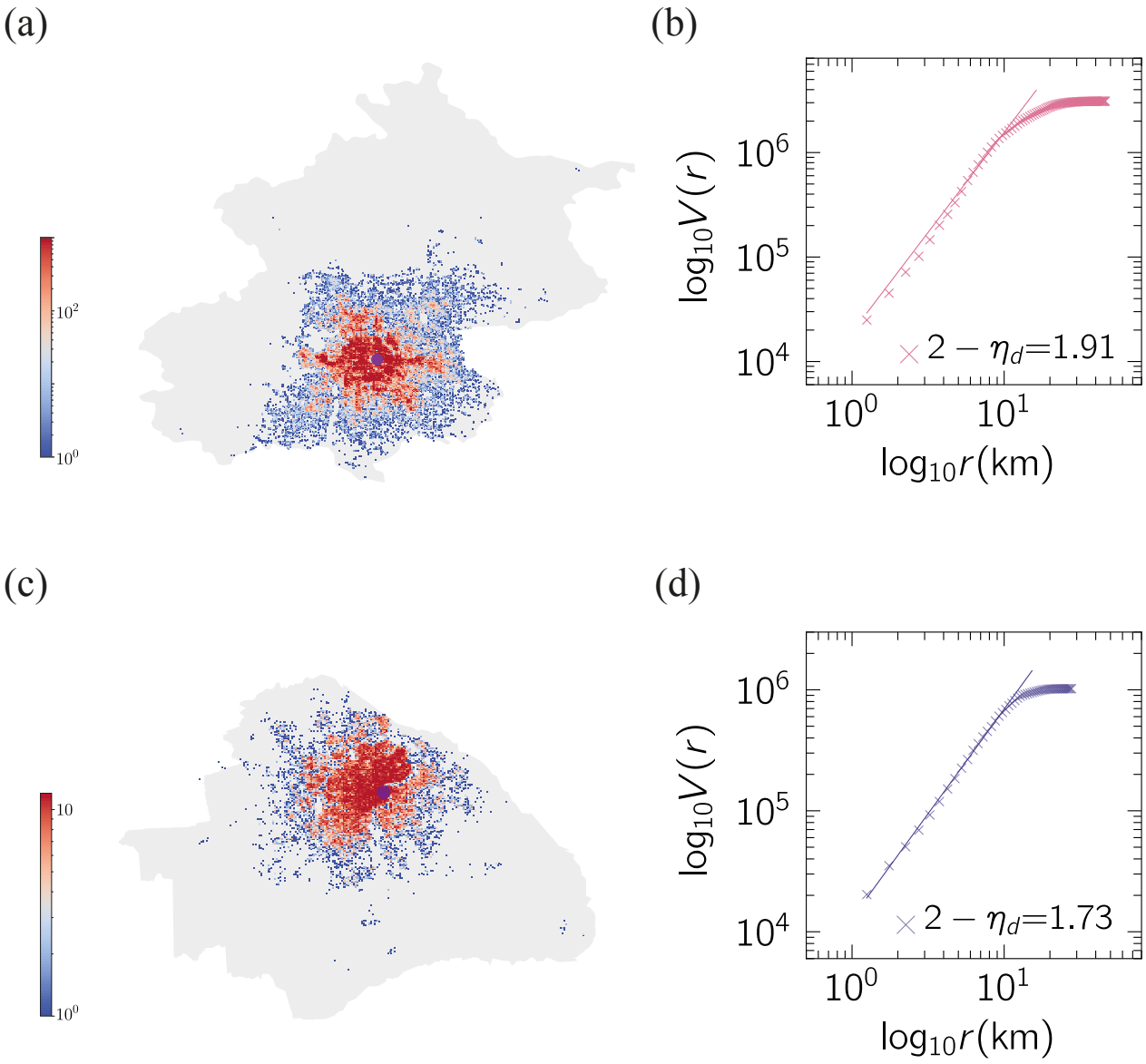


Figure 3.1: Spatial distribution of riding activities and spatial scaling relation between the cumulative volume of riding activities and the corresponding distance to the city centre in Beijing (a, b) and Shanghai (c, d). The purple dot indicates the city centre. The power-law exponent η for Beijing and Shanghai is roughly 0.09 and 0.27, respectively, which indicates that Shanghai has a faster decay on the density of riding activities along with the distance away from the city centre. It's worth noting that the spatial scale of cities and the total volume of riding activities are quite different.

central urban area and decays very fast in suburban regions [182] (the fast density decay part corresponds to a saturation phenomenon in Fig. 3.1(b) and Fig. 3.1(d)). As the η_d for Beijing and Shanghai is roughly 0.09 and 0.27, respectively, it indicates that Shanghai has a faster decay on the density of riding activities. Here, we only take the case of departure activities as

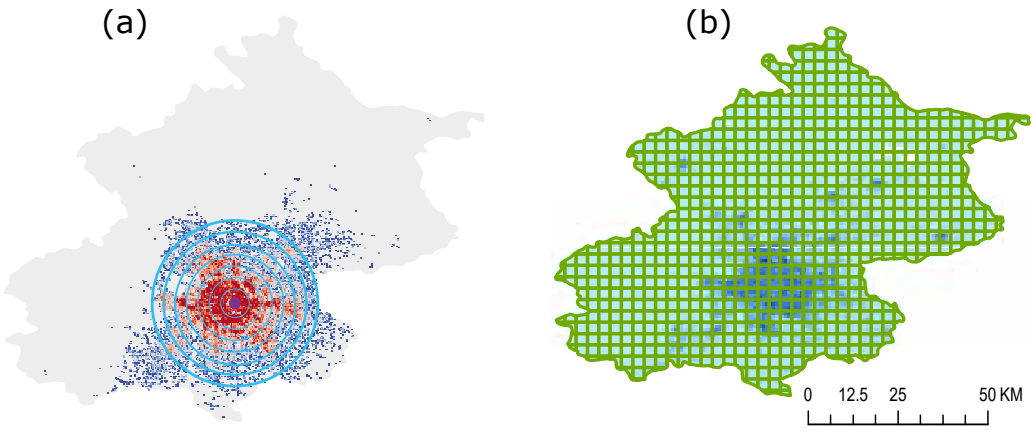


Figure 3.2: (a) Illustration of Spatial buffering and (b) the rasterization of the whole urban space. For (a), the purple dot indicates the city centre. All circles are monocentric with an increasing radius. (b) The rasterization of the whole urban space, each square corresponds to a location. In this example, the spatial resolution is 5 km, which means each location is of 5 km × 5 km. The underlying blue shading is the population data as of 2017, which is obtained from WorldPop dataset [183] with a 100 m spatial resolution. The population of each location is integrated from the underlying WorldPop layer. Here we only take Beijing as an example, the process for Shanghai's part is exactly the same, with only changing the WorldPop population mapping as of 2016.

an example, we find that within the same city, η for departure and arrival activity are almost identical.

3.3.2 Distributions of biking flux

After having a glance at the whole picture of the spatial distribution of riding activities within cities, we then investigate the flux of biking traffic between locations. For both Beijing and Shanghai, we rasterize the whole urban area into square-grid lattices with certain spatial resolutions, for example, each location (i.e., a lattice) would refer to a region of 500 m × 500 m (see Fig. 3.2(b)). Then we integrate all individual trips to obtain the biking flow between locations, and denote the volume of flow from location i to location j as T_{ij}. The distribution of T_{ij} values (i.e., "link weight" in complex network terminology) manifests a power-law with an exponent equals -2.32 ± 0.06 and -2.13 ± 0.06 at a 500 m spatial resolution of locations in Beijing and Shanghai, respectively (see Fig. 3.3(a) and Fig. 3.3(c)). This indicates an inherent heterogeneity between locations that there are few fluxes between most locations, yet a considerable fraction of location pairs are of relatively large volume of biking traffic. We also

relate the volume of biking traffic with the travel distance, and discover a fat tailed distribution above 1km in both Beijing and Shanghai (see Fig. 3.3(b, d)). This means that relatively long distance (i.e. $\geq$ 10 km) travel by bike are not that rare compare to many standard deviation from normal distributions since they follow power-law distributions.

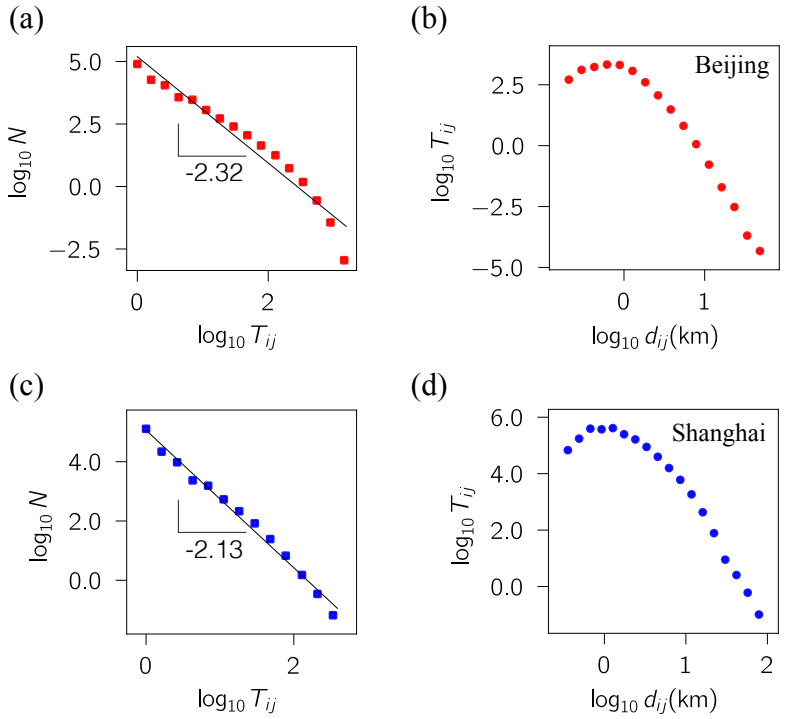


Figure 3.3: Distribution of biking flux between locations T_{ij} at a 500 m × 500 m resolution (a, c), and the dependence of biking flux distribution on the distance travelled (b, d) in Beijing (a, b) and Shanghai (c, d).The power-law exponents in (a) and (c) are -2.32 ± 0.06 and -2.13 ± 0.06, respectively. The distributions in (b) and (d) also have a fat tail.

3.3.3 Average travel distance of locations

In order to better understand the dynamics of the biking flows of a location, we then further look at the average outflow travel distance from an origin location i to all its destinations weighted by the biking traffic volume between them, and vice versa for average inflow distance.

$$\left\langle d_i^{\text{in}} \right\rangle_w = \sum_{j=1}^{N} w_{ji} d_{ji}, \tag{3.3a}$$

$$\left\langle d_i^{\text{out}} \right\rangle_w = \sum_{j=1}^{N} w_{ij} d_{ij}, \tag{3.3b}$$

where $w_{ij} = T_{ij}/T_i^{\text{out}}$ is the weight of the travel path (i, j) measured by the ratio between traffic volume from location i to j and the total outflow of location i (similarly, $w_{ji} = T_{ji}/T_i^{\text{in}}$), and N is the total number of locations in the city. We observed that the weighted average outflow travel distance $\langle d_i^{\text{out}} \rangle_w$ of most locations (95%) is roughly less than 4.5km in Shanghai and 6km in Beijing, such difference might be originated from different geography and urban layout of these two cities (see Fig. 3.4(a) and Fig. 3.4(c)). Those top 5% locations with long average travel distance tend to have relatively low flow volume in Shanghai, but some of them in Beijing are of high volume. We also observe that most of them are scattered at the fringe of Shanghai, which means the riders in these locations tend to travel a long distance with bikes (see Fig. 3.4(b)), with exceptions in Beijing that there are some locations relatively closer to the city centre. This suggests that bikes might be quite crucial for the users in such locations, where dockless bikes are served as more than just short distance commuting connectors. Accordingly, the locations with a short weighted average travel distance are almost distributed everywhere. The cases for weighted inflow travel distance of locations (see Fig. 3.4(b) and Fig. 3.4(d)) are similar to the situations of outflow.

3.4 Gravity model at various spatial scales

The gravity model has been widely applied to predict traffic between locations, which assumes the traffic T_{ij} is proportional to the product of the origin population P_i and destination population P_j, and inversely proportional to the distance between the centroids of these two locations d_{ij}. In this work, the population of locations are obtained from aggregating the raw data from WorldPop [183] (see Fig. 3.2(b)). The general Gravity Model can be formulated as follows [165, 169, 170, 176, 184]:

$$T_{ij} = C\frac{P_i^{\alpha} P_j^{\beta}}{f(d_{ij})} = C\frac{P_i^{\alpha} P_j^{\beta}}{d_{ij}^{\gamma}}, \tag{3.4}$$

where α, β and γ are exponents that can be fine tuned to fit the data, C is a proportionality constant term, and $f(d_{ij})$ is the deterrence function that best fit data, whose common form

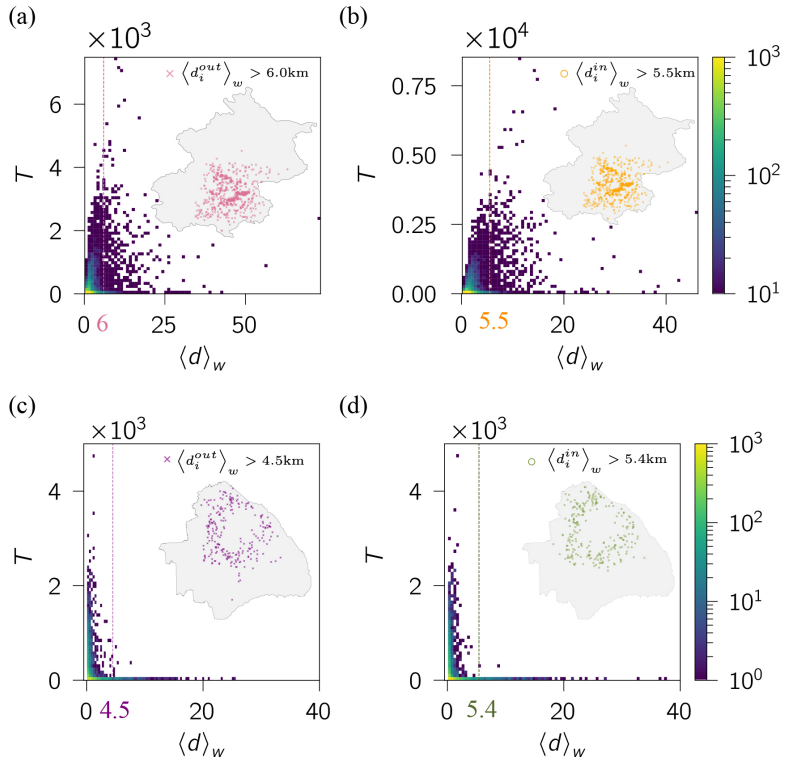


Figure 3.4: The distribution of weighted average (a, c) outflow and (b, d) inflow travel distance of locations in (a, b) Beijing and (c, d) Shanghai, where each point corresponds to a location (500 m × 500 m). The vertical dashed lines indicate 95% percentile, and the corresponding spatial distributions of top 5% locations with long average travel distance are shown as insets. We can clearly see that locations with long average travel distance tend to scatter at the fringe of Shanghai, while in Beijing, some of them are closer to the city centre. For the case in Beijing, we also increased the distance threshold to 7.5 km, and the corresponding spatial patterns are still similar but only sparser.

would be a power-law function with an exponent γ[1]. After normalisation, T_{ij} can also be interpreted as the probability of an individual transit from i to j.

We find that, for the biking traffic, such general Gravity Model (Eq. (3.4)) always holds at different spatial resolutions (see Fig. 3.5) in cities – the R^2 increases with the area of locations systematically ranging from 500 m × 500 m, 1km×1km, and all the way up to 5km×5km. The fitted values of parameters α, β ,γ, and C are shown in Table 3.1. Note that some dots on the left and right side are omitted from linear fitting in Fig. 3.5 due to the challenge on analysis posed by poor statistics, which might be originated from the finite-size effect. With the increasing of spatial scale, the flat region on the left side is becoming smaller and narrower. The saturation

[1]Sometimes, an exponential function is also applied to the distance part, e.g., $d_{ij} = exp(d_{ij}/\kappa)$ where κ is a characteristic travel length governs the decay of flow, beyond which the resistance effect posed by the increase of distance would be stronger

region (see the right side of the plot in Beijing at fine spatial resolutions Fig. 3.5(a, b)) also vanishes as the scale increases. The trend for both Beijing and Shanghai are consistent, if a larger spatial scale, then a stronger linear relation is observed (see R^2 in Table 3.1).

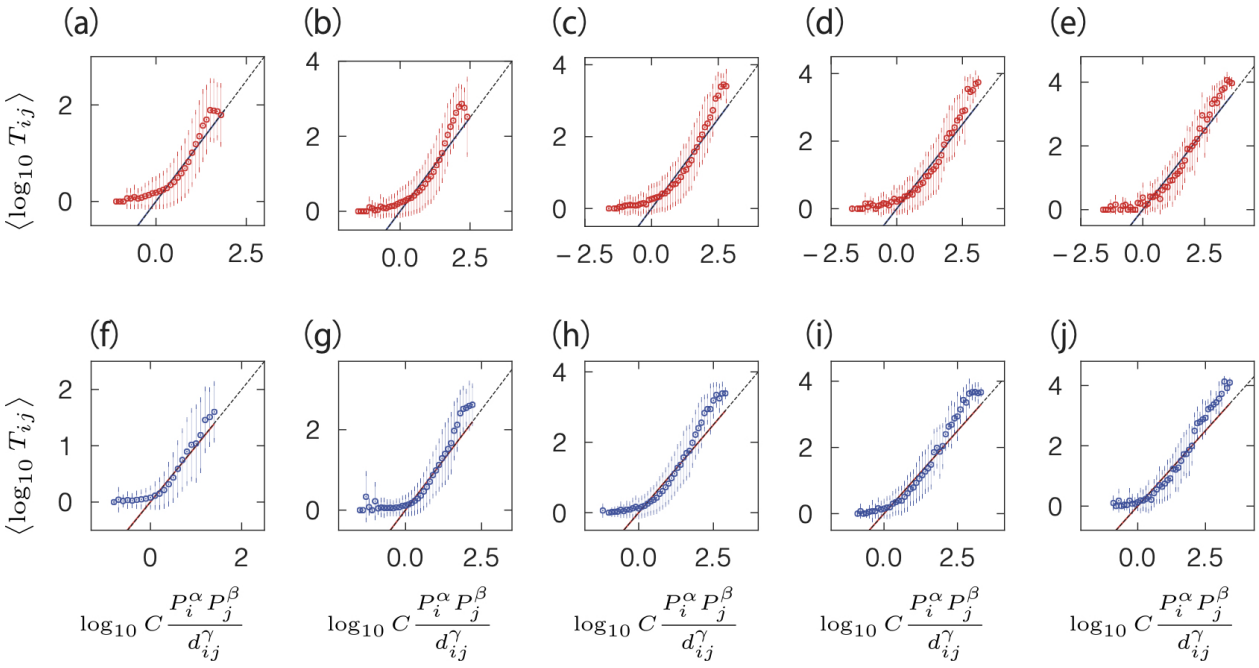


Figure 3.5: General Gravity Model on traffic volume between locations within (a-e) Beijing and (a-e) Shanghai at different spatial resolutions ranging from 5 00m to 5 km (from left to right). The fitted values of exponents are detailed in Table 3.1. The solid lines are direct regression from all data, and the black dashed lines are diagonal (i.e., $Y = X$). We can observe that solid lines and black dashed lines are almost identical, and most of data are around the diagonal, proving the effectiveness of the general Gravity Model. It's worth noting that at larger spatial scales, the right side tends to be higher than prediction which might be due to stronger nonlinear interaction effect induced by agglomeration.

Different from the commuting flow between counties in the USA where α, β are quite different ($\alpha = 0.30, \beta = 0.64, \gamma = 3.05$ reported in [169, 176]), for biking traffic, both population related exponents are of similar magnitudes (with α only slightly larger than β), which suggest the impacts of origin and destination population are relatively balanced, and both exponents are smaller than 1 at various spatial resolutions, which indicates that smaller populations are relatively more important per capita on generating biking traffic. In addition, there's a nonlinear increasing of α, β and γ along with the change of spatial scale in both Beijing and Shanghai, see Fig. 3.6(a). Though the magnitudes are quite different between distance exponent γ and population exponents α and β, the increasing trends after mean normalisation (i.e., dividing

the data by its mean) are almost identical, especially for Shanghai, which might suggest some universal relation behind population and distance in terms of spatial scale.

Validating the gravity model on biking traffic at varying scale within cities are important for better urban planning towards a greener and more sustainable urban development. However, to our best knowledge, there has been no such work focusing on the gravity model in dockless bike-sharing systems in cities.

	Shanghai					Beijing				
Resolutions	α	β	γ	C	R^2	α	β	γ	C	R^2
500 m	0.18±0.01	0.17±0.01	0.99±0.01	0.35	0.38	0.26±0.02	0.22±0.02	1.25±0.02	0.2	0.35
1 km	0.37±0.02	0.38±0.02	1.69±0.03	0.03	0.51	0.40±0.02	0.33±0.03	1.99±0.04	0.07	0.45
2 km	0.67±0.04	0.65±0.03	2.54±0.07	0.0003	0.62	0.54±0.04	0.47±0.04	3.02±0.08	0.03	0.54
3 km	0.88±0.07	0.79±0.07	3.21±0.13	1.73	0.67	0.60±0.08	0.56±0.08	3.84±0.18	0.03	0.58
5 km	0.94±0.14	0.91±0.11	3.96±0.26	9.75	0.68	0.73±0.12	0.70±0.11	5.23±0.36	0.07	0.67

Table 3.1: The fitted values of parameters of the general Gravity Model and R^2. The errors quoted are 95% of confidence interval for fitting parameters.

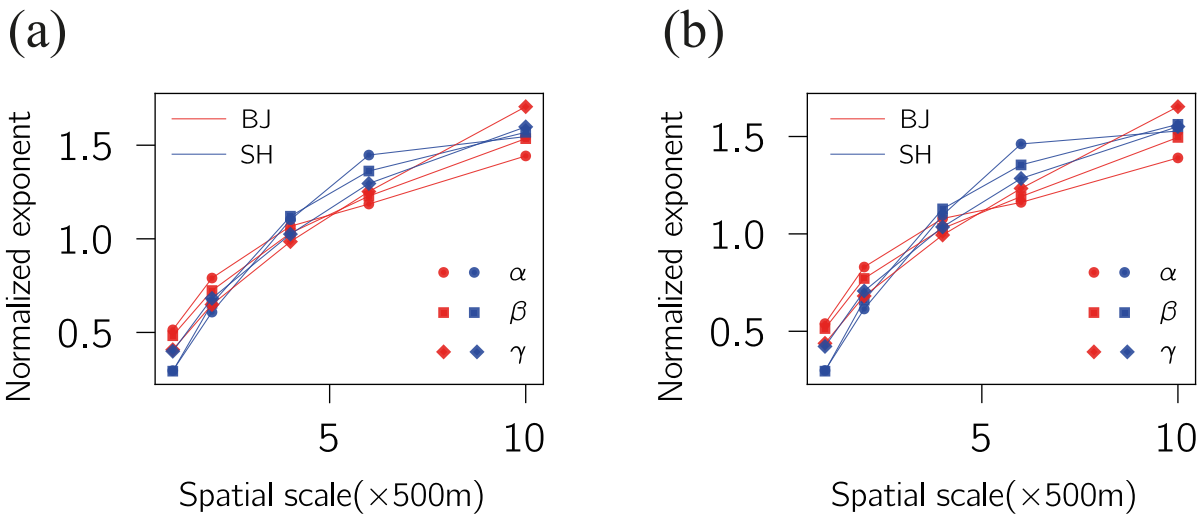


Figure 3.6: The evolution of distance and population related exponents in the (a) general Gravity Model, and (b) Gravity Model II with dimensionless distance in both Beijing and Shanghai. After mean normalization (i.e., dividing the data by mean), we observe that these parameters are almost changing in a similar trend.

In addition, we also test some other variants of the aforementioned general Gravity Model. Sometimes, the distance d_{ij} can be dimensionless by dividing by the spatial resolution d_0 (the spatial scale of locations, e.g., 500m) [181], we name it as Gravity Model II:

$$T_{ij} = C\frac{P_i^\alpha P_j^\beta}{(d_{ij}/d_0)^\gamma} = Cd_0^\gamma \frac{P_i^\alpha P_j^\beta}{d_{ij}^\gamma}. \tag{3.5}$$

We find that such a dimensionless distance setting almost doesn't alter the results, which are quite comparable with the general Gravity Model (see Fig. 3.7). The intrinsic spatial scale still plays an important role. We only observe a little bit larger fluctuations on the right sides at larger spatial scale (3km to 5km), the values of R^2 are also quite comparable with the general Gravity Model. Moreover, the trends of evolution on distance and population related exponents are almost identical with the general Gravity Model, though the magnitudes are not precisely the same, see Fig. 3.6(b).

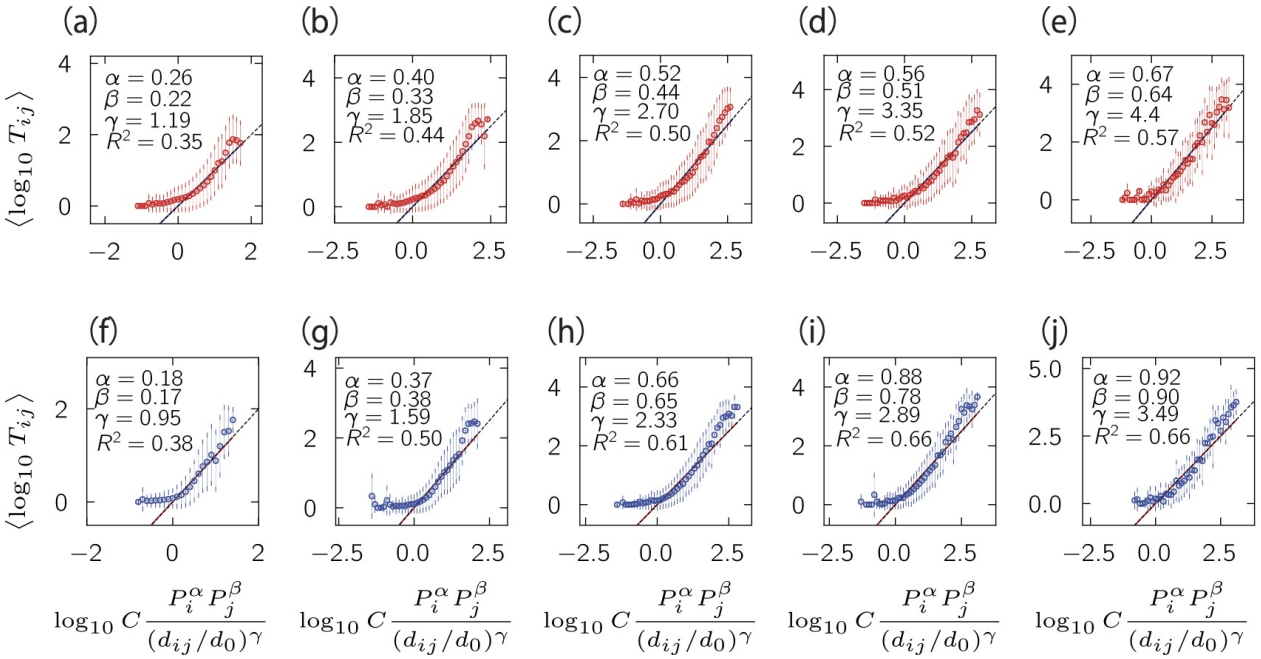


Figure 3.7: Gravity Model II with dimensionless distance on traffic volume between locations within (a-e) Beijing and (f-j) Shanghai. All the settings of figures are the same with Fig. 3.5. The effectiveness of Gravity Model II is also validated with most data around diagonal lines and the regression line is collapsing with the diagonal. We only observe a little bit larger fluctuations on the right sides at larger spatial scale (3km to 5km), the values of R^2 are also quite comparable with the general Gravity Model.

For other variations, we can fix both population exponents α, β as 1 (named as Gravity Model III), which can be derived from entropy maximization [185]:

$$T_{ij} = C \frac{P_i P_j}{d_{ij}^\gamma}, \tag{3.6}$$

when $\gamma = 2$, it simplifies to the form of the Newton's law of the gravity. Yet as indicated in several datasets, $\gamma \in [0.5, 3]$ which is not necessarily equal to 2 [184]. We find the original

Gravity Model and Gravity Model III doesn't hold well for smaller spatial scales (from 500 m to 1 km) with larger fluctuations and larger deviations from the diagonal lines. Besides, the flat region on the left side is longer and wider. And it's worth noting that the regression lines are not overlapping with diagonal lines anymore, which further indicates a systematical deviation between the prediction and reality (see Fig. 3.8).

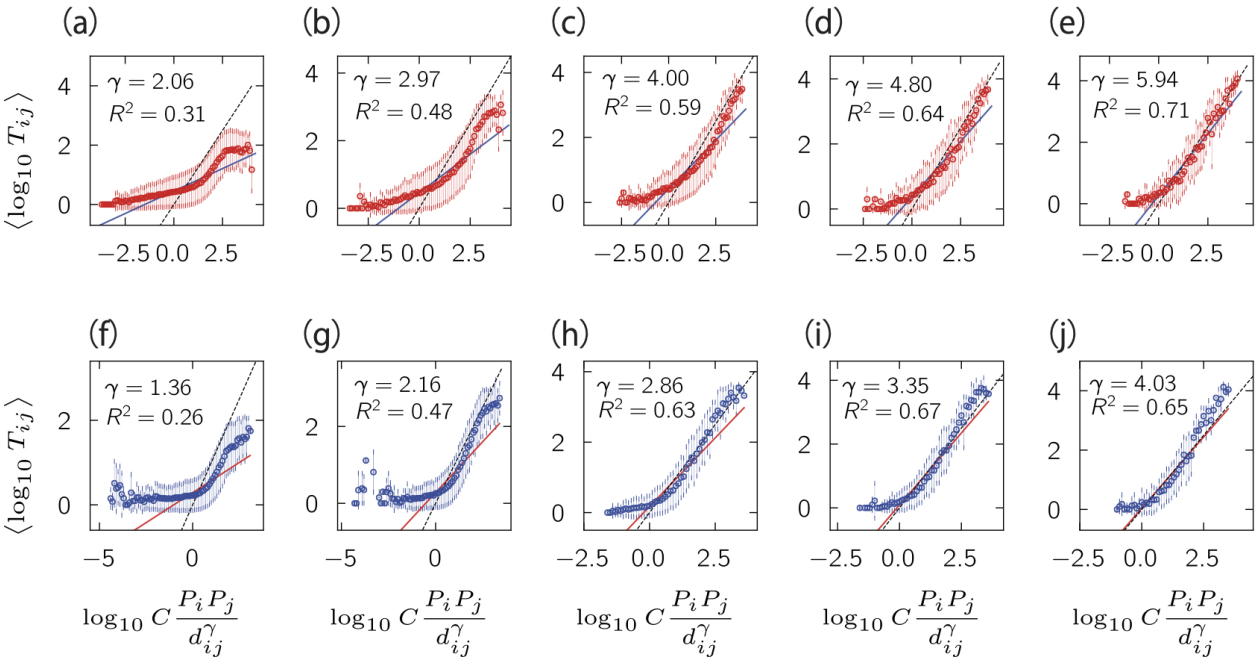


Figure 3.8: Gravity Model III on traffic volume between locations within (a-e) Beijing and (f-j) Shanghai. All the settings of figures are the same with Fig. 3.5. Gravity Model III doesn't hold very well, especially at finer spatial resolutions (500 m to 1 km).

A further variation of Gravity Model III with a dimensionless distance (we label it as Gravity Model IV) has a similar bad performance (even worse in Beijing at a 500 m spatial resolution).

By comprehensive comparisons between results reported in Figs. 3.5-3.9, we find that the form of distance (whether geometric or dimensionless) doesn't pose a significant influence, the most important factors are the population related exponents at finer spatial resolutions (500 m to 2 km), where a linear population related attractions are not a good approximation of the real situations. This indicates that at smaller scale, with a less than unity population related exponents, smaller populations are relatively more important per capita on generating biking traffic. Only when the spatial scales are close to 3km to 5km, then the population exponents are closer to one.

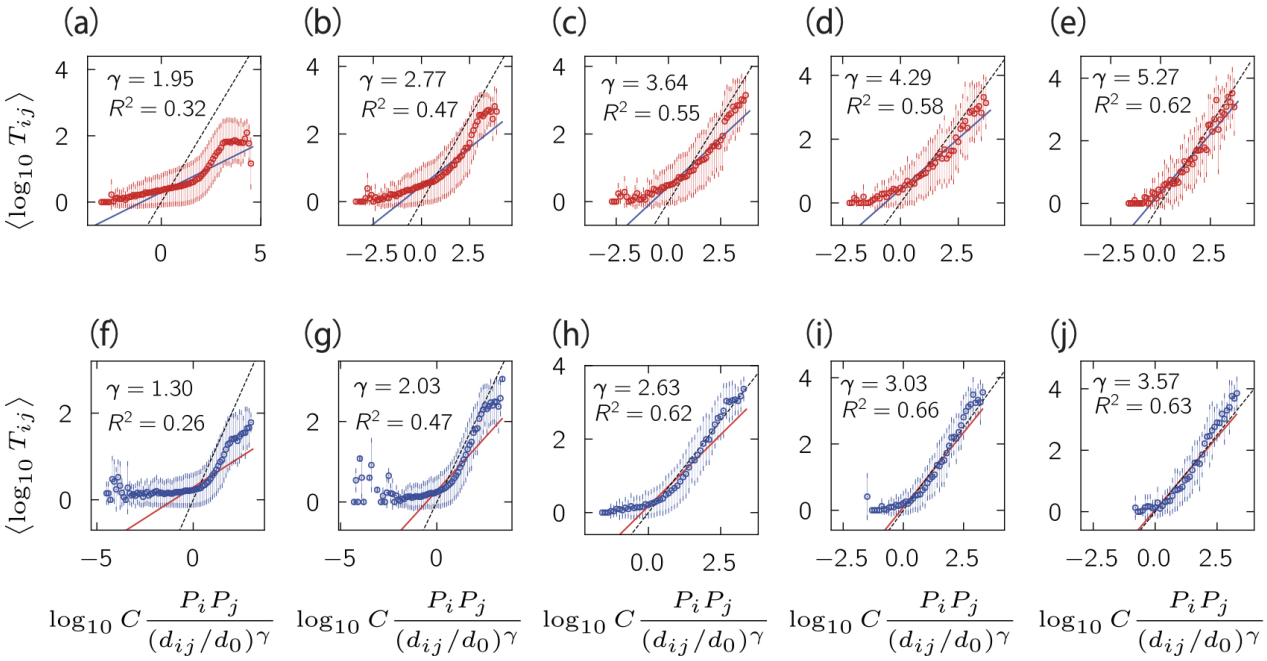


Figure 3.9: Gravity Model IV with dimensionless distance on traffic volume between locations within (a-e) Beijing and (f-j) Shanghai. All the settings of figures are the same with Fig. 3.5. Gravity Model IV also doesn't hold very well, especially at finer spatial resolutions (500 m to 1 km).

3.5 Sources and sinks

The gravity model assumes the interaction between locations is bidirectional and equivalent if population related parameters $\alpha = \beta$, i.e., the traffic volume from location i to j would be the same as j to i. However, among all locations, there are two types of locations – sources and sinks – with only unidirectional flows but worth further studies due to their challenges posed on the operation of dockless bike-sharing systems, and on enhancing the user experience by providing more user friendly and easier access to transport means.

Based on the flow profile of locations (i.e., whether a location has inflow, outflow and inner-flow traffic or not), we divide them into seven distinct types of flows(see Fig. 3.10(a, f), note that we didn't take locations without any type of biking traffic into consideration): locations with only inflow (sinks, see Fig. 3.10(a) and Fig. 3.10(f)), only outflow (sources), only inner-flow, and locations with both inflow and outflow with inner-flow or without inner-flow, and locations with inner-flow and with just either inflow or outflow. Locations with both inflow and outflow (type D and G in Fig. 3.10(a) and Fig. 3.10(f)) are quite dominant. And inflow and outflow

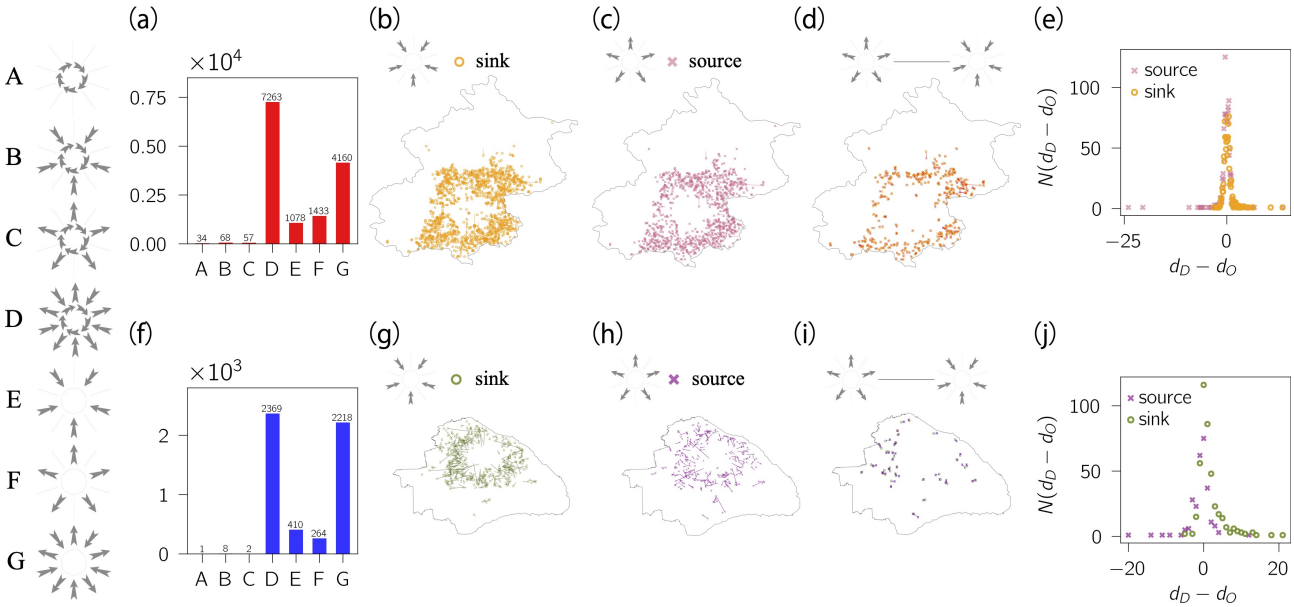


Figure 3.10: (a, f). The spatial distribution of sinks and their directional inflow (b, g); and the spatial distribution of sources and their directional outflow (c, h). The connections between sources and sinks (d, i), where source locations are marked as ×, and sinks as ○; if there's biking flow between any of them, they will be connected. Travel direction and distance relative to the city centre (e, j) in both Beijing (a-e) and Shanghai (f-j).

of locations are in a great linear relation at various spatial scales (especially in Shanghai, the case in Beijing are of relatively larger fluctuations, see Fig. 3.11).

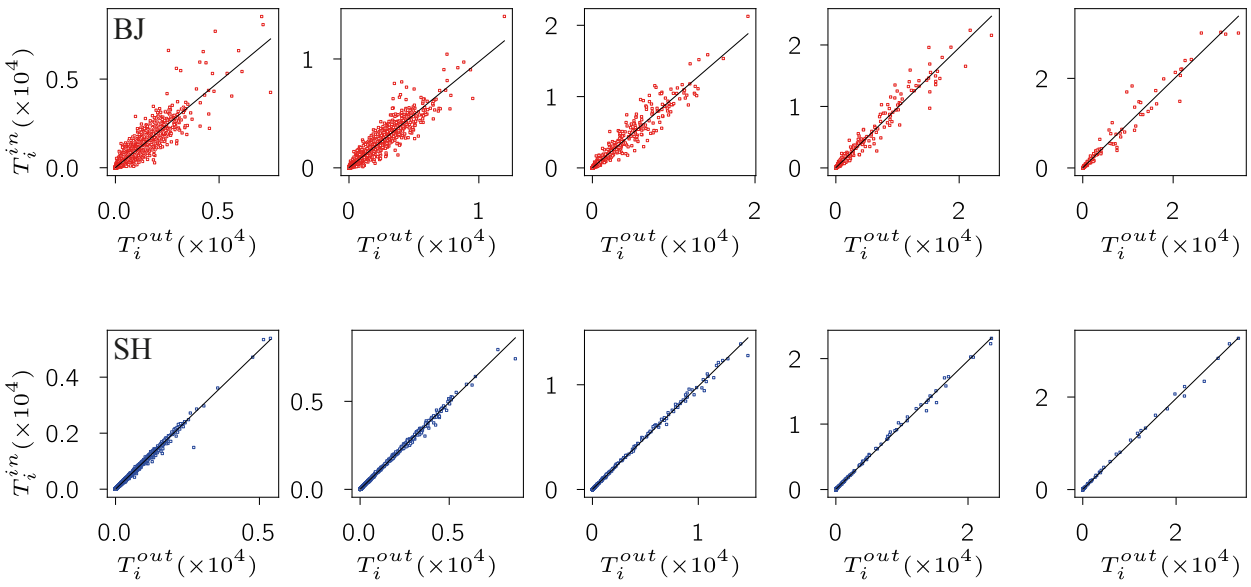


Figure 3.11: (The inflow and outflow of locations with varying spatial resolutions ranging from 500 m to 5 km, from left to right) in both (a) Beijing and (b) Shanghai, which is in a great linear relation.

Sources and sinks (type E and F in Fig. 3.10) are also non-negligible at a spatial resolution of

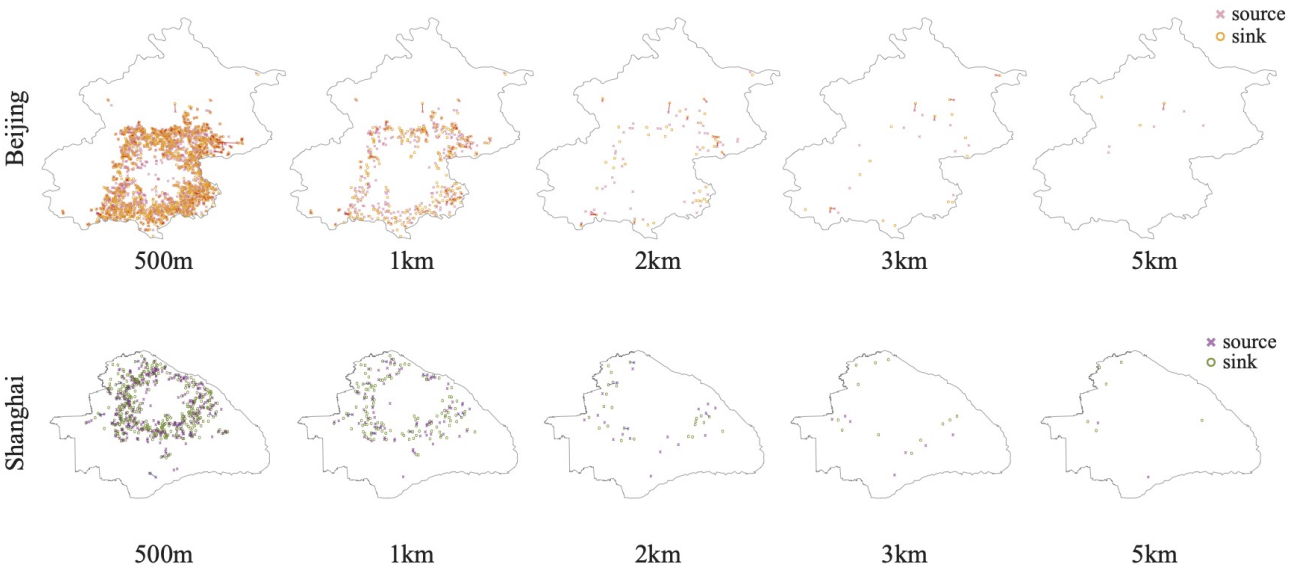


Figure 3.12: The evolution of sources and sinks with varying spatial resolutions of locations, ranging from 500 m × 500 m to 5 km × 5 km (from left to right). We can observe that both source and sink locations are vanishing rapidly, mainly due to the increase of inner flows.

500 m × 500 m. By looking at the spatial distribution of the sources and sinks in both Beijing and Shanghai, we find that most of them are distributed at the fringe of the city. If a sink location is one of the destinations of a source location, then we connect them with a line in the figure, yet we can see that only a few sources and sinks can be linked together (see Fig. 3.10(d, i)).

As a source location may have several destinations, we calculate the directional flow from a source location weighted by the biking traffic as $\vec{F}_i = \sum_j w_{ij}(\vec{l}_j - \vec{l}_i)$, where $\vec{l}_i = (\text{longitude}_i, \text{latitude}_i)$ is the position vector of location i, or vice versa for a destination location (see Fig. 3.10(b), Fig. 3.10(c), Fig. 3.10(g) and Fig. 3.10(h)). We also looked at the average travel direction of sources and sinks relative to the city centre, i.e., whether the riders there tend to move toward or away from the city centre. We calculate the distance between the city centre and the origin (destination) of each trip that starts from a source location (or ends at a sink location), denoting it as d_O (d_D). Then $d_D - d_O$ can reflect whether riders are moving farther away ($d_D - d_O > 0$) or closer to ($d_D - d_O < 0$) the city centre. Their distributions do not show strong trends or significant differences at small values, see Fig. 3.10(e) and Fig. 3.10(j). However, for sinks, we can observe some long distance travel away from the city centre (indicated by positive values of $d_D - d_O$), while the opposite for sources. This indicates that there are still some users in

such locations are riding for long distance towards the city centre, and back home at suburbs probably from the workplace near downtown.

It would be natural to raise following questions: Why do sources and sinks exist? Does a rider depart from a source location never come back, or vice versa? By varying the spatial resolution from 500m to 5km, we can observe that both source and sink locations are vanishing rapidly, see Fig. 3.12, mainly due to the increasing of inner-flow, as can be observed in Fig. 3.10(b), Fig. 3.10(c), Fig. 3.10(d), Fig. 3.10(g), Fig. 3.10(h) and Fig. 3.10(i), that most of $|\vec{F}_i|$, as indicated by the length of the line connected to it, is not too large, and most of connected sources and sinks are not too far away from each other. Yet a 500 m resolution is more informative on riding behaviours, the existence of a non-negligible amount of sources and sinks may originate from the nature of riding behaviours and sharing economy, and require a further study on related operation strategies.

3.6 Conclusions and discussions

In this chapter, we find that the spatial distribution of biking activities is in line with spatial scaling theory, and further investigate the distribution of biking flow in cities at a fine spatial resolution, which manifests a power-law. The dependence of biking flow distribution on distance are also fat tailed above 1km in both Beijing and Shanghai. Dockless sharing bikes may serve as more than just short distance commuting connectors for users in some locations as indicated by a long average weighted travel distance. We validate that the general Gravity Model can be well applied for predicting biking traffic within cities at fine spatial resolutions, which indicates that the human mobility patterns to a large extend have been preserved even the transportation means are changed. The scale effect of location is also studied, though distance related exponent is of larger magnitude, it grows in a similar trend as population exponents as the spatial scale of locations increases, which suggests that there might be some underlying relation between them. And the population related exponents are less than unity especially at finer spatial resolutions which is further confirmed by comparing with Gravity Model 3 and Gravity Model 4 where

population related exponents $\alpha = \beta = 1$. This indicates that smaller population are relatively more important per capita on generating biking traffic. The Gravity Models provide us with a method to understand the biking behaviours in different cities, and is of great importance to related urban practices, for instance, how to relocate number of bikes at different locations. In addition, we also look at the directionality issue [186], as the Gravity Models predict an equivalent mutual flow between two locations, yet in reality, this is not always the case, which requires further studies. Taking a closer look at the source and sink locations, we notice most of them are not connected with each other, and they don't show a clear directional flow relative to the city centre for short-distance travel; most of long-distance trips to sinks are away from the city centre, while most of long-distance trips from sources are towards downtown. However, we do not have the detailed information about the region so we can not know the destination of each short distance ride, for instance, it is a pub or restaurant. Overall, we can see the success of the gravity model.

3.7 Summary

In this chapter, we

- used spatial data from the mobike to study bicycle mobility pattern in Shanghai and Beijing.
- revealed a spatial scaling relation between the cumulative volume of riding activities $V(r)$ and corresponding distance to the city centre r, $V(r) \sim r^{2-\eta}$ for the region far away from the city boundary $r \ll R$.
- found a power-law distribution on the volume of biking flows between fine-grained locations in both Beijing and Shanghai.
- validated the effectiveness of the general Gravity Model on predicting biking traffic at fine spatial resolutions, where population-related parameters were less than unity indicating

that smaller populations are relatively more important per capita on generating biking traffic.

- further studied the impacts of spatial scale on the Gravity Model.
- revealed that the distance-related parameter grows in a similar way as population-related parameters when the spatial scale of locations increases. The flows between different location is proportional to the population product across different scales, however, the scaling exponents are not identical across different scales in terms of self-similarity.

Chapter 4

Linking centrality measures: Closeness and degree

The following chapter is based on:

T., Evans, & B., Chen

Linking the Network Centrality Measures:Closeness and Degree.

Submitted to Communication Physics.

B. Chen contributed to all aspects of this paper and this chapter.

4.1 Introduction

Centrality measures look at the importance of nodes in a network, from social science to epidemiology studies to web site ranking and are used in almost every application of network (see more descriptions in section 1.4.1 in Chapter 1). There are a vast number of centrality indices, as Schoch's "Periodic Table of Network Centrality" illustrates [16] or see [21, 24]. It is clear that many different centrality indices encode similar information about node importance so there is a great deal of redundancy as quantified by the correlation found between centrality indices [21, 24]. Studies focus on linear correlations between centrality measures, e.g. Pearson correlation coefficients. However, the explanations for these correlations are missing.

In this chapter, we focus on two of the most popular centrality measures. Degree is a local measure that counts the number of neighbours of a node. Closeness probes the global properties of a network as it is the inverse of the sum of shortest path distances to other nodes. Since closeness is a measure of shortest paths, one of the most natural structures to think about is the tree structures. Based on the tree structure approximation, we conjecture that the inverse of closeness of a node is linearly dependent on the logarithm of the degree of that node. Our conjecture explains why high values of linear correlation measures are often reported between degree and closeness centrality but at the same time our conjecture shows that linear correlation indices have missed important non-linear features in the landscape of centrality measures.

In the current chapter, we use the construction that the shortest paths from any one node to all other nodes can be arranged as a spanning tree. We then conjecture that the branches of this tree are statistically similar, implying that the closeness of a node can only depend on the number of such branches, i.e., the degree of the node. While assuming that the number of nodes in each branch grows exponentially, we arrive at the non-linear form of our conjecture. This allow us to derive the non-linear relationship between closeness and degree. We then test our conjecture on many artificial and real data sets and show that our relationship works in most real networks.

4.2 Theory

4.2.1 General definitions

For simplicity, we will assume throughout this chapter that we are analysing a simple graph $\mathcal{G}$ with just one connected component. We will denote the number of nodes as N and the degree of each node v as k_v. In a network, a path length ℓ is a sequence of $(\ell + 1)$ distinct nodes $\{v_i\}$ $(i = 0, 1, 2, \ldots, \ell)$ such that each consecutive pair of nodes in the path is connected by an edge. We will define the distance between two nodes u and v in a network to be the length of a shortest path between two nodes, denoted here as d_{uv}.

The *closeness* c_v [53, 193, 194] of a vertex v is then defined to be the inverse of the average distance from v to every other vertex in the graph, so

$$\frac{1}{c_v} = \frac{1}{(N-1)} \sum_{u \in \mathcal{V}} d_{uv}, \tag{4.1}$$

where $\mathcal{V}$ is the set of nodes, $|\mathcal{V}| = N$. Clearly the closer a vertex v is to other vertices in the network, the larger the closeness, so this measure mimics the properties we expect when defining the centre of a geometric shape so making closeness a natural centrality measure.

4.2.2 Estimate of closeness

We start from the idea that some of the statistical properties of real-world networks may be captured by spanning trees [192]. Trees [193, 194] are connected networks with no loops so the number of edges is always one less than the number of nodes. A *spanning tree* [192] is a connected subgraph of the original graph $\mathcal{G}$ containing all the original vertices $\mathcal{V}$ but a subset of $N-1$ edges that are just sufficient to keep every node connected to all others. In particular, we work with *rooted trees* $\mathcal{T}(r)$ in which we have singled out one special node, the *root* r of the tree. Since we are interested in closeness which uses the lengths of shortest paths between nodes, the most useful trees for this work are the *shortest-path trees*, $\mathcal{T}(r)$, that contain one shortest path from a root node r to each remaining node in the network. As our networks are unweighted, the shortest-path trees always exists and are easily defined as part of a breadth-first search algorithm. Breadth-first search algorithm traverse each node once until all the nodes in the network are connected from the root. Every node can act as a root node so there is at least one shortest-path tree, $\mathcal{T}(r)$, for every node r. These trees are not in general unique as there can be many shortest paths between a pair of nodes.

The picture used here, as shown in Figure 4.1, is that close to the root node the structure of these shortest-path trees will vary. In many networks, as we move further away from the root node the number of nodes $n_r(\ell)$ at some distance ℓ from root node r grows exponentially with each step in most networks. This is the origin of the small-world effects seen in many networks,

the way the distance between nodes is typically much smaller than would be found in networks constrained by Euclidean geometry [71]. This means that regardless of the local context of a root node, the trees quickly access a similar set of nodes in the main bulk of the network. Thus we conjecture that the structural and statistical properties of these trees away from the root node are likely to be similar for all possible root nodes. The contribution to closeness of each node in the bulk is bigger as they are further from the root and more numerous. So we might expect that the largest contributions to closeness always come from the same bulk regions where we can expect statistical similarity.

The most important difference while comparing different root nodes is the initial value for the exponential growth in the number of nodes at distance ℓ from the root. This will depend on the local structure with the simplest effect coming from the number of immediate neighbours the root node has, i.e., the degree of the root node k_r. That is the simplest approximation for the growth of these shortest-path trees is:

$$n_r(\ell) = k_r \bar{z}^{\ell}, \tag{4.2}$$

where $\bar{z}$ is some measure of the rate of growth of the shortest-path tree. Note that our assumption of statistical similarity suggests that the branching factor of these trees is, on average, the same so we use a single parameter $\bar{z}$ to represent the growth from any root node r.

One key assumption we make is that *all* these branches have similar statistical properties because the vast majority of such branches are in the 'bulk' of the network. To start with, we will assume that in terms of our measurements $\mathcal{T}(v_\ell, r) \equiv \mathcal{T}(v'_\ell, r)$ for any two nodes v_ℓ and v'_ℓ distance ℓ from our root. For the same reason, in most graphs we might expect these subgraphs to be similar whatever the root node was so $\mathcal{T}(v_\ell, r) \equiv \mathcal{T}_\ell$.

The exponential growth in the number of nodes distance ℓ from *any* root node, as encoded in Eq.(4.2) is our second key assumption. This will not be true for networks embedded on a plane or other Euclidean spaces as there we expect $n(\ell)$ to measure the surface area of a shape of radius ℓ which would follow a power law $n_\ell \sim \ell^{D-1}$ for a D-dimensional Euclidean space.

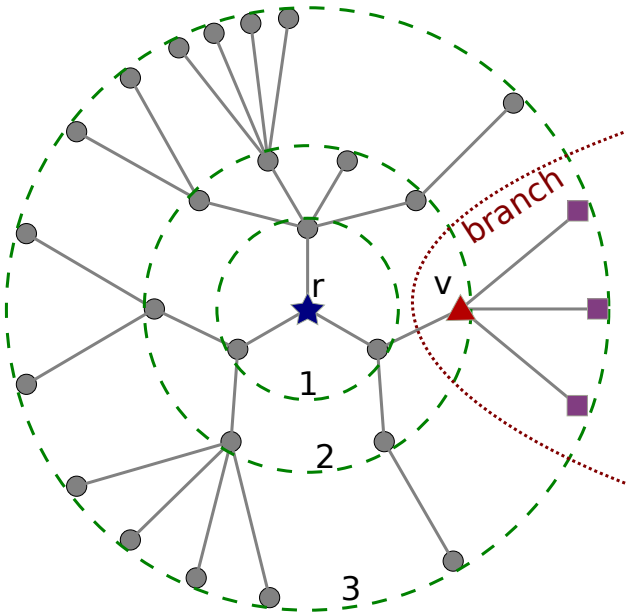


Figure 4.1: An example of a rooted tree $\mathcal{T}(r)$ defined in terms of a root node r, the blue star at the centre. All nodes at the same network distance from the root node, are placed at the same distance from the root node in this visualisation, as indicated by the green dashed circles. The red triangle, node v, is the root of a *branch* $\mathcal{T}(v,r)$, a smaller tree containing all the nodes u which lie on a path from the root through v (which is also a shortest path from r to u in the full graph $\mathcal{G}$). These nodes u in the branch are therefore further from the root than v, $d_{ur} \geq d_{vr}$. The branch $\mathcal{T}(v,r)$ illustrated here includes v (red triangle), the nodes indicated with purple squares and the edges between these nodes. The degree of a node is the number of neighbours so here node v has degree 4.

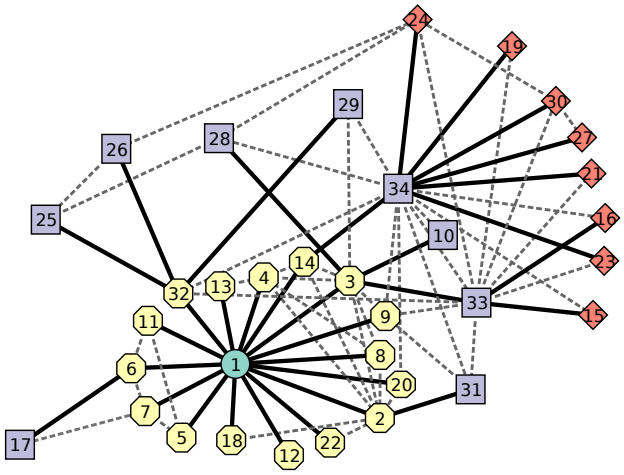


Figure 4.2: The Zachary Karate club network [85]. The nodes of the same colour and shape have shortest paths to node 1 of the same length. The thick black lines are edges which are part of one possible shortest-path tree $\mathcal{T}(1)$ with node index 1 as the root node of the tree. The dashed grey lines indicate edges not in the tree. These shortest-path trees are not unique as can be seen here since we can include edge $(7,17)$ in $\mathcal{T}(1)$ instead of the edge $(6,17)$ used here. Node labels correspond to those used in [85].

To get a network of significant size, we need the branching factor in each layer l, $\bar{z}_\ell > 1$. However, the total number of nodes N is given by:

$$N = \sum_{\ell=0}^{L} n_\ell(r), \tag{4.3}$$

where L is the cut-off distance. Therefore, a model with constant branching factor for all layer $\bar{z}_\ell = \bar{z} > 1$, the number of nodes grows exponentially fast with finite maximum distance. The crudest solution, and the one we will follow here, is to assume that $\bar{z}_\ell = \bar{z}$ for $2 \leq \ell \leq L_r$ where L_r is some long distance cutoff which may depend on the root vertex r considered with $\bar{z}_\ell = 0$ for larger ℓ. The $\bar{z}_\ell = \bar{z}$ assumptions implies effective branching number at the each layer are

the same[1]. So we work with the following model

$$n_\ell(r) = \begin{cases} 1 & \text{if} \quad \ell = 0\,, \\ \bar{z}^{\ell-1}k_r & \text{if} \quad 1 \le \ell \le L_r\,, \\ 0 & \text{if} \quad \ell > L_r\,. \end{cases} \tag{4.4}$$

To improve clarity of the expressions, we will now drop the explicit dependence on the root vertex r chosen and write $c \equiv c_r$, $k \equiv k_r$, and $L \equiv L_r$.

We will determine the distance cutoff L by imposing Eq.(4.3). Our model for $\bar{z}_\ell$ now becomes

$$N = 1 + \sum_{\ell=1}^{L} \bar{z}^{\ell-1}k = 1 + k\frac{(\bar{z}^L - 1)}{(\bar{z} - 1)}\,. \tag{4.5}$$

Inverting this we see that the distance cutoff L we need is given by, for large N,

$$L(N,k) \approx \frac{\ln\left(N(\bar{z}-1)/k\right)}{\ln \bar{z}}\,. \tag{4.6}$$

Even in this simplest approximation, it is clear that the distance cutoff L depends on our choice of root vertex through the degree of the root node.

In principle L in Eq.(4.4) and Eq.(4.5) is an integer but it is clear from the form in Eq.(4.6) that we need L to be a real number, in some sense an average over the actual distances from the root to the leaves (nodes with degree one) of the tree. So the real number valued L given by Eq.(4.6) sets the scale of the distance beyond which the terms in these sums become negligible since almost all the nodes are connected. We also note as an aside that L depends on the size of the network through a $\ln(N)$ factor, not as a power $N^{1/D}$, and this is the classic "small-world" effect seen in many network length scales such as diameter and average distance. We will assume our network is connected so a shortest path exists between all node pairs in the network. We can now rewrite the closeness c of a root vertex r using Eq.(4.4) and Eq.(4.10) to

[1]This is different from the degree number but counting how many new nodes a node connect, thus, called effective.

gives us:

$$\frac{1}{c} = \frac{1}{\Omega}\sum_{\ell=1}^{L} \ell n_\ell \,, \tag{4.7a}$$

$$\Omega = \sum_{\ell=1}^{L} n_\ell \,. \tag{4.7b}$$

For the normalisation Ω we have that $\Omega = N-1$ and using Eq.(4.5) we can express this in terms of our other variables

$$\Omega \;=\; \sum_{\ell=1}^{L} k\bar{z}^{\ell-1} = k\frac{(\bar{z}^L-1)}{(\bar{z}-1)} = N-1\,. \tag{4.8}$$

The exponential branching are unrealistic in many real networks. From Eq.(4.6), we derived an upper cutoff L_r and assume that $n(\ell) = 0$ for $\ell > L$, which means, after all nodes are reached, the branching process stopped.

Note this gives us a link between L, N and $\bar{z}$. We will eventually use Eq.(4.8) to eliminate L as we assume N is known. However the expressions are simpler in terms of L so here we will use Eq.(4.8) to eliminate N and Ω.

Using Eq.(4.4) and Eq.(4.10) gives us that

$$\begin{aligned}
\frac{1}{c} &= \frac{1}{\Omega}\sum_{\ell=0}^{L} k\ell\bar{z}^{\ell-1} \\
&= \frac{k}{\Omega}\frac{d}{d\bar{z}}\sum_{\ell=0}^{L}\bar{z}^{\ell} \\
&= \frac{k}{\Omega}\frac{d}{d\bar{z}}\left(\frac{\bar{z}^{L+1}-1}{\bar{z}-1}\right) \\
&= \frac{k}{\Omega}\left(\frac{(L+1)\bar{z}^{L}}{\bar{z}-1} - \frac{(\bar{z}^{L+1}-1)}{(\bar{z}-1)^2}\right).
\end{aligned} \tag{4.9}$$

Using Eq.(4.8) to eliminate Ω (i.e. N) in terms of L and $\bar{z}$, we have that

$$\begin{aligned}\frac{1}{c} &= \frac{\bar{z}-1}{\bar{z}^L-1}\left[\frac{(L+1)\bar{z}^L}{\bar{z}-1} - \frac{(\bar{z}^{L+1}-1)}{(\bar{z}-1)^2}\right] \\ &= \frac{L\bar{z}^L}{(\bar{z}^L-1)} + \frac{\bar{z}^L}{(\bar{z}^L-1)} - \frac{(\bar{z}^{L+1}-1)}{(\bar{z}^L-1)}\frac{1}{(\bar{z}-1)} \\ &= L\left(1+\frac{1}{\bar{z}^L-1}\right) + \frac{1}{(\bar{z}-1)}\frac{1}{(\bar{z}^L-1)}\left(\bar{z}^L(\bar{z}-1) - (\bar{z}^{L+1}-1)\right) \\ &= L\left(1+\frac{1}{\bar{z}^L-1}\right) - \frac{1}{\bar{z}-1}.\end{aligned} \tag{4.10}$$

Now we can use Eq.(4.10) to produce a prediction of the relationship between the closeness of a node and its degree, also showing how closeness should vary with the size of the network and we find that

$$\frac{1}{c} \approx \left(-\frac{1}{\bar{z}-1} + \frac{\ln(\bar{z}-1)}{\ln\bar{z}}\right) + \frac{1}{\ln\bar{z}}\ln N - \frac{1}{\ln\bar{z}}\ln k + O\left(\frac{\ln N}{N}\right). \tag{4.11}$$

We now restore the dependence on the root vertex in our notation to emphasises which quantities depend on this choice and which are fixed network values. The prediction is that the inverse of closeness c_v of any node v should show a linear dependence on the logarithm of the degree k_v of that node with a slope that is the inverse of the log of the branching ratio parameter, that is

$$\frac{1}{c_v} = -\frac{1}{\ln\bar{z}}\ln k_v + \beta\,. \tag{4.12}$$

Our calculation suggests that the parameter β is a function of other known parameters but that it is also independent of the vertex v chosen, so that

$$\beta = \beta(\bar{z}, N) = \left(-\frac{1}{(\bar{z}-1)} + \frac{\ln(\bar{z}-1)}{\ln\bar{z}}\right) + \frac{1}{\ln\bar{z}}\ln N\,. \tag{4.13}$$

In our analysis we will not assume the parameter β is given by Eq.(4.13). By adding one additional parameter we lose a little predictive power but with many parameters needed to characterise a network the loss is negligible. This leaves us with a conjecture based on the number of nodes N and degree of each nodes k_v which are usually known. Then in principle

we have two unknown global parameter values which we find from a linear fit to our data for c_v and k_v giving $\bar{z}^{(\mathrm{fit})}$ and $\beta^{(\mathrm{fit})}$.

One of the earliest questions asked about networks was about the typical length scale of a network. The length d_{uv} of the shortest path between two nodes u and v provides the natural measure of distance, as it satisfies both the mathematical criteria for a distance function and our own intuition about the importance of short paths to minimise the costs and losses in communication in real networks.

Characteristic length scales are important in any system and for networks the average distance between all nodes pairs $\langle \ell \rangle$ is such a length scale where

$$\langle \ell \rangle = \frac{1}{N(N-1)} \sum_{(u,v)\in\mathcal{E}} d_{rv} \,. \tag{4.14}$$

This is the mathematical quantity that represents Milgram's "six-degrees of separation" [197, 198] and investigated on large scales in modern data sets, for instance [199, 200, 201]. It has been the focus of great interest in some of the earliest theoretical papers such as [202, 203, 204].

The average distance of a network $\langle \ell \rangle$ has a simple relationship to the closeness considered in this chapter as average distance $\langle \ell \rangle$ is simply half the average over all nodes of the inverse closeness

$$\langle \ell \rangle = \frac{1}{2N} \sum_{v\in\mathcal{V}} (c_v)^{-1} \,. \tag{4.15}$$

Remember that the edge $(r, v) \equiv (v, r)$ appears once in the edge set $\mathcal{E}$ but the term with $d_{rv} = d_{vr}$ appears twice in the sum over root nodes r in the second expression. If we insert our expression Eq. (4.12) we have that

$$\begin{aligned} \langle \ell \rangle &= -\frac{1}{N}\frac{1}{\ln \bar{z}} \sum_v \ln k_v + \beta & (4.16) \\ &= \frac{1}{\ln \bar{z}} \ln N - \frac{1}{\ln \bar{z}} \langle \ln k \rangle - \frac{1}{\bar{z}-1} + \frac{\ln \bar{z} - 1}{\ln \bar{z}} & (4.17) \end{aligned}$$

where we have used β from Eq.(4.13). This gives for large N and fixed degree distribution that

$$\lim_{N\to\infty} \langle \ell \rangle = \frac{1}{\ln \bar{z}} \ln N \,. \tag{4.18}$$

4.3 Descriptions of datasets

Before we analyse the data set, we need to describe the data set. We use a variety of networks for which data is openly available. In our case all but one can be found on KONECT [83] but many can be found on other sites with network data. Our aim is to find networks of different sizes representing contrasting types of interaction which we break down into five broad categories: social networks (`social-...`), communication networks (`commun-...`), citation networks (`citation-...`), co-author networks (`coauth-...`), and hyperlink networks (`hyperlink-...`). These networks have been used in many contexts in other publications but we will only give a brief summary of each one.

In each case, we create a simple graph, ignoring edge directions and weights, node types, time stamps, and any other such information. We take the largest connected component (LCC) of the graph and performed our analysis on this. Some basic statistics on each graph is given in Table 4.1 and then more detailed information on each data set follows.

Social networks

Social networks capture the social interactions between actors, such as friends, colleagues, clients and students. We use six data sets, the size of networks ranged from 34 to 2539 nodes. On average, we find the mean shortest distance are quite small compare other type of networks (apart from `social-health` dataset).

The `social-karate-club` is the well-known and much-used Zachary karate club dataset. The original data was collected from the members of a university karate club by Wayne Zachary in 1977 [85] and each edge represents some type of social interaction between two members of the

Network Name	Number of nodes	Number of edges	Mean distance
social-karate-club	34	78	2.44
social-jazz	198	2472	2.21
social-hamster	1858	12534	3.39
social-oz	217	2672	2.33
social-highschool	70	366	2.66
social-health	2539	12969	4.52
commun-email	1133	5451	3.65
commun-UC-message	1899	59835	3.07
commun-EU(core)-email	1005	25571	2.59
commun-DNC-email	2029	39264	3.37
commun-DIGG-reply	30398	87627	4.68
citation-DBLP-cite	12590	49759	4.37
citation-Cora	23166	91500	5.74
coauthor-astro-ph	16046	121251	5.10
coauthor-netscience	1461	2742	6.28
coauthor-pajek	6927	11850	3.79
hyperlink-polblog	1224	33430	2.75
hyperlink-blogs	1224	19025	2.72

Table 4.1: Summary statistics for the data sets used in Chapter 4. The mean distance is the average length of the shortest paths between all pairs of nodes in the largest connected components in the data.

club.

The social-highschool network represents friendships between boys in a small highschool in Illinois, USA. Each boy was asked once in the fall of 1957 and the spring of 1958. This dataset aggregates the results from both dates. A node represents a boy and an edge between two boys shows that at least one boy chose the other as a friend. The original network [86] is directed, weighted and allows multiple edges but we use a simple graph version here.

The social-hamster network comes from the Koblenz Network Collection (KONECT) [83] where it is described as the "Hamsterster households network dataset" but no further information is provided.

The social-jazz network is the collaboration network between Jazz musicians. Each node is a Jazz musician and an edge denotes that two musicians have played together in a band [87].

The social-oz network is a network recording the friendships between 217 residents living at a residence hall located on the Australian National University campus [88]. A node represents

a person and an edge represents the friendship between them.

The `social-health` network is a network created from a survey of students in 1994/1995 [89]. Each student was asked to list their five best female and five best male friends. A node represents a student and an edge between two students shows at least one chose the other as a friend.

Communication networks

Communication networks describe the individual messages exchanged between people. Communication networks are often directed and typically contain multiple edges each with distinct time stamps so we are neglecting a lot of information when working with simple graphs representations.

The `commun-email` network is based on emails sent between members of the University Rovira i Virgili in Tarragona in the south of Catalonia in Spain [90]. Nodes are users and each edge represents that at least one email was sent between two users.

The `commun-DNC-email` network is built from the emails from the Democratic National Committee, the formal governing body for the United States Democratic Party. A dump of emails of the Democratic National Committee was leaked in 2016. Nodes in the network correspond to persons in the dataset. An edge in the dataset denotes that at least one email has been sent between the two linked nodes.

The `commun-UC-message` network represents messages sent between the users of an online community of students from the University of California, Irvine [80]. An edge connects two users if they exchanged at least one message.

The `commun-EU(core)-email` is a network representing emails sent between two members of a large European research institution [84]. An edge represents an email sent between members of the institution. This data was downloaded from Stanford Large Network Dataset Collection [82].

The `commun-DIGG-reply` data [91] gives a network of users of the social news website Digg. Each node is a user of the site and two users are connected by an edge if one of those users replied to another user at any point.

Citation networks

Citation networks represent documents as nodes in the network, with two nodes linked if one document cites another. These are direct acyclic graphs in principle but here we use a simple graph representation.

The `citation-DBLP-cite` is the citation network built from the DBLP database of computer science publications [92].

The `citation-Cora` network uses another database of computer science papers, CORA [93, 94]. Our simple network is constructed as for the DBLP network.

Co-authorship network

Co-authorship networks are networks connecting authors who have written articles together. Co-authorship networks are normally weighted but we ignore that here.

The `coauthor-astro-ph` network is the co-authorship network from the astrophysics section (`astro-ph`) of arXiv preprint archive constructed in [95]. Nodes are authors and an edge denotes a collaboration on at least one paper.

The `coauthor-netscience` network is a network of co-authors in the area of network science [96]. Nodes represent authors and edges denote collaborations.

The `coauthor-pajek` is the co-authorship graph around Paul Erdős [97] which is used to define the "Erdős number".

Hyperlink networks

In hyperlink networks the nodes are pages or documents. These are linked by an edge if there is at least one hyperlink between these two documents in either direction as here we ignore the direction inherent to hyperlinks.

We use two examples from hyperlinks between blogs about politics during the U.S. Presidential Election of 2004 [98], `hyperlinks-blogs` and `hyperlink-polblog`.

4.4 Numerical results

4.4.1 Theoretical models

We looks at the relationship between closeness and degree using simple networks produced from three different theoretical models: the Erdős-Réyni (ER) model [195] (also see section 12.2 [194]), the Barabási-Albert model with preferential attachment [70] (also see section 14.3 [194]), and the configuration model [135] (also see section 13.2 [194]) network starting from a network generated with the same Barabási-Albert model. In the ER networks and the configurational BA networks, the edges are completely randomised so there are no vertex-vertex correlations. The last two models both have fat-tailed degree distributions. Results are shown in Figure 4.3 and Table 4.2. Our networks built from artificial models were created using standard methods in the `networkx` package [54] which is open source.

For single network, we get several nodes with the same degree and we use this variation to find a mean and standard error in the mean shown. The fit is done using Eq.(4.12) with two free parameters $\bar{z}^{(\mathrm{fit})}$ and $\beta^{(\mathrm{fit})}$ and the goodness of fit measures in Table 4.2 show this is a good fit. Roughly speaking we find that mean inverse closeness values for any one degree are typically within 2% of the prediction made from the best fit. The small deviations from our best fit for higher degree values are in a region where the data is sparse and uncertainties are large so no firm conclusions can be drawn about higher order-corrections to our form in Eq.(4.12).

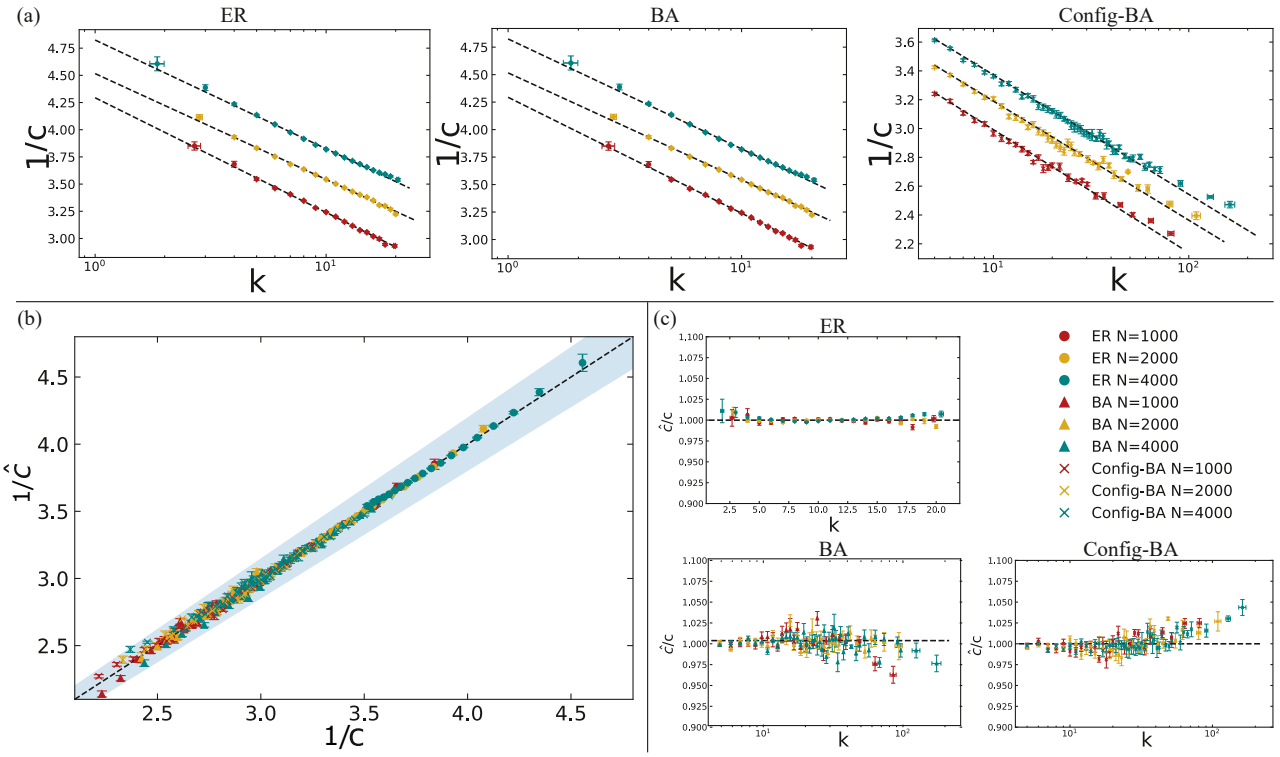


Figure 4.3: Numerical investigations of the conjecture relationships on artificial models. The results are for three different sized networks: $N = 1000$ (red points) $N = 2000$ (blue points) and $N = 4000$ (yellow points) where N is the number of nodes. All networks have average degree 10.0 and 100 realisations were taken for each case. The values of closeness for each value of degree are binned, the mean is shown as the data point with error bars the standard error of the mean. On the top row each plot shows results for networks formed from one artificial model: the Erdős-Réyni (ER) model on the left, the Barabási-Albert (BA) model in the centre, and the configuration model network starting from a Barabási-Albert model (Config-BA) on the right. The dashed lines shows the best linear fit of $1/c$ to $\ln k$ using Eq.(4.12). The same data from all nine artificial networks is shown in the scatter bottom left with data $1/\hat{c}$ against predicted value c obtained from the best best fit Eq.(4.12) and the shaded region corresponds to a 5% deviation from the theoretical prediction. On the bottom right, the fractional error, the fitted value of closeness divided by data value, is shown. The results show that the non-linear correlation of closeness and degree predicted in Eq.(4.12) works most of the time within a 2% variation. There are some hints of small deviations but systematic at higher degree value but the data is sparse and less reliable here.

We now turn to look at the actual values obtained from these fits of data on closeness and degree from the artificial networks to Eq.(4.12). As Table 4.2 shows there is a small amount of variation in value of $\bar{z}^{(\mathrm{fit})}$, the fit for the shortest-path tree growth factor, with the size of the network. What is of more interest are the differences in values between these three types of artificial networks. All these networks had an average degree of about 10.0 and an infinite tree with constant degree 10 (a Bethe lattice) would have a growth factor $\bar{z} = 9$, one less than

Network type	N	$1/\ln \bar{z}^{(\mathrm{fit})}$	$\beta^{(\mathrm{fit})}$	$\bar{z}^{(\mathrm{fit})}$	$\beta(\bar{z}^{(\mathrm{fit})}, N)$	$\rho(c,k)$	χ^2_{r}
ER	1000	0.46 ± 0.01	4.29 ± 0.01	8.87 ± 0.20	3.98 ± 0.03	0.94	1.02
	2000	0.42 ± 0.01	4.52 ± 0.01	10.64 ± 0.18	4.07 ± 0.03	0.93	1.02
	4000	0.43 ± 0.01	4.82 ± 0.01	9.99 ± 0.12	4.45 ± 0.02	0.93	1.03
BA	1000	0.30 ± 0.01	3.59 ± 0.02	28.03 ± 3.11	3.02 ± 0.13	0.75	1.16
	2000	0.32 ± 0.01	3.86 ± 0.01	22.76 ± 2.22	3.37 ± 0.07	0.70	1.29
	4000	0.31 ± 0.01	4.03 ± 0.01	25.17 ± 2.61	3.52 ± 0.08	0.65	1.16
Config-BA	1000	0.35 ± 0.01	3.76 ± 0.02	17.41 ± 1.42	3.34 ± 0.14	0.75	1.19
	2000	0.36 ± 0.01	4.01 ± 0.02	16.08 ± 1.24	3.65 ± 0.15	0.70	1.28
	4000	0.35 ± 0.01	4.19 ± 0.01	17.41 ± 1.41	3.82 ± 0.08	0.66	1.19

Table 4.2: Results for one example of a simple graph with average degree 10.0 produced using one of three artificial models with the same average degree $\langle k \rangle = 10.0$ but with a different number of nodes, N. Each 'ER' network is a standard Erdős-Réyni network, a 'BA' network is produced using pure preferential attachment in the Barabási-Albert model, and the 'Config-BA' network is a configuration model version of a Barabási-Albert model network. The results for $1/\ln \bar{z}^{(\mathrm{fit})}$ and $\beta^{(\mathrm{fit})}$ come from linear fits of inverse closeness, $1/c_v$ to the logarithm of degree, $\ln k_v$ for each vertex v, i.e. $1/c_v = (\ln \bar{z})^{-1} \ln k_v + \beta$ in Eq.(4.12). The value of β derived from $\bar{z}^{(\mathrm{fit})}$ and N using Eq.(4.13) is also shown for comparison. The fits are very good as indicated by the column for the reduced chi-square χ^2_{r}.

the average degree. So the best fit values for the growth factor $\bar{z}$ in the Erdős-Réyni networks are a little higher than this while the Barabási-Albert network and its randomised version are a lot bigger.

Another possible reference value for the shortest-path tree growth factor $\bar{z}$ is the average degree of a neighbour in a random graph with the same degree distribution which is $\langle k \rangle_{\mathrm{nn}} = \langle k^2 \rangle / \langle k \rangle$. This is the relevant value for diffusive processes on a random graph. For our finite Erdős-Réyni networks we have that $\langle k \rangle_{\mathrm{nn}} \approx \langle k \rangle$ so again the growth factor found to give the best fit, $\bar{z}^{(\mathrm{fit})}$, in actual Erdős-Réyni networks is still a bit higher than this estimate. For the Barabási-Albert networks and their randomised versions, the $\langle k \rangle_{\mathrm{nn}}$ is around twenty-two to twenty-five for the networks in Figure 4.2. This value is much closer but still not in complete agreement. This suggests our shortest-path trees are sampling nodes in a different way from diffusion but still with a bias to higher degree nodes.

Since spanning trees have many fewer edges than the original graphs, it is perhaps somewhat surprising that we find that the growth factors comparable with any measures of the average degree in the original network. So the high values of $\bar{z}$ are telling us that the shortest-path trees are sampling the nodes of their networks with a large bias towards high degree nodes in

the parts of the tree close to the root node and that is why we need such a high growth rate $\bar{z}^{(\text{fit})}$ when we fit our data for closeness. That way when we prune the edges to produce a tree we will still have high degrees in the tree close to the root node. The corollary is that the outer parts of shortest-path trees are dominated by leaves (degree one nodes) and other low degree nodes, and these also correspond to low degree nodes in the original network. What we see is consistent with the pictures used to understand the small world nature of these models, where the high degree nodes play a key role in acting as hubs for the shortest paths in the network, for example see the discussion by Bollobás [196].

It is also clear that node correlations play an important role as these are present in the Barabási-Albert model but absent in the randomised version. The large difference in $\bar{z}$ values for these two cases show such node correlations are important and yet, the non-linear relationship Eq.(4.13) still holds remarkably well in these artificial networks, with or without these correlations.

The β parameter in Eq.(4.12) is harder to interpret but Table 4.2 shows a comparison between the two values of β. The first is $\beta^{(\text{fit})}$ derived from a two-parameter fit of the data to Eq.(4.12). The second value is $\beta(\bar{z}^{(\text{fit})}, N)$ the value predicted using Eq.(4.13) where we use the $\bar{z}$ value obtained from the same two parameter fit and the number of nodes N. What we can see is that the values derived using Eq.(4.13), $\beta(\bar{z}^{(\text{fit})}, N)$, are consistently poorer than the values $\beta^{(\text{fit})}$ derived from a two-parameter fit. It highlights that the details of our theoretical form, such as the precise formula for β, here Eq.(4.13), can be improved. However, our simple calculation has captured the important features of the problem so that the form Eq.(4.12) does work in these theoretical models provided we treat both $\bar{z}$ and β in Eq.(4.12) as free parameters to be determined.

4.4.2 Real-world data

We aim for a wide range of networks both in terms of size and in terms of the type of interaction encoded in these real-world networks. We have examined eighteen data sets based on real-world data from five broad categories: social networks (`social-...`), communication

networks (`commun-...`), citation networks (`citation-...`), co-author networks (`coauth-...`), and hyperlink networks (`hyperlink-...`).

Summary statistics are given in Table 4.3. The reduced chi-square $\chi^2_{\rm r}$ measure is between 1.05 and 1.61 for ten networks, more than half, of our examples and another four networks have values between 2.09 and 2.86. Given the wide range of both size and nature of these networks and the simplicity of our theoretical derivation, this agreement is remarkable. We also give the Pearson correlation measure between closeness and degree, $\rho(c,k)$, and this is generally high as has been noted before [17, 18, 19, 20, 21, 22, 23, 24, 25, 26]. The success of our non-linear relationship between closeness and degree is not incompatible with high $\rho(c,k)$ values.

Network	N	$1/\ln(\bar{z}^{(\rm fit)})$	$\beta^{(\rm fit)}$	$\bar{z}^{(\rm fit)}$	$\beta(\bar{z}^{(\rm fit)},N)$	$\rho(c,k)$	$\chi^2_{\rm r}$
`social-karate-club`	34	0.460 ± 0.066	2.997 ± 0.095	8.81 ± 2.76	2.44 ± 0.25	0.77	1.09
`social-jazz`	198	0.367 ± 0.015	3.349 ± 0.048	15.28 ± 1.72	2.84 ± 0.08	0.86	12.16
`social-hamster`	1788	0.353 ± 0.009	4.129 ± 0.020	17.05 ± 1.22	3.56 ± 0.07	0.68	1.11
`social-oz`	217	0.403 ± 0.010	3.492 ± 0.038	11.96 ± 1.02	3.04 ± 0.08	0.89	2.86
`social-highschool`	70	0.561 ± 0.039	3.734 ± 0.079	5.95 ± 0.74	3.08 ± 0.17	0.87	1.17
`social-health`	2539	0.537 ± 0.008	5.605 ± 0.016	6.43 ± 0.17	4.94 ± 0.06	0.75	1.05
`commun-email`	1133	0.394 ± 0.007	4.309 ± 0.014	12.64 ± 0.54	3.65 ± 0.05	0.84	1.06
`commun-UC-message`	1893	0.264 ± 0.003	3.526 ± 0.008	43.92 ± 2.16	2.96 ± 0.03	0.72	2.23
`commun-EU(core)-email`	986	0.259 ± 0.004	3.324 ± 0.012	47.63 ± 2.67	2.76 ± 0.03	0.84	29.21
`commun-DNC-email`	1833	0.222 ± 0.010	3.499 ± 0.012	91.16 ± 18.84	2.65 ± 0.08	0.41	1.34
`commun-DIGG-reply`	29652	0.388 ± 0.002	5.078 ± 0.003	13.12 ± 0.18	4.89 ± 0.02	0.61	1.61
`citation-DBLP`	12494	0.361 ± 0.003	4.856 ± 0.004	15.98 ± 0.35	4.31 ± 0.03	0.54	1.31
`citation-Cora`	23166	0.503 ± 0.004	6.639 ± 0.008	7.31 ± 0.13	5.82 ± 0.04	0.48	1.15
`coauthor-astro-ph`	14845	0.441 ± 0.004	5.735 ± 0.010	9.67 ± 0.21	5.07 ± 0.04	0.61	14.30
`coauthor-netsci`	379	0.382 ± 0.080	6.553 ± 0.119	13.74 ± 7.51	3.16 ± 0.48	0.35	1.47
`coauthor-pajek`	6927	0.259 ± 0.002	3.894 ± 0.002	47.45 ± 1.33	3.26 ± 0.02	0.64	12.79
`hyperlink-polblog`	1222	0.240 ± 0.004	3.316 ± 0.011	64.84 ± 4.44	2.68 ± 0.03	0.72	2.12
`hyperlink-blogs`	1222	0.239 ± 0.004	3.316 ± 0.011	65.07 ± 4.48	2.68 ± 0.03	0.72	2.09

Table 4.3: Results for a variety of friendship networks derived from real-world data. The results for $1/\ln\bar{z}^{(\rm fit)}$ and $\beta^{(\rm fit)}$ come from linear fits of inverse closeness, $1/c_v$ to the logarithm of degree, $\ln k_v$ for each vertex v, i.e. $1/c_v = (\ln\bar{z})^{-1}\ln k_v + \beta$ in Eq.(4.12). The value of β derived from $\bar{z}^{(\rm fit)}$ and N using Eq.(4.13) is also shown as $\beta(\bar{z}^{(\rm fit)},N)$ for comparison. The fits are very good as reduced chi-square $\chi^2_{\rm r}$ values show.

The data for each network is shown in more detail in Figure 4.4. Again, we can see that within the error bars the average closeness at each degree generally follows the form we predict within 5% when the best fit parameters are used. Further, the uncertainties estimated for these data points suggest that the vast majority of average closeness values are statistically consistent with the predicted value for that degree, something already captured by the reduced chi-square values in Table 4.3.

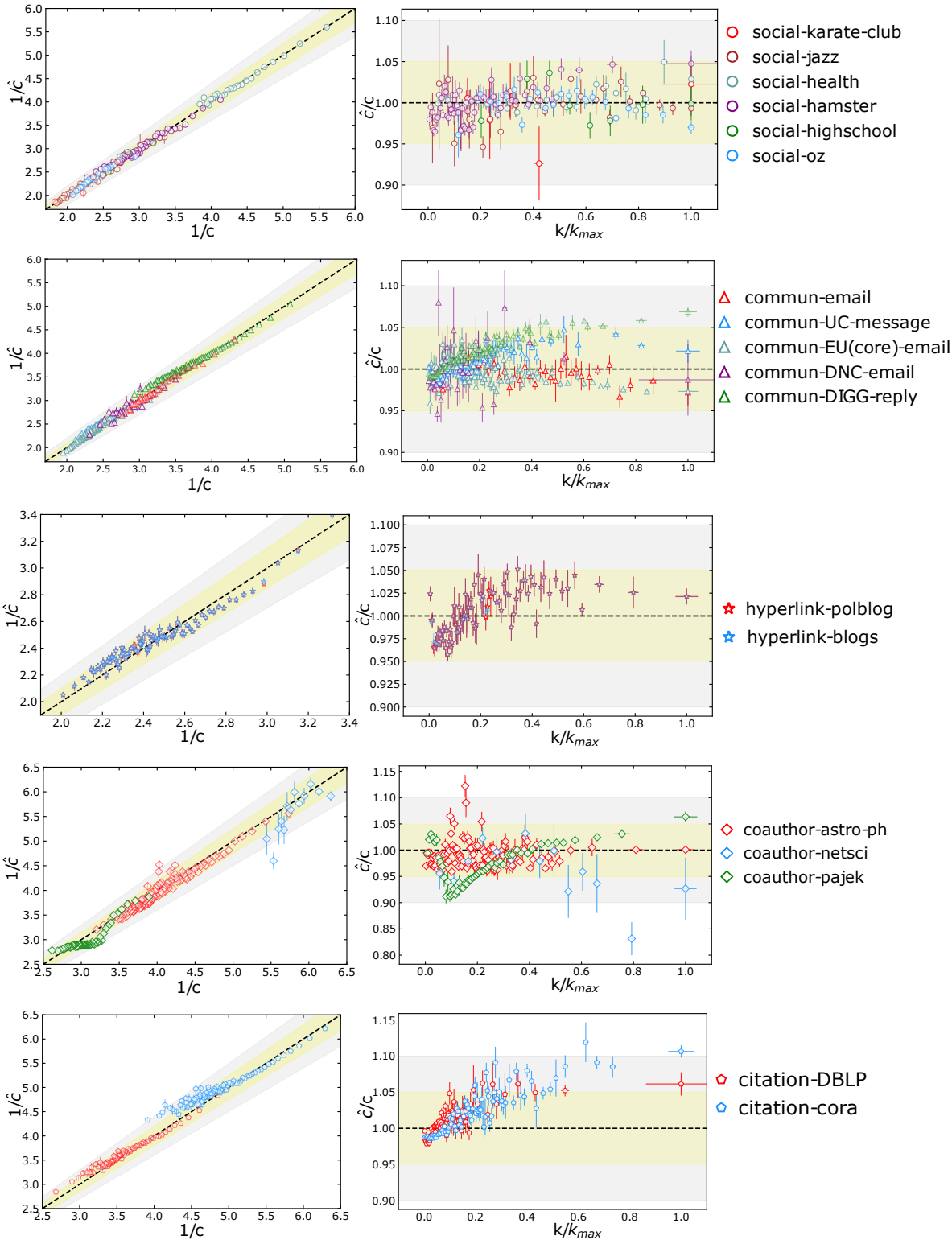


Figure 4.4: Results for eighteen real networks derived from real-world data, see Table 4.3 for the statistics of each dataset. The yellow shaded region corresponds to 5% deviation and grey region corresponds to a 10% deviation. The left plot shows the the inverse of the predicted result $1/\hat{c}$ from the best fit against the inverse of the mean measured value $1/c = 1/\langle c \rangle_k$ averaged over nodes with the same degree k. If the prediction matched data perfectly, the point will lie on the dashed line. Both axes are essentially $\ln k$. The error bars represent from standard error of mean of the inverse closeness. For majority of points, we can see our prediction Eq.(4.12) captures the relation between closeness and degree, usually with within a 5% margin.

4.5 Discussion and conclusions

Our results confirm our conjecture that the inverse of closeness depends linearly on the degree Eq.(4.12) for most networks. This is a correlation, true on average but not an exact for every node. This non-linear relationship has been missed in previous studies which focused on linear correlations. Our work suggests that in the majority of networks, closeness captures little more information on average than is contained in the degree.

An important use of our result is that it allows us to factor out this trivial dependence to extract the information contained in closeness that is independent of degree. For instance, we could start by examining the degree centrality of every node. This would be the primary measure of centrality. Then we could fit our closeness values using Eq.(4.12) to produce an expected value of closeness $c_v^{(\mathrm{fit})}$ for each node and we would use this to find nodes which are more or less central than expected, for instance using the normalised closeness

$$c_v^{(\mathrm{norm})} = \frac{c_v}{c_v^{(\mathrm{fit})}} \,. \tag{4.19}$$

This would highlight the outliers which would then indicate any interesting behaviours.

The success of our conjecture also suggests that most networks satisfy the two key assumptions built into our derivation. First we assumes that the number of nodes a distance ℓ from any node grows exponentially and this must be reasonable for most networks. That exponential growth is common is not a surprise as it is essentially the mechanism behind the concepts of the "six degrees of separation" and the "small world" effect. More formally, length scales in most real-world networks grow as $O(\ln N)$ and so much slower than networks embedded in d-dimensional Euclidean space where $n(\ell) \sim \ell^{1/d}$. For instance, when we averaged the inverse closeness Eq.(4.7b) over all vertices we found our prediction for the average path length in (4.18), where the N dependence comes from the $\ln N$ term in the expression for β of Eq.(4.13).

Our work shows that our Euclidean intuition regarding closeness breaks down for most networks with their small world, non-Euclidean features. If we look at the original context where closeness was developed, Bavelas [53] only uses planar graph examples and the initial applications of

closeness centrality measures were on very small networks. It appears the importance of the small-world vs Euclidean properties of a network when interpreting closeness was lost when, much later, closeness was used to analyse networks which were much larger and no longer constrained by geography.

Our second assumption that our results support is that the branches of the shortest-path trees are statistically similar, as illustrated in Figure 4.1. The success of our results suggests this assumption works well whenever we are looking at measurements that depend on the bulk of the network, rather than one special path (e.g. betweenness) or the immediate neighbourhood (e.g. community detection). This simple approximation may therefore help analyse other network measurements.

Though generally very successful, there are examples of networks where our form Eq.(4.12) fails to capture the behaviour well or where we can see some clear, if small, trends in the deviations. These cases highlight the limitations of our approach but also suggests how this approach may be improved. At the simplest level, we could replace the sharp cutoff used for $n_\ell(r)$ where $n_\ell(r) = 0$ for $\ell > L$. That may well lead to better predictions for β as we used fitted values rather than our prediction Eq.(4.13) but fitting one rather than two parameters while theoretically satisfying does not seem a gain in practice. More serious changes will be needed to the calculation if other effects neglected here, such as community structure or degree assortivity, are to be included.

However, another option might be to calculate a different network parameter, namely the second degree $k_r^{(2)} = n_{\ell=2}(r)$ [223] for each node r. By finding the number of nodes two steps away we can make a better approximation for $n_\ell(r)$, that is $n_0(r) = 1$, $n_1(r) = k$, and $n_\ell(r) = k_r^{(2)}\bar{z}^{\ell-2}$ for $2 \le \ell \le L_r$ and $n_\ell(r) = 0$ for $\ell > L_r$. This approach cannot be worse than the method used here as the latter is included as a special case where the second degree $k_r^{(2)} = \bar{z}k_r$ for all nodes r. To leading order we get the same type of result, namely that $1/c_r = (\bar{z})^{-1} \ln k_r^{(2)} + \beta$ since the degree k_r now only contributes a small number of terms to closeness. So in this approach using second degree we need a different set of N parameters to find. Finding second degree is slower than degree but both scale in the same way with increasing network size. The success of our

simpler method here points to the idea that second degree and degree may often be correlated so using second degree may only enhance results in a few cases.

Finally, it is interesting to note that the logarithm of degree $\ln k$ has been found to play an important role in network analysis before. A large fraction of papers on the topic will show degree distributions where the horizontal axis is the dependent variable $\ln k$ and not simply the degree k. A more specific example comes from [224] where the ratio of the degrees of nodes at the two ends of each edge (largest value in the numerator) is used to assign a 'distance' $\eta(u, v)$ to each edge (u, v). This is equivalent to defining $\lambda(u, v) = \ln \eta(u, v) = |\ln k_u - \ln k_v|$. In fact one can quickly see that while both η and λ are semi-distances on the set of edges in the formal mathematical sense, only λ is also a semi-metric and so λ is in some sense the more natural 'distance' measure in a qualitative sense. Our work suggests that an alternative view is to replace the logarithm of degree by the inverse of closeness. Since $(c_u)^{-1}$ is the actual average of the shortest-path distances from u to all other nodes, we can immediately see it is natural to work with inverse closeness while considering distances. For instance in [224] we could look at a different edge measure $\tilde{\lambda}(u, v) = |(c_u)^{-1} - (c_v)^{-1}|$. While the inverse closeness is a more natural distance, the degree is much easier to calculate in practice. Our work allows researchers to move between these two pictures.

4.6 Summary

In this chapter, we

- investigated correlations between closeness and degree.
- gave an analytic derivation that shows the inverse of closeness is linearly dependent on the logarithm of degree, based on the shortest-path tree approximation.
- showed that our hypothesis works well for a range of networks produced from stochastic network models including the Erdős-Réyni and Barabási-Albert models.

- tested our relation on networks derived from a wide range of real-world data including social networks, communication networks, citation networks, co-author networks, and hyperlink networks.
- found the relationship holds true within a few percent in most, but not all, cases.
- connected the results with those on average distance and suggested some ways that this relationship can be used to enhance network analysis.

Chapter 5

Conclusion and future work

Complexity science is an emerging field[2, 1] with wide applications in many fields, but there is no universal mathematical framework to solve many questions. For each question, complexity science provides a way to understand emergent behaviours, for instance, from earthquake, rainfall, rice piles, atrial fibrillation [225, 226, 227, 228]. In addition, complexity science also provides an alternative way to understand other fields, including ecology and epidemiology [229, 230, 231]. However, challenges still exist since other fields do not widely accept the ideas or comments from the models from complexity science, though models are mathematically precise and rigorous.

This thesis uses complex networks to study three different questions. Two of these topics investigate the structures of networks and one uses networks to model the flows of bicycles using the gravity model. Firstly, we looked through popular higher-order network models in recent years. To be specific, we designed a combinatorial model, which used three-node interactions to analyse how temporal networks evolve. We also designed a statistical method that can verify the significance of the triplet transitions in temporal networks. We analysed several social and communication networks and found it is evident that the triplet transition mechanism appears in temporal networks using our statistical test. We use our triplet transition methods to predict evolution of edges and triplet transition methods in general outperform other nine pair-wise methods in the goal of predicting links. The result implies the evolution of temporal network

is indeed beyond pairwise in many real world networks. It is not simple to detect higher-order mechanisms directly behind observed topology. In fact, what we observed from the triplets may be projections from higher dimension graphlets, for instance, four nodes interactions. Understanding optimal numbers of multi-way interaction mechanism is an important but less well-developed question, which needs to be further investigated. In addition, temporal network reconstruction is an emerging field and lacks models that can explain the evolution of temporal networks. A further question is the network completion problem, where the nodes in the network are not all observable. The state information can be used to infer the structures of networks, one of these tasks is to construct explainable likelihood inference, instead of relying on methods such as neural networks or embedding methods. Network completion problems have many applications in different fields, for instance, in biological experiments, determining protein-protein interactions usually cost a lot and only about 70% interactions can be obtained. Inferring the hidden structures can help to suggest if critical nodes are determined, so any further experiment is not necessary. These fields are still relatively new and need to be better understood.

Human mobility is an important question to understand, especially during a pandemic. Understanding mobility patterns can help cities to allocate resources. We looked at the flow of bicycles in Shanghai and Beijing. We found the gravity model can capture human mobility behaviour. Gravity model or other recently developed mobility models [176, 180, 175], reflect the interactions between spatially distributed populations, in which spatial attraction and spatial constraints play an important role [182]. The gravity model originally was inspired by Newton's Law of Gravity. In Newton's law of Gravity, the strength of attraction force between two objects is proportional to the product of two masses and inversely proportional to distance squared between them. In the Gravity model, it is stated that the flows between two locations are proportional to the product of populations in two locations, inversely proportional to distances between two locations. Though each individual human being mobility behaviour is highly complex, at a large scale the model can predict correct flow behaviours. However, if we look at locations from different spatial scales, i.e. using 500 m × 500 m, instead of 1000 m × 1000 m, we found the Gravity models still hold, however, the scaling exponents are no longer the

same. Since there exists transitions across two very near locations, if we coarse-grain these two locations into one, these transitions across will be neglected and treated as self-loops. Therefore, the system is no longer scale-invariant due to information loss if we choose a larger spatial resolution. In addition, spatial models are limited since the transitions between locations in a city have time dependence, especially at intra-city level, a more precise spatio-temporal mapping of population would be essential to mobility models. When the mobility is not limited to commuting (which is the case for biking traffic), we assume that the active population, instead of a pure residential or a working population would be a more suitable proxy for the attraction of locations. The active population measures within a certain region how many people have ever been there in one day and this can be more accurately estimated from cellphone data [167, 182, 187]. Yet due to limitations on getting access to large scale cellphone data, we still use the residential population as proxy of attraction of locations. From such an interaction perspective, the relationship between the gravity model and other recently developed mobility models (including the radiation model [176], population-weighted opportunities model [180]) might be better revealed [188]. Furthermore, comparing the effectiveness of the aforementioned mobility models on biking behaviors and developing a more suitable mechanistic model would be important future works.

Centrality has been applied widely in many different fields, from biological science, medical science to social networks, for instance, it has been found that the road with high betweenness is normally regarded as an optimal route, thus nodes with high betweenness should expect to receive more traffics[232, 233]. High correlations between different centrality measures are really common, but the reason behind it is rarely studied. We looked at closeness, which measures the average distance of a node to all reachable nodes in the network and degree, which measures the number of neighbours. The intuition behind the strong correlation is that if a node has a high degree, then there are immediately more shortest paths to access rest of other nodes compared to low degree nodes. One natural structure which captures this concept is a tree structure. The tree is a structure that uses minimum numbers of edges to maintain connections in the network. To answer the question why the global distance measure is related to local connectivity, we choose the shortest-path tree as our starting point to approximate the network. We used a

simple homogeneous assumption, that is, on average, each node in the network, apart from the root node and leaves, connects to $\bar{z}$ new nodes that have not connected to original nodes before. The parameter $\bar{z}$ is also known as the effective branching number which explains how quick a network grows. We derived an analytical equation that relate the logarithm of degree and inverse of the closeness and we tested the relationship on several different types networks, including communication networks, social networks, coauthor networks, hyperlink networks and citation networks. For most of our real networks, our relationship performs well. To understand intuitions behind the effective branching number, we relate the average distance to an effective branching number and find an analytical expression that describes how the length scales grows as the size of networks grows. The small world model states that the average shortest path length of real networks grows as $O(\ln N)$. Our model provides a more precise description, that is, the average shortest path length is $\langle l \rangle = \ln N / \ln \bar{z}$, which is different from the average degree $\langle k \rangle - 1$ in the network, since it only consider average number of new nodes a node brings to. We test our models in many networks and most agreed with our theory. The results bring some new insights, that a higher density of connections in the networks does not imply that shorter average path lengths, but the effective branching number do. Our analysis can enhance network analysis. If we find there exist a good fit between inverse of closeness and logarithms of degree($\ln k$). Further extension can be made to extend the study effective branching number $\bar{z}$ in different networks, for instances, the effectiveness of edges should be larger in social network compare to communication networks. This can help to detect whether there exist abnormal structures behind observed topology in networks.

Bibliography

[1] Jensen, H. J. . Self-organized criticality: emergent complex behavior in physical and biological systems. *Cambridge university press*, Vol. 10 (1998).

[2] Christensen, K., & Moloney, N. R. Complexity and criticality, *World Scientific Publishing Company*, Vol. 1 (2005).

[3] Newman, M. *Networks*. Oxford university press (2018).

[4] Gerlach, M., Peixoto, T. P., & Altmann, E. G. A network approach to topic models. *Science Advances*, **4**(7), eaaq1360 (2018).

[5] Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen, A.M., Radicchi, F., Sinatra, R., Uzzi, B. & Vespignani, A., Science of science. *Science*, **359**(6379) (2018).

[6] Cheng, F., Kovács, I. A., & Barabási, A. L. Network-based prediction of drug combinations. *Nature communications*, **10**(1), 1-11 (2019).

[7] Kovács, I. A. *et.al.* Network-based prediction of protein interactions. *Nature communications*, **10**(1), 1-8 (2019).

[8] Gysi, D.M., Do Valle, Í., Zitnik, M., Ameli, A., Gan, X., Varol, O., Ghiassian, S.D., Patten, J.J., Davey, R.A., Loscalzo, J. & Barabási, A.L. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proceedings of the National Academy of Sciences*, **118**(19) (2021).

[9] Deville, P., Song, C., Eagle, N., Blondel, V. D., Barabási, A. L., & Wang, D. Scaling identity connects human mobility and social interactions. *Proceedings of the National Academy of Sciences*, **113**(26), 7047-7052 (2016).

[10] Li, R., Dong, L., Zhang, J., Wang, X., Wang, W. X., Di, Z., & Stanley, H. E. Simple spatial scaling rules behind complex cities. *Nature Communications*, 8(1), 1-7(2017).

[11] Scarpino, S. V., Allard, A., & Hébert-Dufresne, L. The effect of a prudent adaptive behaviour on disease transmission. *Nature Physics*, **12**(11), 1042-1046 (2016).

[12] Sahasrabuddhe, R., Neuhäuser, L., & Lambiotte, R. Modelling non-linear consensus dynamics on hypergraphs. *Journal of Physics: Complexity*, **2**(2), 025006 (2021).

[13] Wasserman, S., & Faust, K. Social network analysis: Methods and applications (1994).

[14] Page, L., Brin, S., Motwani, R., & Winograd, T. The PageRank citation ranking: Bringing order to the web. *Stanford InfoLab* (1999).

[15] Schoch, D. *A Positional Approach for Network Centrality.* Ph.D. thesis, Universität Konstanz (2015).

[16] Schoch, D. Periodic table of network centrality (2016). URL `http://schochastics.net/sna/periodic.html`.

[17] Bolland, J. M. Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks. *Social networks* **10**, 233–253 (1988).

[18] Rothenberg, R. B. *et al.* Choosing a centrality measure: epidemiologic correlates in the colorado springs study of social networks. *Social Networks* **17**, 273–297 (1995).

[19] Faust, K. Centrality in affiliation networks. *Social networks* **19**, 157–191 (1997).

[20] Lee, C.-Y. Correlations among centrality measures in complex networks. *arXiv preprint physics/0605220* (2006). `arXiv:physics/0605220`.

[21] Valente, T. W., Coronges, K., Lakon, C. & Costenbader, E. How correlated are network centrality measures? *Connections (Toronto, Ont.)* **28**, 16 (2008). URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2875682/`.

[22] Batool, K. & Niazi, M. A. Towards a methodology for validation of centrality measures in complex networks. *PloS one* **9**, e90283 (2014).

[23] Lozares, C., López-Roldán, P., Bolibar, M. & Muntanyola, D. The structure of global centrality measures. *International Journal of Social Research Methodology* **18**, 209–226 (2015).

[24] Schoch, D., Valente, T. W. & Brandes, U. Correlations among centrality indices and a class of uniquely ranked graphs. *Social Networks* **50**, 46–54 (2017).

[25] Oldham, S. *et al.* Consistency and differences between centrality measures across distinct classes of networks. *PLOS ONE* **14**, e0220061 (2019).

[26] Bringmann, L. F. *et al.* What do centrality measures measure in psychological networks? *Journal of Abnormal Psychology* **128**, 892–903 (2019).

[27] Arnaudon, A., Peach, R. L. & Barahona, M. Scale-dependent measure of network centrality from diffusion dynamics. *Physical Review Research* **2**, 033104 (2020).

[28] Lambiotte, R., Rosvall, M. & Scholtes, I. From networks to optimal higher-order models of complex systems, *Nature Physics*, **15**, 313-320 (2019).

[29] Benson, A., David F., & Jure L. Higher-order organization of complex networks. *Science* **353**, 163-166 (2016).

[30] Rapoport, A Spread of information through a population with socio-structural bias. I. Assumption of transitivity, *Bulletin of Mathematical Biology*, **15**, 523-533 (1953).

[31] Granovetter, M. The Strength of Weak Ties. *American Journal of Sociology*, **78**, 1360 (1973).

[32] Expert, P. *et al.* Self-similar correlation function in brain resting-state functional magnetic resonance imaging. *J R Soc Interface* **8**, 472-479 (2011).

[33] Petri, G.*et al* Homological scaffolds of brain functional networks. *J R Soc Interface* **11**, 20140873 (2014).

[34] Expert, P., Lord, L, Kringelbach, Morten L, & Petri, G. Topological neuroscience, **3**, 653-655, (2019).

[35] Sanchez-Gorostiaga, A., Bajić, D., Osborne, M. L., Poyatos, J. F., & Sanchez, A. High-order interactions dominate the functional landscape of microbial consortia. *PLoS Biol 17* e3000550 (2019).

[36] Bairey, E., Eric D., & Roy K. High-order species interactions shape ecosystem diversity. *Nat. Communication* **7**(1), 1-7 (2016).

[37] Milo, R. *et al.* Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).

[38] Milenkovic, T., & Przulj, N. Uncovering biological network function via graphlet degree signatures. *Cancer informatics*, 6, CIN-S680 (2008).

[39] Isella, L.,*et al.* What's in a crowd? Analysis of face-to-face behavioral networks. *Journal of Theoretical Biology* **271**, 166–180 (2011).

[40] Yao, Q., Evans,T.S.,& Christensen,K. How the network properties of shareholders vary with investor type and country. *PloS One* **14** e0220965; 10.1371/journal.pone.0220965 (2019).

[41] Li, A., Zhou, L., Su, Q., Cornelius, S. P., Liu, Y. Y., Wang, L., & Levin, S. A. Evolution of cooperation on temporal networks. *Nature communications*, 11(1), 1-9 (2020)

[42] Holme, P. and Saramäki, J. Temporal Networks. *Physics Reports* **519**, 97–125 (2012).

[43] Iniguez, G., Battiston, F., & Karsai, M. Bridging the gap between graphs and networks. *Communications Physics*, 3(1), 1-5 (2020).

[44] Lovász, L.. Large networks and graph limits . *American Mathematical Society*,Vol. 60 (2012).

[45] Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C.I., Gómez-Gardenes, J., Romance, M., Sendina-Nadal, I., Wang, Z. & Zanin, M. The structure and dynamics of multilayer networks. *Physics reports*, **544**(1), pp.1-122 (2014).

[46] Masuda, N., & Lambiotte, R. A guide to temporal networks. (2016).

[47] Latora, V., Nicosia, V., & Russo, G. Complex networks: principles, methods and applications. Cambridge University Press (2017).

[48] Barabási, A.L. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **371**.1987 (2013).

[49] Anderson, P. W. More is different. *Science* **177**, (4047), 393-396 (1972).

[50] Battiston, F. *et al.* Networks beyond pairwise interactions: structure and dynamics. *Physics Reports* (2020).

[51] Skiena, S. S. The algorithm design manual (Vol. 2). New York: springer (1998).

[52] Van Loan, C. F., & Golub, G. Matrix computations (Johns Hopkins studies in mathematical sciences) (1996).

[53] Bavelas, A. Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America* **22**, 725–730 (1950).

[54] Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T. & Millman, J. (eds.) *Proceedings of the 7th Python in Science Conference (SciPy2008)*, 11–15, (2008).

[55] Freeman, L. C. A set of measures of centrality based on betweenness. *Sociometry*, 35-41 (1977).

[56] Brandes, U. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, **30**(2), 136-145 (2008).

[57] Lü, L., & Zhou, T. Link prediction in complex networks: A survey. *Physica A* **390**, 1150–1170 (2011).

[58] Newman, M. E. Detecting community structure in networks. *The European physical journal B*, **38**(2), 321-330, (2004).

[59] Abbe, E. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, **18**(1), 6446-6531 (2017).

[60] Plantié, M., & Crampes, M. Survey on social community detection. *In Social media retrieval* (pp. 65-85). Springer, London (2013).

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

[62] Holme, P., Huss, M., & Jeong, H. Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, **19**(4), 532-538 (2003).

[63] Girvan, M., & Newman, M. E. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, **99**(12), 7821-7826 (2002).

[64] Pons, P., & Latapy, M. Computing communities in large networks using random walks. *In International symposium on computer and information sciences* (pp. 284-293). Springer, Berlin, Heidelberg (2005).

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

[66] Yang, Z., Algesheimer, R., & Tessone, C. J. A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, **6**(1), 1-18 (2016).

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

[68] Rosvall, M., & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, **105**(4), 1118-1123 (2008).

[69] Lancichinetti, A., Kivelä, M., Saramäki, J., & Fortunato, S. Characterizing the community structure of complex networks. *PloS one*, **5**(8), e11976 (2010).

[70] Barabási, A. L., & Albert, R. Emergence of scaling in random networks. *Science*, **286**(5439), 509-512 (1999).

[71] Watts, D. J., & Strogatz, S. H. Collective dynamics of 'small-world'networks. *Nature* **393**(6684), 440-442 (1998).

[72] Newman, M. E. J. The structure and function of complex networks. *SIAM Review***45**, 167 (2003).

[73] Holland, P. W., Laskey, K. B., & Leinhardt, S. Stochastic blockmodels: First steps. *Social networks*, **5**(2), 109-137 (1983).

[74] Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., & Ghahramani, Z. Kronecker graphs: an approach to modeling networks. *Journal of Machine Learning Research*, **11**(2) (2010).

[75] Gilbert, E. N. Random graphs. The Annals of Mathematical Statistics, 30(4), 1141-1144(1959).

[76] Erdos, P., & Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, **5**(1), 17-60 (1960).

[77] Goh, K. I., Oh, E., Jeong, H., Kahng, B., & Kim, D. Classification of scale-free networks. *Proceedings of the National Academy of Sciences*, **99**(20), 12583-12588 (2002).

[78] Broido, A. D., & Clauset, A. Scale-free networks are rare. *Nature communications*, **10**(1), 1-10 (2019).

[79] Chen, B., Lin, Z., & Evans, T.S. Analysis of the Wikipedia Network of Mathematicians. Preprint at `https://arXiv.org/abs/1902.07622` (2019).

[80] Opsahl, T. & Panzarasa, P. Clustering in weighted networks. *Social networks* **31**, 155–163 (2009).

[81] Panzarasa, P., Opsahl, T., & Carley, K. M. Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *J. American Society for Information Science and Technology*, **60**, 911–93 (2009).

[82] Leskovec, J. & Krevl, A. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data` (2014).

[83] Kunegis, J. KONECT-The Koblenz Network Collection. *Proc. Int. Conf. on World Wide Web Companion*,1343–1350 (2013).

[84] Leskovec,J., Kleinberg, J., & Faloutsos, C. Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)* **1** (2007).

[85] Wayne Zachary. An information flow model for conflict and fission in small groups. *J. of Anthropol. Res.*, **33**:452–473 (1977).

[86] Coleman, J. S. *Introduction to Mathematical Sociology* (London Free Press Glencoe 1964).

[87] Gleiser, P. M. & Danon, L. Community Strucure in Jazz. *Advances in Complex Systems* **06**, 565–573 (2003).

[88] Freeman, L. C., Webster, C. M. & Kirke, D. M. Exploring social structure using dynamic three-dimensional color images. *Social Networks* **20**, 109–118 (1998).

[89] Moody, J. Peer influence groups: Identifying dense clusters in large networks. *Soc. Netw.* **23**, 261–283 (2001).

[90] Guimerà, R., Danon, L., Díaz-Guilera, A., Giralt, F. & Arenas, A. Self-similar community structure in a network of human interactions. *Physical Review E* **68**, 065103 (2003).

[91] Choudhury, M. D., Sundaram, H., John, A. & Seligmann, D. D. Social synchrony: Predicting mimicry of user actions in online social media. In *Proc. Int. Conf. on Comput. Science and Engineering*, 151–158 (2009).

[92] Ley, M. The DBLP computer science bibliography: Evolution, research issues, perspectives. In *Proc. Int. Symposium on String Process. and Inf. Retr.*, 1–10 (2002).

[93] McCallum, A. K., Nigam, K., Rennie, J. & Seymore, K. Automating the construction of internet portals with machine learning. *Information Retrieval* **3**, 127–163 (2000).

[94] Šubelj, L. & Bajec, M. Model of complex networks based on citation dynamics. In *Proc. of the WWW Workshop on Large Scale Network Analysis*, 527–530 (2013).

[95] Newman, M. E. J. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 404–409 (2001).

[96] Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **74** (2006).

[97] Batagelj, V. Pajek datasets. `http://vlado.fmf.uni-lj.si/pub/networks/data/` (2017).

[98] Adamic, L. A. & Glance, N. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, 36–43 (ACM, 2005).

[99] Battiston, S., Caldarelli, G., & D'Errico, M. The financial system as a nexus of interconnected networks. *In Interconnected networks* (pp. 195-229). Springer, Cham (2016).

[100] Grover, A., & Leskovec, J. node2vec:Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'16* 855-864 (2016).

[101] Scholtes, I. When is a Network a Network? Multi-Order Graphical Model Selection in Pathways and Temporal Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'17* 1037–1046 (2017).

[102] Freeman, L. C. The Sociological Concept of "Group": An Empirical Test of Two Models. *American Journal of Sociology*, **98**, 152 (1992).

[103] Evans, T.S., Clique Graphs and Overlapping Communities. *J.Stat.Mech* P12037; 10.1088/1742-5468/2010/12/P12037 (2010).

[104] Derényi, I, Palla, G, & Vicsek, T Clique Percolation in Random Networks. *Physical Review Letters* **94**, 160202 (2005).

[105] Muhammad, A., & Egerstedt, M. Control using higher order Laplacians in network topologies. In *Proc. of 17th International Symposium on Mathematical Theory of Networks and Systems*, pp. 1024-1038. (2006).

[106] Horak, D., Maletić, S & Rajković, M. Persistent homology of complex networks. *Journal of Statistical Mechanics: Theory and Experiment* (03), P03034 (2009).

[107] Giusti, C., Ghrist, R., & Bassett, D. S. Two's company, three (or more) is a simplex. *Journal of Computational Neuroscience* **41**, 1–14 (2016).

[108] Bianconi, G., & Rahmede, C. Emergent hyperbolic network geometry. *Scientific Reports* **6**, 1-9 (2017).

[109] Petri, G., & Barrat, A. Simplicial activity driven model. *Physical Review Letters* **121**, 228301 (2018).

[110] Millán, A. P., Torres, Joaqu. J. & Bianconi. G. Complex network geometry and frustrated synchronization. *Scientific Reports* **8**, 1–10 (2018).

[111] Millán, A. P., Torres, Joaqu. J. & Bianconi. G. Synchronization in network geometries with finite spectral dimension. *Physical Review E* **99**, 218301 (2019).

[112] Skardal, P. S. & Arenas, A. Abrupt desynchronization and extensive multistability in globally coupled oscillator simplexes. *Physical Review Letters* **122**, 248301 (2019).

[113] Millán, A. P., Torres, J. J. & Bianconi, G. Explosive higher-order Kuramoto dynamics on simplicial complexes. *Physical Review Letters* **124**, 022307 (2020).

[114] Iacopini, I., Petri, G., Barrat, A., & Latora, V. Simplicial models of social contagion. *Nature Communications* **10**, 1–9 (2019).

[115] Matamalas, J. T., Gómez, S. & Arenas, A. Abrupt phase transition of epidemic spreading in simplicial complexes. *Physical Review Research* **2**, 108701 (2020).

[116] Berge C., *Graphes et hypergraphes*, (Dunod, Paris, 1967).

[117] Berge C., *Graphs and Hypergraphs*, (North-Holland Publishing Co, 1973).

[118] Johnson, J., *Hypernetworks in the science of complex systems*, (World Scientific, 2013).

[119] Lentz, H. H. K., Selhorst, T. & Sokolov, I. M. Unfolding accessibility provides a macroscopic approach to temporal networks. *Phys. Rev. Lett.* **110**, 118701 (2013).

[120] Koher, A., Lentz, H. H. K., Hövel, P. & Sokolov, I. M. Infections on temporal networks — a matrix-based approach. *PLoS ONE* **11**, e0151209 (2016).

[121] Starnini, M., Baronchelli, A., & Pastor-Satorras, R. Modeling human dynamics of face-to-face interaction networks. *Physical Review Letters* **110**, 012049 (2013).

[122] Van Mieghem, P., & Van de Bovenkamp, R. Non-Markovian infection spread dramatically alters the susceptible-infected-susceptible epidemic threshold in networks. *Physical Review Letters* **110**, 169701 (2013).

[123] Scholtes, I., Wider, N., Pfitzner, R., Garas, A., Tessone, C. J., & Schweitzer, F. Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks *Nature Communications* **5**, 1–9 (2014).

[124] Delvenne, J.C., Lambiotte, R., & Rocha, L. E. Diffusion on networked systems is a question of time or structure. *Nature Communications* **6**, 1–10 (2015).

[125] Williams, O. E., Lillo, F. & Latora, V. Effects of memory on spreading processes in non-Markovian temporal networks. *New Journal of Physics* **21**, 043028 (2019).

[126] Williams, O.E., Lacasa, L., Millán, A. P., & Latora, V. The shape of memory in temporal networks. *Nature communications*, 13(1), 1-8(2022).

[127] Pfitzner, R., Scholtes, I., Garas, A., Tessone, C. J., & Schweitzer, F. Betweenness preference: Quantifying correlations in the topological dynamics of temporal networks *Physical Review Letters* **110**, 198701 (2013).

[128] Scholtes, I., Wider, N. & Garas, A. Higher-order aggregate networks in the analysis of temporal networks: path structures and centralities, *The European Physical Journal B* **89**, 1–15 (2016).

[129] Krackhardt, D. & Mark H. Heider vs Simmel: Emergent Features in Dynamic Structure. In *The Network Workshop Proceedings. Statistical Network Analysis: Models, Issues and New Directions*, 14–27 (New York: Springer, 2007).

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

[131] Liben-Nowell, D. and Kleinberg, J. The link-prediction problem for social networks. *Journal of the American society for information science and technology.* **58**, 1019–1031 (2007).

[132] Lü, L., Jin, C.H., & Zhou, T. Similarity index based on local paths for link prediction of complex networks. *Physical Review E* **80**, 046122 (2009).

[133] Abuoda, G., Morales, G. D. F., & Aboulnaga, A. Link prediction via higher-order motif features. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 412–429 (2019).

[134] Salton, G. & McGill, M. J., Introduction to Modern Information Retrieval. (McGrawHill, 1983).

[135] Molloy, M., & Reed, B. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms* **6**, 161–180 (1995).

[136] Newman, M. E. Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E* **64**, 016131 (2001).

[137] Barabási, A. L., Jeong, H., Néda, Z. Ravasz, E., Schubert, A., & Vicsek, T. Evolution of the social network of scientific collaborations. *Physica A* **311**, 590–614 (2002).

[138] Zhou, T., Lü, L. & Zhang, Y.-C. Predicting missing links via local information. *The European Physical Journal B* **71**, 623–630 (2009).

[139] Leicht, E. A., Holme, P., & Newman, M. E. Vertex similarity in networks. *Phys. Rev. E* **73**, 026120 (2006).

[140] Chebotarev, P., & Shamis, E. The matrix-forest theorem and measuring relations in small social groups. *Autom. Remote Control* **58**, 1505 (1997).

[141] Barandela, R., Sánchez, J. S., Garca, V., & Rangel, E., Strategies for learning in class imbalance problems. *Pattern Recognition* **36**, 849–851 (2003).

[142] Sundararajan, A. *The sharing economy: The end of employment and the rise of crowd-based capitalism* (MIT Press, 2016).

[143] Shaheen, S. & Chan, N. Mobility and the sharing economy: Potential to facilitate the first-and last-mile public transit connections. *Built Environment* **42**, 573–588 (2016).

[144] Sun, Y. Sharing and riding: how the dockless bike sharing scheme in china shapes the city. *Urban Science* **2**, 68 (2018).

[145] Tu, Y., Chen, P., Gao, X., Yang, J. & Chen, X. How to make dockless bikeshare good for cities: Curbing oversupplied bikes. *Transportation Research Record* **2673**, 618–627 (2019).

[146] Monge, G. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris* (1781).

[147] Carey, H. C. *Principles of social science*, vol. 3 (JB Lippincott & Company, 1867).

[148] Szell, M., Mimar, S., Perlman, T., Ghoshal, G., & Sinatra, R. Growing urban bicycle networks. *Scientific Reports*, 12(1), 1-14 (2022).

[149] Zipf, G. K. The p1 p2/d hypothesis: on the intercity movement of persons. *American sociological review* **11**, 677–686 (1946).

[150] Jung, W.-S., Wang, F. & Stanley, H. E. Gravity model in the korean highway. *EPL (Europhysics Letters)* **81**, 48005 (2008).

[151] Balcan, D. *et al.* Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences* **106**, 21484–21489 (2009).

[152] Krings, G., Calabrese, F., Ratti, C. & Blondel, V. D. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment* **07**, 1–9 (2009).

[153] Expert, P., Evans, T. S., Blondel, V. D. & Lambiotte, R. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences* **108**, 7663–7668 (2011).

[154] Palchykov, V., Mitrović, M., Jo, H.-H., Saramäki, J. & Pan, R. K. Inferring human mobility using communication patterns. *Scientific Reports* **4**, 1–6 (2014).

[155] Kaluza, P., Kölzsch, A., Gastner, M. T. & Blasius, B. The complex network of global cargo ship movements. *Journal of the Royal Society Interface* **7**, 1093–1103 (2010).

[156] Ravenstein, E. G. The laws of migration. *Journal of the Royal Statistical Society* **52**, 241–305 (1889).

[157] Tobler, W. Migration: Ravenstein, thornthwaite, and beyond. *Urban Geography* **16**, 327–343 (1995).

[158] Park, H. J., Jo, W. S., Lee, S. H. & Kim, B. J. Generalized gravity model for human migration. *New Journal of Physics* **20**, 093018 (2018).

[159] Fagiolo, G. The international-trade network: gravity equations and topological properties. *Journal of Economic Interaction and Coordination* **5**, 1–25 (2010).

[160] Pan, R. K., Kaski, K. & Fortunato, S. World citation and collaboration networks: uncovering the role of geography in science. *Scientific Reports* **2**, 902 (2012).

[161] Batty, M. The size, scale, and shape of cities. *Science* **319**, 769–771 (2008).

[162] Batty, M. *The new science of cities* (MIT Press, 2013).

[163] Xu, Y., Olmos, L. E., Abbar, S. & González, M. C. Deconstructing laws of accessibility and facility distribution in cities. *Science Advances* **6**, eabb4112 (2020).

[164] Olmos, L. E. *et al.* A data science framework for planning the growth of bicycle infrastructures. *Transportation research part C: emerging technologies* **115**, 102640 (2020).

[165] Erlander, S. & Stewart, N. F. *The gravity model in transportation analysis: theory and extensions*, vol. 3 (Vsp, 1990).

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

[167] Dong, L., Li, R., Zhang, J. & Di, Z. Population-weighted efficiency in transportation networks. *Scientific Reports* **6**, 26377 (2016).

[168] Ortuzar, J. & Willumsen, L. *Modelling transport.* (New York: John Wiley & Sons., 2011).

[169] Viboud, C. *et al.* Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science* **312**, 447–451 (2006).

[170] Deville, P. *et al.* Scaling identity connects human mobility and social interactions. *Proceedings of the National Academy of Sciences* **113**, 7047–7052 (2016).

[171] Li, R., Wang, W. & Di, Z. Effects of human dynamics on epidemic spreading in côte d'ivoire. *Physica A: Statistical Mechanics and its Applications* **467**, 30–40 (2017).

[172] Li, R., Richmond, P. & Roehner, B. M. Effect of population density on epidemics. *Physica A: Statistical Mechanics and its Applications* **510**, 713–724 (2018).

[173] Lu, X., Bengtsson, L. & Holme, P. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences* **109**, 11576–11581 (2012).

[174] Bagrow, J. P., Wang, D. & Barabasi, A.-L. Collective response of human populations to large-scale emergencies. *PLoS One* **6**, e17680 (2011).

[175] Yan, X.-Y., Wang, W.-X., Gao, Z.-Y. & Lai, Y.-C. Universal model of individual and population mobility on diverse spatial scales. *Nature Communications* **8**, 1–9 (2017).

[176] Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 96–100 (2012).

[177] Masucci, A. P., Serras, J., Johansson, A. & Batty, M. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E* **88**, 022812 (2013).

[178] Goh, S., Lee, K., Park, J. S. & Choi, M. Modification of the gravity model and application to the metropolitan seoul subway system. *Physical Review E* **86**, 026102 (2012).

[179] Noulas, A., Scellato, S., Lambiotte, R., Pontil, M. & Mascolo, C. A tale of many cities: universal patterns in human urban mobility. *PLoS One* **7**, e37027 (2012).

[180] Yan, X.-Y., Zhao, C., Fan, Y., Di, Z. & Wang, W.-X. Universal predictability of mobility patterns in cities. *Journal of The Royal Society Interface* **11**, 20140834 (2014).

[181] Lee, M. & Holme, P. Relating land use and human intra-city mobility. *PLoS One* **10**, e0140152 (2015).

[182] Li, R. *et al.* Simple spatial scaling rules behind complex cities. *Nature Communications* **8**, 1–7 (2017).

[183] WorldPop. `https://www.worldpop.org.uk/` (Accessed: 6th December 2019) (2017).

[184] Barthélemy, M. Spatial networks. *Physics Reports* **499**, 1–101 (2011).

[185] Wilson, A. G. The use of entropy maximising models, in the theory of trip distribution, mode split and route split. *Journal of transport economics and policy* **3**, 108–126 (1969).

[186] Biazzo, I. & Ramezanpour, A. Efficiency and irreversibility of movements in a city. *Scientific Reports* **10**, 1–8 (2020).

[187] Dong, L., Huang, Z., Zhang, J. & Liu, Y. Understanding the mesoscopic scaling patterns within cities. *arXiv preprint arXiv:2001.00311* (2020).

[188] Hong, I., Jung, W.-S. & Jo, H.-H. Gravity model explained by the radiation model on a population landscape. *PLoS One* **14**, e0218028 (2019).

[189] Bettencourt, L. M., Lobo, J., Helbing, D., Kühnert, C. & West, G. B. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences* **104**, 7301–7306 (2007).

[190] Bettencourt, L. M. The origins of scaling in cities. *Science* **340**, 1438–1441 (2013).

[191] Li, R., Dong, L., Wang, X. & Zhang, J. The geometric origins of complex cities. In *Proceedings of ECCS 2014*, 45–57 (Springer, 2016).

[192] Šubelj, L. Algorithms for spanning trees of unweighted networks. Tech. Rep., University of Ljubljana (2021).

[193] Wasserman, S. & Faust, K. *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)* (Cambridge University Press, 1994).

[194] Newman, M. *Networks: an introduction* (Oxford University Press, 2010). [71] Watts, D. J., & Strogatz, S. H. . Collective dynamics of 'small-world'networks. *Nature*, 393(6684), 440-442(1998).

[195] Erdős, P. & Réyni, A. On random graphs. i. *Publicationes Mathematicae* **6**, 290–297 (1959). URL `http://www.renyi.hu/~p_erdos/1959-11.pdf`.

[196] Bollobás, B. Mathematical results on scale-free random graphs. In *Handbook of Graphs and Networks*, 1–37 (Wiley, 2003).

[197] Milgram, S. The small world problem. *Psychology Today* (1967).

[198] Travers, J. & Milgram, S. An experimental study of the small world problem. *Sociometry* **32**, 425 (1969).

[199] Leskovec, J. & Horvitz, E. Planetary-scale views on an instant-messaging network. Tech. Rep., Microsoft Research (2007). `arXiv:0803.0939v1`.

[200] Backstrom, L., Boldi, P., Rosa, M., Ugander, J. & Vigna, S. Four degrees of separation. In *Proceedings of the 3rd Annual ACM Web Science Conference on - WebSci '12* (ACM Press, 2012).

[201] Boldi, P. & Vigna, S. Four degrees of separation, really. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (IEEE, 2012).

[202] Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).

[203] Chung, F. & Lu, L. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences* **99**, 15879–15882 (2002).

[204] Dorogovtsev, S. N., Mendes, J. F. F. & Samukhin, A. N. Metric structure of random networks. *Nuclear Physics B* **653**, 307–338 (2003).

[205] Pržulj, N., Corneil, D. G., & Jurisica, I., Modeling interactome: scale-free or geometric?, *Bioinformatics* **20**,3508–3515 (2004).

[206] Pržulj.N. Biological network comparison using graphlet degree distribution. *Bioinformatics* **23**, e177–e183 (2007).

[207] Burt, R.S. Structural Holes and Good Ideas. *American Journal of Sociology* **110**, 349–399 (2004).

[208] Zhang, X., Moore, C., & Newman, M.E. Random graph models for dynamic networks. *The European Physical Journal B* **90**, 1-14 (2017).

[209] Hagberg, A., Swart, P., & S Chult, D. and Exploring network structure, dynamics, and function using NetworkX. Los Alamos National Lab.(LANL), Los Alamos, NM (United States) (2008).

[210] Brandes, U., Robins, G. McCranie, A., & Wasserman, S. *Network Science* **1**, 1–15 (2013).

[211] Kovanen, L., Kaski, K., Kertész, J., & Saramäki, J. Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. *Proceedings of the National Academy of Sciences* **110**, 18070–18075 (2013).

[212] Hagberg ,A., Schult A. & Swart, P., Exploring network structure, dynamics, and function using NetworkX, in *Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA)*, 11-15(2008).

[213] Sarukkai, R. R. Link prediction and path analysis using markov chains. *Computer Networks* **33**, 377–386 (2000).

[214] Popescul, A., & Ungar, L. H. Statistical relational learning for link prediction. In *IJCAI workshop on learning statistical models from relational data 2003* (2003).

[215] Bilgic, M., Namata, G. M., & Getoor, L. Combining collective classification and link prediction. In *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)* 381–386, IEEE (2007).

[216] Yu, K., Chu, W., Yu, S., Tresp, V., & Xu, Z. Stochastic relational models for discriminative link prediction. In *NIPS* **6**, 1553–1560 (2006).

[217] Clauset, A., Moore, C., & Newman, M. E. Hierarchical structure and the prediction of missing links in networks. *Nature* **453**, 98–101 (2008).

[218] Freeman, L. C. Centrality in social networks conceptual clarification. *Social networks* **1**, 215–239 (1978).

[219] Borgatti, S. P. & Everett, M. G. A graph-theoretic perspective on centrality. *Social Networks* **28**, 466–484 (2006-10).

[220] Landherr, A., Friedl, B. & Heidemann, J. A critical review of centrality measures in social networks. *Business & Information Systems Engineering* **2**, 371–385 (2010).

[221] Brandes, U. & Hildenbrand, J. Smallest graphs with distinct singleton centers. *Network Science* **2**, 416–418 (2014).

[222] Das, K., Samanta, S. & Pal, M. Study on centrality measures in social networks: a survey. *Social Network Analysis and Mining* **8** (2018).

[223] Falkenberg, M. *et al.* Identifying time dependence in network growth. *Physical Review Research* **2**, 023352 (2020).

[224] Zhou, B., Meng, X. & Stanley, H. E. Power-law distribution of degree-degree distance: A better representation of the scale-free property of complex networks. *Proceedings of the National Academy of Sciences* **117**, 14812–14818 (2020).

[225] Olami, Z., Feder, H. J. S., & Christensen, K. Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. *Physical review letters*, **68**(8), 1244 (1992).

[226] Peters, O., & Christensen, K. Rain: Relaxations in the sky. *Physical Review E*, **66**(3), 036120 (2002).

[227] Christensen, K., Manani, K. A., & Peters, N. S. Simple model for identifying critical regions in atrial fibrillation. *Physical review letters*, **114**(2), 028104, (2015).

[228] Frette, V., Christensen, K., Malthe-Sørenssen, A., Feder, J., Jøssang, T., & Meakin, P. Avalanche dynamics in a pile of rice. *Nature* **379**(6560), 49-52 (1996).

[229] Jansen, V. A., Stollenwerk, N., Jensen, H. J., Ramsay, M. E., Edmunds, W. J., & Rhodes, C. J.. Measles outbreaks in a population with declining vaccine uptake. *Science*, **301**(5634), 804-804 (2003).

[230] Christensen, K., Di Collobiano, S. A., Hall, M., & Jensen, H. J. Tangled nature: a model of evolutionary ecology. *Journal of theoretical Biology*, **216**(1), 73-84 (2002).

[231] Dolan, D., Sloboda, J., Jensen, H. J., Crüts, B., & Feygelson, E. The improvisatory approach to classical music performance: An empirical investigation into its characteristics and impact. Music Performance Research, **6**, 1-38 (2013).

[232] Holme, P., Congestion and centrality in traffic flow on complex networks. *Advances in Complex Systems*, **6**(02), 163-176(2003).

[233] Ashton, Douglas J., Timothy C. Jarrett, and Neil F. Johnson. "Effect of congestion costs on shortest paths through complex networks." *Physical review letters* **94**(5) , 2005.