KYOTO UNIVERSITY

DOCTORAL THESIS

Modeling and Statistical Inference of Preferential Attachment in Complex Networks: Underlying Formation of Local Community Structures

Author:
Masaaki INOUE

Supervisor:
Hidetoshi SHIMODAIRA

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Informatics

in the

Department of Systems Science Graduate School of Informatics

January, 2022

Abstract

This thesis discusses the modeling and statistical inference of preferential attachment (PA) in complex networks that exhibit social community characteristics. Growth models of PA explain the scale-free nature of degree distribution in real-world networks, including social and human networks such as scientific co-authorship. However, the PA mechanism depends only on the degree of each individual, which makes it difficult to form community clusters under the cooperative behavior of multiple individuals.

In Chapter 1, we present the motivation and overview of this thesis. This chapter also describes the following issues we tackle in this thesis:

(i) PA and transitivity, the classic and simple mechanisms, are widely used to explain the formation of the heavy tail of the degree distribution and the high clustering in real-world networks, respectively. Since one of the above simple mechanisms is not well suited to capture both features, many existing studies have attempted to reveal the formation of the two features by considering both PA and transitivity. Existing approaches either estimate one mechanism in isolation or jointly estimate both mechanisms assuming some functional forms. Each of them has the problem of poor fitting or risks losing the fine details of the two mechanisms.

(ii) There exist social networks where the collectivity of interactions is lost when expressed in graphs. Since group interactions such as collaborative behavior may contain more than two individuals, in graph expression, each of them is decomposed into multiple edges: the pre-specification that all interactions are pairwise relationships.

For solving the above issues, we present our methodologies, findings, and discussions in Parts I- III, which are the three main parts of this thesis.

In Chapter 2, we describe the background knowledge of the network growth models related to the subsequent chapters of this thesis.

In Part I, we address issue (i). We present our methodology in Chapter 3, and the real data analysis in Chapter 4. We propose a statistical method for estimating the non-parametric PA and transitivity functions simultaneously in a growing network, in contrast to conventional methods that either estimate each function in isolation or assume a certain functional form for these. Our model is demonstrated to exhibit a good fit to two real-world co-authorship networks and can illuminate several intriguing details of the PA and transitivity phenomena that would be unavailable under traditional methods. Moreover, we introduce a method for quantifying the amount of contributions of these phenomena in the growth process of a network based on the probabilistic dynamic process induced by the model formula. By applying this method, we found that transitivity dominated PA in both co-authorship networks. This suggests the importance of indirect relations in scientific creative processes. The proposed method is implemented and publicly available in the R package FoFaF.

In Part II, we address issue (ii). We present our methodology in Chapter 5, and the real data analysis in Chapter 6. We propose a new hypergraph growth model with a data-driven PA mechanism estimated from observed data. A key component of our method is a recursive formula that allows us to overcome a bottleneck in computing the normalizing factors in our model. We also treat an often-neglected selection bias in modeling the emergence of new edges with new nodes. Fitting the proposed hypergraph model to 13 real-world datasets from diverse domains, we found that all estimated PA functions deviate substantially from the linear form. This demonstrates the need to do away with the linear PA assumption and adopting a data-driven approach. We also showed that our model outperformed conventional models in replicating the observed first-order and second-order structures in these real-world datasets. The proposed method is implemented in the R package HyperPA, which will be published on IEEE Xplore Code Ocean.

We present the conclusion of this thesis and the future directions in Part III, which includes Chapter 7.

Acknowledgements

I deeply thank Professor Hidetoshi Shimodaira, a supervisor in my master's and doctoral course, for giving me a great many insightful comments and suggestions on the thesis. I am grateful to Dr. Thong Pham for his daily advice and encouragement on my research activities in complex networks. I am also grateful to Professor Kazunori Hayashi, a supervisor in my bachelor's course, for teaching me the basics of research. I would also like to thank my thesis committee members, Professor Toshiyuki Tanaka and Professor Manabu Kano, for all of the time they have taken to be on my committee and provide their feedback. I warmly thank all the past and present members of our laboratory, Statistical Intelligence. Finally, I thank my family for all the support that I have received in my life.

List of Publications

Related Journal Papers

1. **Masaaki Inoue**, Thong Pham, & Hidetoshi Shimodaira. (2022). A Hypergraph Approach for Estimating Growth Mechanisms of Complex Networks [Early Access Version], *IEEE Access*. doi:10.1109/ACCESS.2022.3143612 (Copyright © 2022 IEEE)

2. **Masaaki Inoue**, Thong Pham, & Hidetoshi Shimodaira. (2020). Joint Estimation of Non-parametric Transitivity and Preferential Attachment Functions in Scientific Co-authorship Networks, *Journal of Informetrics*, 14(3), 101042. doi:10.1016/j.joi.2020. (Copyright © 2020 Elsevier)

Related International Conference Proceeding

3. **Masaaki Inoue**, Thong Pham, & Hidetoshi Shimodaira. (2018). Transitivity vs Preferential Attachment: Determining the Driving Force behind the Evolution of Scientific Co-authorship Networks, In: *Unifying Themes in Complex Systems IX. ICCS 2018. Springer Proceedings in Complexity*, Springer, 262-271. doi:10.1007/978-3-319-96661-8_28 (Copyright © 2018 Springer Nature)

Other Journal Paper

4. **Masaaki Inoue**, Kazunori Hayashi, Hiroki Mori, & Toshihisa Nabetani. (2018). A DOA Estimation Method With Kronecker Subspace for Coherent Signals, *IEEE Communications Letters*, 22(11), 2306-2309. doi:10.1109/LCOMM.2018.2870824

Other International Conference Presentation

5. **Masaaki Inoue**, Thong Pham, & Hidetoshi Shimodaira. (2019, November 17). Statistical Estimation of the Effects of First and Second Order Local Structures on Growth of Complex Networks [Poster session], In: *ACML 2019 Workshop on Statistics and Machine Learning Researchers in Japan*, Nagoya, Japan.

Copyright Notice

Contents

Chapter 1

Introduction

In this chapter, we first describe the motivation of this thesis in Section 1.1. Secondly, in Section 1.2, we describe the two main issues that we tackle in this thesis. We then summarize our contributions to the issues in Section 1.3. Finally, we briefly describe the organization of this thesis in Section 1.4. Descriptions in this section are based on our related journal papers (Inoue et al., 2020b, 2022) and international conference proceedings (Inoue et al., 2018). The figures in this section are newly created for this thesis.

1.1 Motivation

This thesis originated from some issues about preferential attachment (PA) that we realized when trying to capture the formation process of the community structures in scientific co-authorship networks. Investigating the PA mechanism, the"rich get richer" phenomenon in complex networks, is one of the most classical and important topics in network science (Barabási & Albert, 1999). PA explains the scale-free nature of degree distribution in real-world networks, including social and human networks such as scientific co-authorship. However, the PA mechanism depends only on the degree of each individual, which makes it difficult to form community clusters under the cooperative behavior of multiple individuals.

Cooperation is among the most fundamental behaviors of living creatures (Nowak, 2006). Animals cooperate in various activities: from hunting and forming territories to grooming and child raising (Dugatkin, 1997).

Humans are the experts of cooperation. From cooperation between states (Watson, 1984), companies (Hamel et al., 1989) to cooperation between individuals (David W. Johnson, 1991), it is the bedrock of our society.

As a form of human cooperation that builds social communities, scientific collaboration is the backbone of the scientific world. In this process, scientists share their ideas, their time, and their skills with each other in order to push the boundary of knowledge. Since the start of the twentieth century, the number of scientific articles with more than one author has grown to more than three times the number of single-author articles (Larivière et al., 2015). There is accumulating evidence that articles resulting from collaborations are cited more frequently than non-collaborated ones (Bornmann, 2017; Larivière et al., 2015). Since the number of citations is the main metric of scientific impact (Tahai & Meyer, 1999), collaborations thus lead to high impact research. Therefore, understanding the evolution of a scientific co-authorship network, in particular, understanding how new collaborations are fostered, is significantly important for policy makers, funding agencies, university managers as well as each scientist. To understand the evolution of social and human networks, it is beneficial to consider the evolution in a larger context of the evolution of complex networks.

1.2 Issues in Preferential Attachment

In this section, we describe the two issues in PA in terms of the formation of local community structures. We tackle these two issues in this thesis.

1.2.1 Poor Fit of Preferential Attachment to High Clustering Features in Real-world Networks

Two defining characteristics of an evolving complex network are the heavy-tail of the degree distribution and the high value of the clustering coefficient (Holme & Kim, 2002); both are often represented at the same time in social and human networks such as scientific co-authorship networks (Newman, 2001b, 2004). Previous studies suffered from difficulties in capturing the driving forces that form these two co-existing features with a single comprehensive growth model. Following studies have discussed the formation of each feature through the data-driven growth models and estimation of the functions that determine the growth.

On one hand, to explain the heavy tail property, complex network studies have proposed the PA mechanism where the probability that a node with degree d receives a new link is proportional to the PA function A_d (Krapivsky et al., 2001; Pham et al., 2015). When A_d is an increasing function on average, nodes with higher numbers of links will receive more new links, and thus hubs are formed and the heavy-tail degree distribution emerges.

On the other hand, one of the simplest mechanisms to explain the high value of the clustering coefficient is transitivity where the probability that a pair of nodes with b common neighbors receives a new link connecting them is proportional to the transitivity function B_b (Newman, 2001a). When B_b is an increasing function on average, more triangles are formed between sets of three nodes, and this leads to an increase in the clustering coefficient.

Existing approaches either estimate one mechanism in isolation (Jeong et al., 2003; Newman, 2001a; Pham et al., 2015) or estimate jointly the two mechanisms assuming some parametric forms for A_d and B_b (Krivitsky & Handcock, 2019; Ripley et al., 2018). On one hand, estimating either mechanism in isolation often leads to poor fit, since many real-world networks simultaneously exhibit heavy-tail degree distribution and high clustering. On the other hand, it is difficult to justify a particular choice of functional forms used in parametric estimation methods. A non-parametric estimation method would allow the functional forms to be learned from the observed data.

Estimating A_d and B_b is the first step towards answering the question of what matters more in the evolution of a complex network: transitivity or PA. While there is some research studying a similar question regarding

PA and fitness (Kong et al., 2008), the question regarding PA and transitivity has curiously remained unexplored, despite its potential to provide deeper understandings on how new cooperation is fostered.

1.2.2 Preferential Attachment Under the Limitation of Graph Representations

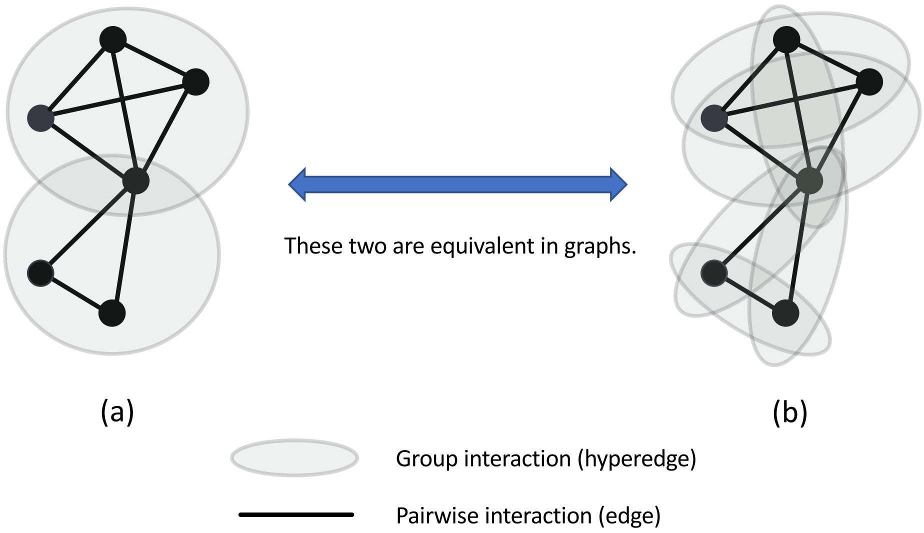

These two are equivalent in graphs.
(a)
(b)
Group interaction (hyperedge)
Pairwise interaction (edge)

FIGURE 1.1: Examples of two hypergraphs that are projected onto the same graph. Graphs do not preserve the dependencies of edges.

Graphs are widely used in various fields, including complex network theory, to analyze interactions among individuals and their dynamics. However, there are some data where the range of interactions is lost when expressed in graphs. For example, in the co-authorship data of scientific papers, a new co-authorship relationship is added by giving a paper to the set of authors. In such data, the interaction may contain more than two individuals, but graphs cannot capture the feature. This is because, in graphs of such data, an original group interaction is projected as multiple edges, each of which is a pairwise interaction. Figure 1.1 shows two instances where the arrangement and the number of original interactions are different, but are equivalent in graph representation. Most of the existing growth models for complex networks rely on graph representations and thus fail to capture this feature in the growth process. The limitations of the graph representation in the case of growth processes are discussed in detail in Section 2.2.2. In those cases, hypergraph representation may be preferable since it can preserve higher-order relationships with hyperedges. However, existing hypergraph models of temporal complex networks employ a data-independent growth mechanism, which is the linear PA (Do et al., 2020). In principle, this pre-specification is undesirable since it completely ignores the data at hand.

1.3 Contributions

In the subsequent parts of this thesis, we provide the methodologies and result of real data analysis as solutions for the two issues of PA described above in Section 1.2.1 and Section 1.2.2, respectively. Our contributions are as follows.

1.3.1 FoFaF: Joint Estimation of Non-parametric Preferential Attachment and Transitivity

To address the issues discussed in Section 1.2.1, we propose a statistical method for estimating the non-parametric PA and transitivity functions simultaneously in a growing network, in contrast to conventional methods that either estimate each function in isolation or assume a certain functional form for these. Our model is demonstrated to exhibit a good fit to two real-world co-authorship networks and can illuminate several intriguing details of the PA and transitivity phenomena that would be unavailable under traditional methods. Moreover, we introduce a method for quantifying the amount of contributions of these phenomena in the growth process of a network based on the probabilistic dynamic process induced by the model formula. By applying this method, we found that transitivity dominated PA in both co-authorship networks. This suggests the importance of indirect relations in scientific creative processes. The proposed method is implemented in the R package FoFaF (Inoue et al., 2020a). The methodology and results of real data analysis are based on our papers (Inoue et al., 2018, 2020b), and presented in Part I.

1.3.2 HyperPA: A Hypergraph Approach for Estimating Preferential Attachment

To address the issues discussed in Section 1.2.2, we propose a new hypergraph growth model with a data-driven PA mechanism estimated from observed data. A key component of our method is a recursive formula that allows us to overcome a bottleneck in computing the normalizing factors in our model. We also treat an often-neglected selection bias in modeling the emergence of new edges with new nodes. Fitting the proposed hypergraph model to 13 real-world datasets from diverse domains, we found that all estimated PA functions deviate substantially from the linear form. This demonstrates the need to do away with the linear PA assumption and to adopt a data-driven approach. We also showed that our model outperformed conventional models in replicating the observed first-order and second-order structures in these real-world datasets. The proposed method is implemented in the R package HyperPA and will be available in (Inoue et al., 2022). The methodology and results of real data analysis are based on our paper (Inoue et al., 2022), and presented in Part II.

1.4 Organization

We describe the organization of this thesis in this section. Figure 1.2 illustrates the structure of this thesis. So far, we have provided the motivations, issues, and contributions in this chapter. In Chapter 2, we first review the backgrounds of graph growth models of complex networks, and temporal hypergraph, which is another representation of complex network data.

Our methodology and findings are discussed in the following three parts. In Part I, we describe the joint estimation method of non-parametric

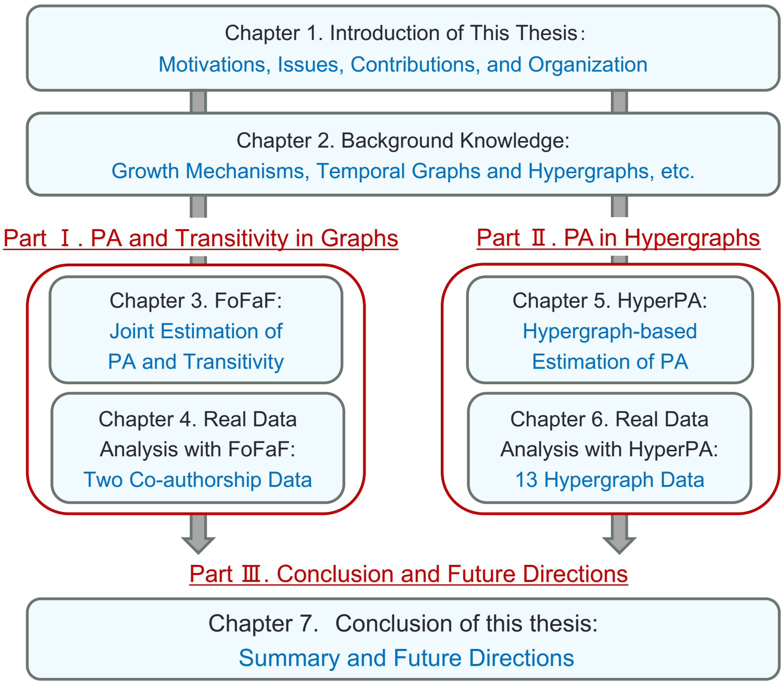

Chapter 1. Introduction of This Thesis:
Motivations, Issues, Contributions, and Organization
Chapter 2. Background Knowledge:
Growth Mechanisms, Temporal Graphs and Hypergraphs, etc.
Part Ⅰ. PA and Transitivity in Graphs
Part Ⅱ. PA in Hypergraphs
Chapter 3. FoFaF:
Joint Estimation of
PA and Transitivity
Chapter 4. Real Data
Analysis with FoFaF:
Two Co-authorship Data
Chapter 5. HyperPA:
Hypergraph-based
Estimation of PA
Chapter 6. Real Data
Analysis with HyperPA:
13 Hypergraph Data
Part Ⅲ. Conclusion and Future Directions
Chapter 7. Conclusion of this thesis:
Summary and Future Directions

FIGURE 1.2: Organization of this thesis.

PA function and transitivity function in Chapter 3, and present the result of real-world data analysis in Chapter 4. In Part II, we describe a hypergraph approach to estimate the non-parametric PA function which relieves the assumption of edge independence that is required in conventional graph models in Chapter 5, and present the result of real-world data analysis in Chapter 6. Summary of this thesis and future directions are presented in Part III which includes Chapter 7.

Chapter 2

Background

In this chapter, we describe background knowledge and some preliminaries of this thesis. We first briefly review the history of preferential attachment (PA) and transitivity modeling, which are classic growth mechanisms of complex networks. We then describe General Temporal (GT) model that encompasses important existing graph growth models. Finally, we describe the temporal graph model, which is a new approach to growth modeling of complex networks. Descriptions and figures in this chapter are based on our papers (Inoue et al., 2020b, 2022).

2.1 Graph-based Growth Model

In this section, we describe the background of PA mechanism, transitivity mechanism and GT model as the preliminaries of Part I and Part II.

2.1.1 Preferential Attachment and Transitivity

The concept of the rich-get-richer phenomenon has its roots in the theoretical works of Yule (Yule, 1925) and Simon (Simon, 1955). Its status as a fundamental process in informetrics was cemented by the revolutionary works of Merton (Merton, 1968) and Price (Price, 1965, 1976). The term "preferential attachment" was coined by Barabási and Albert when they re-discovered the mechanism in the context of complex networks (Barabási & Albert, 1999). In this thesis, we refer to the individuals and their interactions represented by graph structures as "networks", and the networks observed on a large scale in the real world are called "complex networks". In order to concentrate on applications to complex networks, we will leave the detailed mathematical description of graph theory to the textbooks such as (Godsil & Royle, 2001). However, it is necessary to clarify here that, for practical application, we add the following conditions to graphs in our mathematical models. We assume that the edges of the graphs have no directions and allow multiple edges to the graphs. The allowance for multiple edges is in accordance with the repeated occurrence of interactions between individuals in real data. In this thesis, we refer to "undirected graph without self-loops" as "graph".

In PA, the probability that a node with degree d will receive a new edge is proportional to its PA function A_d. When A_d is an increasing function on average, the PA effect exists: a node with a large degree d is more likely to receive more new connections. Estimating the PA phenomenon in a network amounts to estimating the function A_d given the growth data of that network. Various non-parametric approaches (Newman, 2001a; Pham et al., 2015) and parametric methods (Gómez et al., 2011; Massen & Jonathan, 2007) have been proposed. Power-law function forms, such as $A_d = (d+1)^{\alpha}$, are often employed in parametric methods (Krapivsky & Redner, 2001).

Transitivity originated as a concept in psychology (Heider, 1946) and was developed theoretically in the framework of social network analysis by Holland and Leinhardt in the 1970s (Holland & Leinhardt, 1970, 1971, 1976). It was introduced into the informetrician modeling toolbox in 2001 when Newman provided a heuristic method to estimate the transitivity function in real-world co-authorship networks (Newman, 2001a). Independently at the same time, Snijders introduced his now-famous stochastic actor-based models that include transitivity as a network formation mechanism (Snijders, 2001).

In transitivity, the probability that a pair of two nodes with b common neighbors will receive a new edge between them is proportional to the transitivity function B_b. When B_b is an increasing function on average, the transitivity effect exists: when a pair of nodes shares more common neighbors, it is easier for them to connect. Similarly to the case of PA, non-parametric approaches (Newman, 2001a) and parametric approaches (Ferligoj et al., 2015; Kronegger et al., 2012; Zinilli, 2016) have been proposed to estimate B_b from observed network growth data.

PA and transitivity are among the simplest and most comprehensive mechanisms using the first-order and second-order structures of graphs, respectively, which is a main reason why we analyze them in Part I to get a closer look at the growth mechanisms behind the real-world networks. We re-emphasize that all existing methods consider either PA or transitivity in isolation, or are of a parametric nature.

2.1.2 General Temporal Model

We describe an undirected graph version of the General Temporal (GT) model with PA (Pham et al., 2015). This model is a generalization of various classical models such as Barabási-Albert model (Barabási & Albert, 1999) and Price's model (Price, 1976). In the GT model, the probability that a node pair i, j will acquire an edge connecting them at time t is expressed as:

$$P_{i,j}(t) \propto A_{d_i(t)} A_{d_j(t)}, \tag{2.1}$$

where $d_i(t), d_j(t)$ are the degree of node i, j at time-step t, and A_d is the PA value of degree d. The function A_d of d is often called the attachment function or attachment kernel. Note that A_d is assumed to be time-invariant. This graph PA model is hereinafter referred to as "Edge PA". In Edge PA, no functional form is assumed on the PA function A_d. Several methods have been proposed to estimate the PA function A_d from the temporal network data. Parametric estimation methods for estimating A_d based on the assumption that the PA function has the functional form $A_d = d^\alpha$ with a tunable parameter α include regression-based methods (Kunegis et al., 2013), maximum likelihood estimation methods (Gómez et al., 2011), and methods based on Markov chain Monte Carlo (Sheridan et al., 2012). There are also nonparametric estimation methods that do not make assumptions on the functional form of the PA function A_d including methods using histograms (Jeong et al., 2003; Newman, 2001a) and maximum likelihood estimation (Pham et al., 2015).

2.2 Temporal Hypergraph

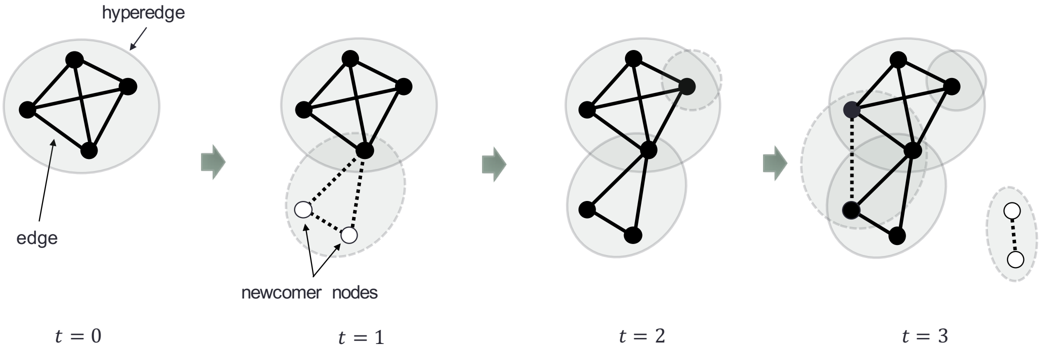


FIGURE 2.1: Hypergraph expression and graph expression of temporal data. Whereas a graph consists of nodes and edges, a hypergraph consists of nodes and hyperedges. Hyperedges and edges are indicated by ovals and line segments, respectively. Dashed ovals and line segments represent newly added hyperedges and edges at each time-step, respectively. White nodes represent newcomer nodes.

In this section, as preliminaries of Part II, we explain some properties of discrete-time hypergraph-based growth models, comparing it with conventional growth models for graphs.

2.2.1 Growth of Hypergraph

Let $G_t = (V_t, E_t)$ be the hypergraph at time-step $t = 0, \ldots, T$. The hypergraph G_t consists of the node set V_t and the hyperedge set E_t existing at time-step t. The temporal hypergraph grows from $G_0 = (V_0, E_0)$, the initial hypergraph, with the emergence of new hyperedges and nodes at each time-step. Similarly to graphs, we add the following conditions to hypergraphs in our mathematical models. We assume that hyperedges do not have any order, and we also allow multiple hyperedges to the hypergraphs. Although a hypergraph where the set of nodes in a hyperedge does not have any order is called an “undirected hypergraph” (Michoel & Nachtergaele, 2012), in this thesis we simply refer to it as a “hypergraph”.

Fig. 2.1 illustrates a discrete-time temporal hypergraph and its graph representation. To handle some features of the real-world hypergraph data collected by discrete-time observations, we explicitly allow the following two points in the temporal hypergraph model.

1. The hyperedge added at each time-step t can include both existing nodes and new nodes. This is illustrated in the hypergraph at time-step $t = 1$ in Fig. 2.1. We call these new nodes newcomer nodes.

2. The number of hyperedges added at each time-step t can be more than one. An example is given in the hypergraph at time-step $t = 3$ in Fig. 2.1.

Guimerà et al. (2005) proposed a probabilistic model of temporal hypergraphs that controls the proportion of newcomer nodes that appear with hyperedges, and analyzed the relationship between this proportion and the success of collaboration. In this paper, both our proposed estimation method and generator for hypergraphs address the above two points.

2.2.2 Information loss of Graph Representation

We next explain some information losses that can occur with graph representations of complex temporal data. The growth of complex network data such as co-authorships of papers is conventionally modeled by ordinary graphs, where each motif (e.g., paper) occurring among nodes (e.g., authors) is represented by edges (e.g., pairwise co-authorships). When a motif contains more than two nodes (e.g., a paper with more than two authors), the group interaction created by the motif is decomposed into multiple edges. On the other hand, in hypergraphs, one motif is represented by one hyperedge. An example is given in the hypergraph at time-step $t = 0$ in Fig. 2.1. In order to present a motif on four nodes, in hypergraph representations, a hyperedge containing these four nodes is used, whereas in ordinary graph representations, six ordinary edges, which constitute a clique on the four nodes, are used instead. Generally, when a motif occurs among m nodes, in ordinary graphs, $m(m-1)/2$ edges are added collectively, whereas one hyperedge of size m is added in hypergraph representations. As can be seen from the hypergraph at time-step $t = 3$, given only one graph representation at one time-step, in general, we cannot identify which edges were added together in the past without hyperedge information. Thus, the information about motifs is not perfectly preserved when one employs a graph representation of the data.

Furthermore, there is another information loss when $m = 1$, i.e., when a motif contains only one node (e.g., a single-author paper) is added. In such cases, a hyperedge of size $m = 1$ is added to the hypergraph, while the edges of the graph remain unchanged, as illustrated at time-step $t = 2$ of Fig. 2.1.

As mentioned above, graphs do not preserve the information about which edges appeared jointly. In other words, the conventional graph-based growth models assume implicitly that all edges are independent. Let m denote the size of a hyperedge. Although data with huge m exist in the real world, for example in multinational projects (Cronin, 2001; Hu et al., 2010), few studies have examined whether this independence assumption is appropriate under such large values of m. The datasets used in the experiments in Sections 5.2 and 6.1 contain hyperedges whose m are greater than 100. Using these datasets, we will investigate the performance of some conventional graph-based growth models in the case of large m, which has not been examined much in existing studies. In addition, it is reported that the modern science collaboration has shown a trend that hyperedge sizes are becoming larger. From the beginning of the 20th century to present, the average number of co-authors per paper has increased in almost all disciplines (Fortunato et al., 2018). Such changes in data over time may also lead to unforeseen problems for graph-based growth models.

The main motivation for considering hypergraph models in Chapter 4 is that it does not require the above independence assumption throughout the growth process. In Chapter 4, we describe our proposed approach to capture the characteristics of temporal hypergraphs.

Part I

PA and Transitivity Functions in Evolving Graphs

Chapter 3

FoFaF: Joint Estimation of Non-parametric PA and Transitivity Functions

Our contributions, descriptions and figures in this chapter are based on our papers (Inoue et al., 2018, 2020b).

3.1 Introduction

Science has never been more collaborative. In this era, which has been witnessing an unprecedented explosion of multi-author scholarly articles (Larivière et al., 2015), collaboration has become increasingly important in the quest for scientific success (Bornmann, 2017; Jones et al., 2008). Promising ideas from numerous analytical fields, including complex network theory, statistics, and informetrics, have been combined to understand the collaborative nature of science (Fortunato et al., 2018; Zeng et al., 2017).

An early attempt was made to analyze the formative process of scientific collaborations in physics when Newman proposed a non-parametric method to estimate the preferential attachment (PA) and transitivity functions from scientific collaboration networks (Newman, 2001a). PA (Barabási & Albert, 1999; Merton, 1968; Price, 1965, 1976) and transitivity (Heider, 1946; Holland & Leinhardt, 1970, 1971, 1976) are two fundamental mechanisms of network growth. PA is a phenomenon concerning the first-order structure of a network. In PA, a higher number of co-authors that a scientist already has will result in more collaborators being formed. Transitivity pertains to the second-order structure: co-authors of co-authors are likely to collaborate. The method of Newman is non-parametric in that it does not assume any forms for either the PA or transitivity function. However, the method considers each phenomenon in isolation, and thus completely ignores any entanglements of the two phenomena, which are entirely plausible in real-world networks.

In addition to this non-parametric, in-isolation approach, a joint estimation approach, in which the two phenomena are considered simultaneously, has been attempted in recent years (Ferligoj et al., 2015; Kronegger et al., 2012; Zinilli, 2016) under the framework of stochastic actor-based models (Snijders, 2001). This approach is inherently parametric: it assumes the forms of the PA and transitivity functions a *priori*, and therefore risks losing the fine details of the two phenomena, which are difficult to capture using any parametric functional forms.

We argue that the ideal method should combine the best of both worlds whenever possible: it should consider both phenomena simultaneously and it should not assume any functional forms for these.

The contributions of this part can be summarized as follows:

1. We propose a network growth model that combines non-parametric PA and transitivity functions for undirected networks, which is the type of scientific collaboration networks. We derive an efficient Minorize-Maximize (MM) algorithm (Hunter & Lange, 2000) for their simultaneous estimation. This iterative algorithm is guaranteed to increase the log-likelihood of the model per iteration. Using simulated examples, we demonstrate that our approach is capable of capturing the complex details of PA and transitivity, as opposed to conventional approaches (see Fig. 3.1). Furthermore, we perform a systematic simulation to confirm the performance of our algorithm.

2. We suggest a method for quantifying the amount of contributions of PA and transitivity in the growth process of a network. Our quantification exploits the probabilistic dynamic process induced by the

network growth formula and can easily be extended to other network growth mechanisms.

3. We apply the proposed method to two real-world co-authorship networks and uncover several interesting properties that would be unavailable under conventional approaches. In particular, the transitivity function appears to be substantially different from the typical power-law functional form. Moreover, we find that transitivity dominated PA in the growth processes of both networks. This suggests the importance of indirect relations in scientific creative processes: it does in fact matter whom your collaborators collaborate with. All of the proposed method is implemented in the R package `FoFaF` (Inoue et al., 2020a).

The remainder of this part is organized as follows. The proposed growth model is discussed in detail in Section 3.2. Section 3.3.2 presents the estimation method, derivation of an efficient MM algorithm for the estimation and the effectiveness of our methodology. In Section 3.4, we discuss how to exploit the probabilistic dynamic process imposed by the model formula to sensibly quantify the amount of contributions of PA and transitivity. We apply the proposed method to two real-world collaboration networks and analyze the results in Section 4.1. Concluding remarks are presented in Section 4.2.

3.2 Joint Non-parametric Modeling of PA and Transitivity

In this section, we first describe our network growth model that incorporates non-parametric PA and transitivity functions. Moreover, we explain its relation to several conventional network models.

3.2.1 Proposed Network Model

Our model for undirected networks can be viewed as a discrete Markov model, which is a popular framework in social network modeling (Holland & Leinhardt, 1977). Let G_t denote the network at time t. Starting from a seed network G_0, at each time step $t = 1, \cdots, T$, $v(t)$ new nodes and $m(t)$ new edges are added to G_{t-1} to form G_t. In particular, at the onset of time step t, let $d_i(t)$ denote the degree of node i, and $b_{ij}(t)$ denote the number of common neighbors between nodes i and j in G_{t-1}. Our model is based on the GT model (2.1) and dictates that the probability of a new edge emerging between nodes i and j at time step t is independent of the other new edges at that time and is equal to

$$P_{ij}(t) \propto A_{d_i(t)} A_{d_j(t)} B_{b_{ij}(t)}, \tag{3.1}$$

where A_d is the PA function of degree d and B_b is the transitivity function of the number b of common neighbors. That is, the unordered pair of the two ends (i, j) of a new edge follows a categorical distribution over all unordered pairs of nodes existing at time t. The weight of each pair is proportional to the product of the PA and transitivity values of that pair at t. Thus, this formulation can capture the PA and transitivity effects simultaneously.

Apart from modeling how new edges connect in G_t by Eq. (3.1), we do not explicitly model probability distributions of any other aspects of the growth process. Suppose that the joint distribution of $v(t)$ and $m(t)$ is governed by a certain parameter vector $\boldsymbol{\theta}_t$ ($t \geq 1$), and the distribution of G_0 is governed by another parameter vector $\boldsymbol{\theta}_{\text{init}}$. Note that we do not make any assumptions on forms of these two distributions, essentially allowing a great degree of freedom for them. However, we need to make one standard assumption, which is virtually employed in all network models, that the parameter vectors $\boldsymbol{\theta}_t$ ($t \geq 1$) is independent of A_d and B_b given $v(t)$, $m(t)$, and G_{t-1}; and $\boldsymbol{\theta}_{\text{init}}$ is independent of A_d and B_b. Under this assumption, the probability of the observed data $G_0, \cdots, G_T$ can be expressed as follows:

$$\begin{aligned} P(G_0, \cdots, G_T) &= \prod_{t=1}^{T} P(G_t|G_{t-1})P(G_0) \\ &= \prod_{t=1}^{T} P(G_t|m(t), v(t), G_{t-1}, A_d, B_b) P(m(t), v(t)|G_{t-1}, \boldsymbol{\theta}_t) \\ &\quad \cdot P(G_0|\boldsymbol{\theta}_{\text{init}}). \end{aligned} \tag{3.2}$$

As revealed in Section 3.3.1, Eq. (3.2) enables a partial likelihood approach, whereby one can ignore $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{\text{init}}$ in estimating A_d and B_b. Next, we discuss the relation between the model in Eq. (3.1) and those in the literature.

3.2.2 Related Models

As explained previously, existing non-parametric models either have a non-parametric A_d function (Pham et al., 2015) or a non-parametric B_b function (Newman, 2001a), and the probabilities of PA and transitivity for a new edge emerging between nodes i and j at time step t are respectively as follows,

$$P_{ij}(t) \propto A_{d_i(t)} A_{d_j(t)}, \tag{3.3}$$

$$P_{ij}(t) \propto B_{b_{ij}(t)}. \tag{3.4}$$

while estimating A_d function and B_b function in isolation can lead to bias, Eq. (3.1) is the first to combine both non-parametric functions to deal with it. It includes several well-known complex network models as special cases, such as the Barabási-Albert model (Barabási & Albert, 1999) or the Erdös-Rényi with growth model (Callaway et al., 2001).

The well-known stochastic actor-based model (Ripley et al., 2018; Snijders, 2001, 2017) has been employed in studies on scientific co-authorship networks (Ferligoj et al., 2015; Kronegger et al., 2012; Zinilli, 2016). The actor-based model is a type of generalized linear models with and it is not clear how to convert the PA and transitivity functions in our probabilistic setting into those in the setting of the stochastic actor-based model, because the two models are defined differently. However, for a fair comparison, we can set the joint parametric model of PA and transitivity in the framework of GT model for experiments, as well as an actor-based model. In this case, the functional forms are explicitly given to the growth functions A_d and B_b in Eq. (3.1). For example, when we assume log-linear forms $A_d = (d+1)^{\alpha}, B_b = (b+1)^{\beta}$, the joint parametric model we use in this part can be expressed as

$$P_{ij}(t) \propto (d_i+1)^{\alpha}(d_j+1)^{\alpha}(b_{ij}+1)^{\beta}. \tag{3.5}$$

Note that we here consider only undirected graphs in the growth of networks, the PA and transitivity phenomena are modeled in a parametric manner in the undirected version of the stochastic actor-based model (Snijders, 2017).

One key assumption of the model in Eq. (3.1) is that A_d and B_b are not dependent on t; that is, on a practical level, they change little throughout the growth process. While this time-invariance assumption is standard and is employed in all of the network models mentioned previously, there is an increasing body of models departing therefrom. A time-varying A_d has been discussed in the context of citation networks (Csárdi et al., 2007; Medo et al., 2011; Wang et al., 2008), while different parametric forms for such A_d have been studied by Medo (Medo, 2014). More recently, the R package `tergm` (Krivitsky & Handcock, 2019) has enabled the estimation of time-varying parametric PA and transitivity functions. However, no existing work has employed time-varying and non-parametric modeling simultaneously, presumably because a huge amount of data is probably required in such a model. If time has little importance in the system, time-varying modeling is not necessary. In Section 4.1.4, we demonstrate that the time-invariance assumption indeed holds in all of the real-world networks analyzed in this study.

Finally, we note that, while we model the PA phenomenon directly in this chapter, an alternative approach is to let PA emerging as a consequence of some deeper mechanisms. For example, it has been shown that

PA can emerge from either reference-copying mechanisms in citation networks (Golosovsky & Solomon, 2017; Simkin & Roychowdhury, 2007) or some optimization frameworks (D'Souza et al., 2007).

3.3 Methodology of Estimation

In this section, we discuss maximum partial likelihood estimation for the model and derive an efficient the Minorize-Maximize algorithms for the log-likelihood function. We also provide two simulated examples to demonstrate our method. We conclude the section by presenting a systematic simulation to investigate the performance of the proposed method.

3.3.1 Maximum Partial Likelihood Estimation

Here, we describe how to estimate the parameters of the model in Eq. (3.1). We denote $X = \{G_0, G_1, \cdots, G_T\}$ as the observed data, and let $\boldsymbol{A} = [A_0, A_1, \ldots, A_{d_{max}}]$ with $A_d > 0$ being the PA function and $\boldsymbol{B} = [B_0, B_1, \ldots, B_{b_{max}}]$ with $B_b > 0$ being the transitivity function. In this case, d_{max} is the maximum degree and b_{max} is the maximum number of common neighbors between a pair of nodes. Given X, our goal is to estimate $\boldsymbol{A}$ and $\boldsymbol{B}$ without assuming any specific functional forms, which is an approach we refer to as "non-parametric".

We can rewrite Eq. (3.2) using the new notations:

$$P(X) = \prod_{t=1}^{T} P(G_t|m(t), v(t), G_{t-1}, \boldsymbol{A}, \boldsymbol{B}) P(m(t), v(t)|G_{t-1}, \boldsymbol{\theta}_t) P(G_0|\boldsymbol{\theta}_{\text{init}}).$$

Taking the logarithm of both sides of the previous equation provides us with:

$$\begin{aligned} L(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\theta}_{\text{init}},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_T|X) = &\underbrace{\sum_{t=1}^{T}\log P(G_t|m(t),v(t),G_{t-1},\boldsymbol{A},\boldsymbol{B})}_{L(\boldsymbol{A},\boldsymbol{B}|X)} \\ &+\underbrace{\sum_{t=1}^{T}\log P(m(t),v(t)|G_{t-1},\boldsymbol{\theta}_t)}_{L(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_T|X)} \\ &+\underbrace{\log P(G_0|\boldsymbol{\theta}_{\text{init}})}_{L(\boldsymbol{\theta}_{\text{init}}|X)}, \end{aligned} \tag{3.6}$$

where L denotes the log-likelihood function. Maximization of the total log-likelihood $L(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\theta}_{\text{init}},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_T|X)$ leads to maximization of the three terms on the right-hand side, and the parameters of our interest $\boldsymbol{A},\boldsymbol{B}$ are only involved in the first term. This allows us to ignore the nuisance parameters $\boldsymbol{\theta}_{\text{init}},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_T$ in estimating A_d and B_b, and we only need to maximize $L(\boldsymbol{A},\boldsymbol{B}|X)$. For simplicity, when calculating $P(G_t|m(t),v(t),G_{t-1},\boldsymbol{A},\boldsymbol{B})$, we ignore the new edges that attach to new nodes that emerge at time step t. Starting from Eq. (3.1), with some calculations, we arrive at

$$\begin{aligned} L(\boldsymbol{A},\boldsymbol{B}|X) = &\sum_{t=1}^{T}\sum_{d_1=0}^{d_{max}}\sum_{d_2=d_1}^{d_{max}}\sum_{b=0}^{b_{max}} m_{d_1,d_2,b}(t)\log A_{d_1}A_{d_2}B_b - \\ &\sum_{t=1}^{T} m(t)\log\left(\sum_{d_1=0}^{d_{max}}\sum_{d_2=d_1}^{d_{max}}\sum_{b=0}^{b_{max}} n_{d_1,d_2,b}(t)A_{d_1}A_{d_2}B_b\right), \end{aligned} \tag{3.7}$$

where $n_{d_1,d_2,b}(t)$ is the number of node pairs (i,j) satisfying $(d_i(t),d_j(t),b_{ij}(t)) = (d_1,d_2,b)$ with $d_1 \leq d_2$ at time step t, and $m_{d_1,d_2,b}(t)$ is the number of new edges between such node pairs. The number of new edges at time step t can then be expressed as $m(t) = \sum_{d_1=0}^{d_{max}}\sum_{d_2=d_1}^{d_{max}}\sum_{b=0}^{b_{max}} m_{d_1,d_2,b}(t)$.

Although analytically maximizing $L(\boldsymbol{A},\boldsymbol{B}|X)$ is intractable, we can derive an efficient MM algorithm that iteratively updates $\boldsymbol{A}$ and $\boldsymbol{B}$. Its derivation is presented in the next Section. We also write the final results of the algorithm as $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$, which are estimates of $\boldsymbol{A}$ and $\boldsymbol{B}$.

3.3.2 An MM algorithm for Estimating the Non-parametric PA and Transitivity Functions

We derive an instance of the Minorize-Maximize algorithms (Hunter & Lange, 2000) for maximizing the partial log-likelihood function $l(\boldsymbol{A}, \boldsymbol{B})$ in Eq. (3.7). Denote $A_d^{(q)}$ the value of A_d at iteration q ($q \geq 0$), and $\boldsymbol{A}^{(q)} = [A_0^{(q)}, A_1^{(q)}, \ldots, A_{d_{max}}^{(q)}]$ the value of $\boldsymbol{A}$ at that iteration. Define $B_b^{(q)}$ and $\boldsymbol{B}^{(q)}$ in the same way. Starting from some initial values $(\boldsymbol{A}^0, \boldsymbol{B}^0)$ at iteration $q = 0$, we want to compute $(\boldsymbol{A}^{(q+1)}, \boldsymbol{B}^{(q+1)})$ from $(\boldsymbol{A}^{(q)}, \boldsymbol{B}^{(q)})$. In MM algorithms, such update formulas can be derived by first determining a surrogate function $Q(\boldsymbol{A}, \boldsymbol{B})$ satisfying $l(\boldsymbol{A}, \boldsymbol{B}) \geq Q(\boldsymbol{A}, \boldsymbol{B}), \forall \boldsymbol{A}, \boldsymbol{B}$ and $l(\boldsymbol{A}^{(q)}, \boldsymbol{B}^{(q)}) = Q(\boldsymbol{A}^{(q)}, \boldsymbol{B}^{(q)})$, and then maximizing the surrogate function. It can be proved that, if $(\boldsymbol{A}^{q+1}, \boldsymbol{B}^{q+1})$ maximizes $Q(\boldsymbol{A}, \boldsymbol{B})$, then $l(\boldsymbol{A}^{(q+1)}, \boldsymbol{B}^{(q+1)}) \geq l(\boldsymbol{A}^{(q)}, \boldsymbol{B}^{(q)})$; that is, the objective function is increased monotonically per iteration. Since several surrogate functions satisfying the conditions may exist, the main indicator for evaluating a particular $Q(\boldsymbol{A}, \boldsymbol{B})$ is the ease of its maximization.

Based on previous works (Pham et al., 2015, 2016), the following function is a surrogate function of l:

$$\begin{aligned} Q'(\boldsymbol{A}, \boldsymbol{B}) &= \sum_{t=0}^{T} \sum_{i=0}^{D} \sum_{j=i}^{D} \sum_{l=0}^{B} m_{i,j,l}(t) \log A_i A_j B_l \\ &- \sum_{t=0}^{T} m(t) \log \left(\sum_{i=0}^{D} \sum_{j=i}^{D} \sum_{l=0}^{B} n_{i,j,l}(t) A_i^{(q)} A_j^{(q)} B_l^{(q)} \right) \\ &- \sum_{t=0}^{T} m(t) \frac{\sum_{i=0}^{D} \sum_{j=i}^{D} \sum_{l=0}^{B} n_{i,j,l}(t) A_i A_j B_l}{\sum_{i=0}^{D} \sum_{j=i}^{D} \sum_{l=0}^{B} n_{i,j,l}(t) A_i^{(q)} A_j^{(q)} B_l^{(q)}} + \sum_{t=0}^{T} m(t), \end{aligned} \tag{3.8}$$

where $D := d_{max}$ and $B := b_{max}$.
The product $A_i A_j B_l$ in the numerator of the third term on the right hand side of Eq. (3.8) prevents parallel updating of $\boldsymbol{A}$ and $\boldsymbol{B}$. One approach to handle this product is to apply the AM-GM inequality (Hunter & Lange,

2004):

$$
\begin{aligned}
-A_iA_jB_l &\geq -\frac{1}{2}\left(\frac{A_j^{(q)}}{A_i^{(q)}}A_i^2 + \frac{A_i^{(q)}}{A_j^{(q)}}A_j^2\right)B_l \\
&\geq -\frac{1}{4}\left(\frac{A_j^{(q)}B_l^{(q)}}{\left(A_i^{(q)}\right)^3}A_i^4 + \frac{A_i^{(q)}B_l^{(q)}}{\left(A_j^{(q)}\right)^3}A_j^4\right) - \frac{1}{2}\frac{A_i^{(q)}A_j^{(q)}}{B_l^{(q)}}B_l^2.
\end{aligned}
$$

Inserting this inequality into Eq. (3.8), we obtain the final surrogate function:

$$
\begin{aligned}
Q(\boldsymbol{A},\boldsymbol{B}) = &\sum_{t=0}^{T}\sum_{i=0}^{D}\sum_{j=i}^{D}\sum_{l=0}^{B} m_{i,j,l}(t)\log A_iA_jB_l \\
&- \sum_{t=0}^{T} m(t)\log\left(\sum_{i=0}^{D}\sum_{j=i}^{D}\sum_{l=0}^{B} n_{i,j,l}(t)A_i^{(q)}A_j^{(q)}B_l^{(q)}\right) \\
&- \sum_{t=0}^{T} m(t)\frac{\sum_{i=0}^{D}\sum_{j=i}^{D}\sum_{l=0}^{B} n_{i,j,l}(t)\left(\frac{1}{4}\left(\frac{A_j^{(q)}B_l^{(q)}}{\left(A_i^{(q)}\right)^3}A_i^4 + \frac{A_i^{(q)}B_l^{(q)}}{\left(A_j^{(q)}\right)^3}A_j^4\right) + \frac{1}{2}\frac{A_i^{(q)}A_j^{(q)}}{B_l^{(q)}}B_l^2\right)}{\sum_{i=0}^{D}\sum_{j=i}^{D}\sum_{l=0}^{B} n_{i,j,l}(t)A_i^{(q)}A_j^{(q)}B_l^{(q)}} \\
&+ \sum_{t=0}^{T} m(t).
\end{aligned}
$$

Solving $\frac{\partial Q(\boldsymbol{A},\boldsymbol{B})}{\partial A_d} = 0$ and $\frac{\partial Q(\boldsymbol{A},\boldsymbol{B})}{\partial B_b} = 0$, we obtain the following closed-form formulas:

$$
A_d^{(q+1)} = \sqrt[4]{\frac{\sum_{t=0}^{T}\sum_{i=0}^{D} m_{i,d,\cdot}(t) + \sum_{t=0}^{T}\sum_{j=d}^{D} m_{d,j,\cdot}(t)}{\sum_{t=0}^{T} m(t)\frac{\sum_{j=d+1}^{D}\sum_{l=0}^{B} n_{d,j,l}(t)\frac{A_j^{(q)}B_l^{(q)}}{\left(A_d^{(q)}\right)^3} + \sum_{i=0}^{d-1}\sum_{l}^{B} n_{i,d,l}(t)\frac{A_i^{(q)}B_l^{(q)}}{\left(A_d^{(q)}\right)^3} + \sum_{l=0}^{B} n_{d,d,l}(t)\left(\frac{A_d^{(q)}B_l^{(q)}}{\left(A_d^{(q)}\right)^3} + \frac{A_d^{(q)}B_l^{(q)}}{\left(A_d^{(q)}\right)^3}\right)}{\sum_{i=0}^{D}\sum_{j=i}^{D}\sum_{l=0}^{B} n_{i,j,l}(t)A_i^{(q)}A_j^{(q)}B_l^{(q)}}}},
$$

$$
B_b^{(q+1)} = \sqrt{\frac{\sum_{t=0}^{T} m_{\cdot,\cdot,b}(t)}{\sum_{t=0}^{T} m(t)\frac{\sum_{i=0}^{D}\sum_{j=i}^{D} n_{i,j,b}(t)\frac{A_i^{(q)}A_j^{(q)}}{B_b^{(q)}}}{\sum_{i=0}^{D}\sum_{j=i}^{D}\sum_{l=0}^{B} n_{i,j,l}(t)A_i^{(q)}A_j^{(q)}B_l^{(q)}}}},
$$

where $m_{i,d,\cdot}(t) := \sum_{l=0}^{B} m_{i,d,l}(t)$ and $m_{\cdot,\cdot,b} := \sum_{i=0}^{D}\sum_{j=i}^{D} m_{i,j,b}(t)$.

Based on these formulas, at each iteration $\boldsymbol{A}^{(q+1)}$ and $\boldsymbol{B}^{(q+1)}$ can be computed in parallel without solving any additional optimization problems. This enables the method to work with large datasets. The objective function value $l(\boldsymbol{A}^{(q+1)}, \boldsymbol{B}^{(q+1)})$, as explained previously, is guaranteed to be increasing in q.

3.3.3 Illustrative Examples

We demonstrate the effectiveness of our method using two examples. In the first example, we simulate a network using Eq. (3.1) with

$$A_d = 3(\log(\max d, 1))^2 + 1,$$
$$B_b = 3(\log(\max b, 1))^2 + 1.$$

This functional form, which deviates substantially from the power-law form, has been used to demonstrate the processes of non-parametric PA estimation methods (Pham et al., 2015). The initial number of nodes is 300. At each time step, first $m(t) = 100$ new edges are added between existing nodes, and then five new node are added. The total number of time steps T is 199. In the second example, we use real-world functions as the true functions for a more realistic comparison. We first estimate A_d and B_b by applying our proposed method to a real-world co-authorship network between authors in statistics journals (see Section 4.1), and then use these parameter values to simulate a network based on Eq. (3.1). In the process, we maintain the initial condition, as well as the number of new nodes and new edges at each time step, exactly the same as those observed in the real-world network.

We apply three estimation methods to each simulated network. The first is our proposed method, which jointly estimates the non-parametric functions A_d and B_b. The second is a joint parametric method, which jointly estimates PA and transitivity using the simplistic functional forms $A_d = (d+1)^\alpha$ and $B_b = (b+1)^\beta$. This parametric formation is used extensively in various PA and transitivity estimation methods (Gómez et al., 2011; Massen & Jonathan, 2007). The third method ignores the joint existence of PA and transitivity. It consists of two sub-methods: the first is a non-parametric method for estimating the PA function in isolation (Pham et al., 2015), and the second is a maximum likelihood version of a non-parametric method for estimating the transitivity function in isolation (Newman, 2001a).

The results are presented in Fig. 3.1. In both examples, while the joint parametric method somehow succeeds in obtaining the general trends of A_d and B_b, it fails to capture the deviations from the power-law form in the two functions. The non-parametric, in-isolation method grossly overestimates both the PA and transitivity mechanisms, owing to its complete disregard for their joint existence. The proposed method performs reasonably well, succeeding in capturing the PA and transitivity functions in fine details.

3.3.4 Simulation Study

We performed a systematic simulation to evaluate the effectiveness of the proposed method in estimating A_d and B_b. We selected $A_d = (d+1)^\alpha$ and $B_b = (b+1)^\beta$ as the true functions. This power-law functional form has been used in previous simulation studies on PA estimation methods (Pham et al., 2015, 2020). We considered five values (0, 0.5, 1, 1.5, and 2) for the exponent α and seven values (0, 0.5, 1, 1.5, 2, 2.5, and 3) for the exponent β. These are the ranges of α and β observed in Section 4.1.2. We simulated 10 networks for each combination of α and β. In each network, the initial number of nodes is 100, and 20 new edges and one new node are added at each time step.

For each simulated network, we first estimated A_d and B_b, as described in the previous section, and then fitted $(d+1)^\alpha$ and $(b+1)^\beta$ to the estimation results to determine the estimates of α and β. That is, we indirectly measured how well A_d and B_b are estimated by analyzing the estimation of α and β: if the estimates of α and β are good, the estimations of A_d and B_b are probably also successful.

Figure 3.2 presents the true and estimated values of α and β. The proposed method successfully recovers α and β in all combinations. This implies that the estimation of A_d and B_b is good. When β and α are small, the

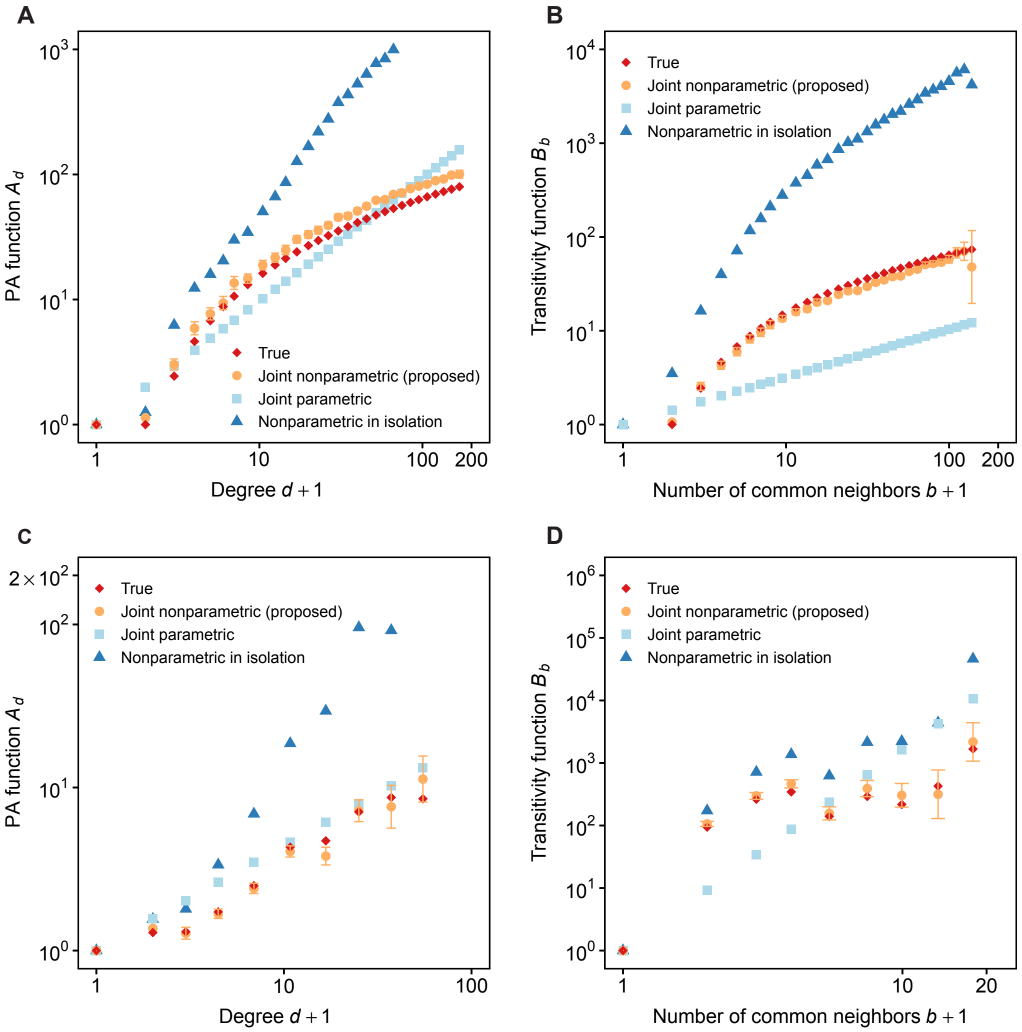


FIGURE 3.1: Proposed method compared to conventional methods for estimating PA and transitivity functions in two simulated networks. **A**, **B**: estimated PA and transitivity functions from a simulated network with $A_d = 3(\log(\max d, 1))^2 + 1$ and $B_b = 3(\log(\max b, 1))^2 + 1$. **C**, **D**: estimated PA and transitivity functions from a simulated network in which true A_d and B_b are A_d and B_b estimated from a real-world co-authorship network between authors in statistics journals. The interval at each point of the proposed method represents the standard deviation obtained as a by-product of the maximum likelihood estimation. In both networks, the proposed method successfully captures the fine details of the PA and transitivity.

standard errors of β are comparatively large. This appears to be caused by the instability of the estimation of B_b for large b; see Section 4.1.5.

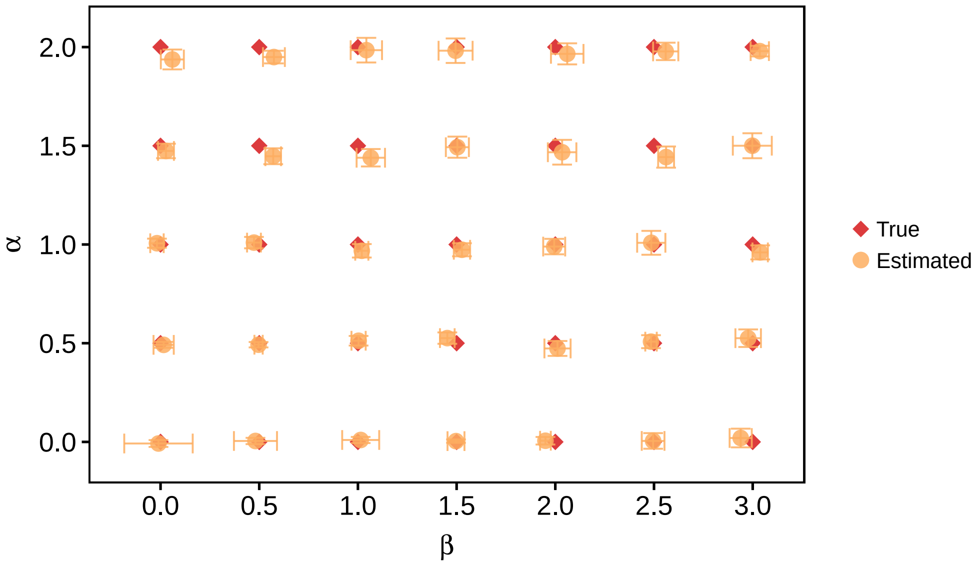


FIGURE 3.2: True and estimated exponents α and β from the power-law forms $A_d = (d+1)^{\alpha}$ and $B_b = (b+1)^{\beta}$. The exponents are estimated by a two-step procedure: first, A_d and B_b are estimated jointly by the proposed method, and then $(d+1)^{\alpha}$ and $(b+1)^{\beta}$ are fitted to the estimated results by least square using the actual values. Each estimated point is the mean of the results of 10 simulations, with the error bars displaying the two standard errors of the mean.

3.4 Quantifying Contribution Amounts of PA and Transitivity

Our model provides a simple answer to a previously un-raised yet fascinating question: how can one compare the amount of contributions of PA and transitivity in the growth process of a network? To the best of our knowledge, no attempt has been made to quantify the amount of contributions of different network growth mechanisms. To answer this question, one must determine a meaningful method to define the amount of contributions, so that they are computable and comparable. We achieve this by considering the dynamic process expressed in Eq. (3.1). This probabilistic dynamic process suggests that the variability of the PA/transitivity values in the set of node pairs is a sensible measure for the amount of contributions of PA/transitivity.

We define the amount of contributions of PA and transitivity at time step t, which are denoted as $s_{\text{PA}}(t)$ and $s_{\text{trans}}(t)$, respectively. Taking the logarithm of both sides of Eq. (3.1), we obtain:

$$\log_2 P_{ij}(t) = \log_2[A_{d_i(t)} A_{d_j(t)}] + \log_2 B_{b_{ij}(t)} - C(t), \tag{3.9}$$

where $C(t) = \log_2 \sum_{i,j} A_{d_i(t)} A_{d_j(t)} B_{b_{ij}(t)}$ is the logarithm of the normalizing constant at time step t, and it is independent of i and j. Equation (3.9) implies that, considering a node pair (i,j) locally, PA and transitivity contribute to $\log_2 P_{ij}(t)$ according to the amounts of $\log_2[A_{d_i(t)} A_{d_j(t)}]$ and $\log_2 B_{b_{ij}(t)}$, respectively; the amount of contributions is measured by $\log_2$ fold changes.

However, the relative sizes of all $\log_2[A_{d_i(t)} A_{d_j(t)}]$ and $\log_2 B_{b_{ij}(t)}$ at that time step t are ultimately what are important. For example, consider the case when $A_d = 1, \forall d$. In this case, the value of $\log_2[A_{d_i(t)} A_{d_j(t)}]$ will be the same for all node pairs, and thus, the PA plays no role in determining which pair receives a new edge. By considering the case when $B_b = 1, \forall b$, it can be observed that the same reasoning should apply to $\log_2 B_{ij}(t)$.

This observation prompts us to define $s_{\text{PA}}(t)$ and $s_{\text{trans}}(t)$ as the standard deviations of $\log_2[A_{d_i(t)} A_{d_j(t)}]$ and $\log_2 B_{b_{ij}(t)}$, respectively, when (i,j) is sampled based on Eq. (3.1). Let $U(t)$ be the set formed by all node pairs (i,j) existing at time step $t-1$; that is, all the combinations of two nodes in G_{t-1}. The probability $P_{ij}(t)$ in Eq. (3.1) can be explicitly expressed as:

$$P_{ij}(t) = \frac{A_{d_i(t)} A_{d_j(t)} B_{b_{ij}(t)}}{\sum_{(i,j)\in U(t)} A_{d_i(t)} A_{d_j(t)} B_{b_{ij}(t)}}.$$

The aforementioned standard deviations can be calculated as follows:

$$s_{\text{PA}}(t) := \left(\sum_{(i,j)\in U(t)} P_{ij}(t) \left(\log_2[A_{d_i(t)} A_{d_j(t)}] - E_{\text{PA}}(t) \right)^2 \right)^{1/2}, \quad (3.10)$$

$$s_{\text{trans}}(t) := \left(\sum_{(i,j)\in U(t)} P_{ij}(t) \left(\log_2 B_{b_{ij}(t)} - E_{\text{trans}}(t) \right)^2 \right)^{1/2}, \quad (3.11)$$

in which $E_{\text{PA}}(t) := \sum_{(i,j)\in U(t)} P_{ij}(t) \log_2[A_{d_i(t)} A_{d_j(t)}]$, and $E_{\text{trans}}(t) := \sum_{(i,j)\in U(t)} P_{ij}(t) \log_2 B_{b_{ij}(t)}$. Although A_d and B_b are only defined up to multiplicative constants, the standard deviations of $\log_2[A_{d_i(t)} A_{d_j(t)}]$ and $\log_2 B_{b_{ij}(t)}$ are invariant to the constant factors in A_d and B_b, and thus, $s_{\text{PA}}(t)$ and $s_{\text{trans}}(t)$ are well defined. The use of base-2 logarithms enables us to interpret $s_{\text{PA}}(t)$ and $s_{\text{trans}}(t)$ as $\log_2$ fold changes; a contribution value of s indicates a change in the probability of 2^s times in Eq. (3.1). We also note that, although A_d and B_b are assumed to be time invariant, $d_i(t)$, $b_{ij}(t)$, and $P_{ij}(t)$ change over time, thereby leading to the dynamic nature of $s_{\text{PA}}(t)$ and $s_{\text{trans}}(t)$.

In real-world situations, the true values $\boldsymbol{A}$ and $\boldsymbol{B}$ are not available to us, but only their estimates $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$. We insert these estimates into Eqs. (3.10) and (3.11) to obtain $\hat{s}_{\text{PA}}(t)$ and $\hat{s}_{\text{trans}}(t)$, which are estimates of $s_{\text{PA}}(t)$ and $s_{\text{trans}}(t)$, respectively.

The requirement that (i,j) is sampled from Eq. (1) is necessary to reflect the probabilistic dynamic process accurately, and leads to the following interpretation of $s_{\text{PA}}(t)$ and $s_{\text{trans}}(t)$. Assume that, at some time step t, we observe $m(t) \geq 2$ new edges, the end points of which are $(i_1, j_1), \cdots, (i_{m(t)}, j_{m(t)})$. Consider the sample standard deviation of $\log_2(B_{b_{i_l j_l}(t)})$ for $l = 1, \cdots, m(t)$, which is defined as

$$h_{\text{trans}}(t) := \left(\frac{1}{m(t)-1} \sum_{l=1}^{m(t)} \log_2(B_{b_{i_l j_l}(t)})^2 - \frac{1}{m(t)(m(t)-1)} \left(\sum_{l=1}^{m(t)} \log_2(B_{b_{i_l j_l}(t)}) \right)^2 \right)^{1/2}.$$

Similarly, define $h_{\text{PA}}(t)$ as the sample standard deviation of $\log_2(A_{d_{i_l}(t)}A_{d_{j_l}(t)})$ for $l = 1, \cdots, m(t)$. Standard calculations then yield $s_{\text{trans}}(t)^2 = \mathbb{E}\, h_{\text{trans}}(t)^2$ and $s_{\text{PA}}(t)^2 = \mathbb{E}\, h_{\text{PA}}(t)^2$. By inserting the estimates $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$, we can regard $\hat{s}_{\text{PA}}(t)$ and $\hat{s}_{\text{trans}}(t)$ as the estimates of the expectations of the sample standard deviations of the PA and transitivity values observed at the end points of the new edges at time step t. As noted in Section 4.1.3, this interpretation also provides a means to visualize how effectively the model fits an observed network.

Finally, we note that this quantification approach is not limited to PA and transitivity. Given a growth formula in which all growth mechanisms are combined in a multiplicative manner; for example, as in Eq. (3.1), the standard deviation of the logarithmic values of each growth mechanism can be used as a measure of the contribution of that mechanism.

Chapter 4

Real Data Analysis with FoFaF

Our contributions, descriptions and figures in this chapter are based on our papers (Inoue et al., 2020b).

4.1 Experiments

4.1.1 Real-world Datasets

We apply our proposed method to two scientific co-authorship networks: SMJ (Ronda-Pupo & Pham, 2018) and STA (Ji & Jin, 2016), in which the nodes represent authors and the links represent co-authorship in papers. SMJ includes papers published in the *Strategic Management Journal*, which is considered as one of the top journals in strategy and management, from 1980 to 2017. STA includes papers in four statistics journals: the *Journal of the American Statistical Association*, the *Journal of the Royal Statistical Society (Series B)*, the *Annals of Statistics*, and *Biometrika*, from 2003 to 2012. These are generally considered as leading journals in statistics. New and repeated collaborations are pooled together in both networks. The time resolution is selected as one year in SMJ and six months in STA.

Table 4.1 presents the summary statistics for the two networks. The ratios $\Delta|V|/|V|$ and $\Delta|E|/|E|$ are both close to 1, indicating that each network grows from a relatively small initial network, compared to the size of the final network. As the number of new edges $\Delta|E|$ loosely corresponds to the amount of available data in our statistical model, STA has the larger amount of data. The clustering coefficients in both networks are rather high, but nevertheless fall within the normal range observed in real-world networks (Newman, 2001c). The clustering coefficient C of the final snapshot can be calculated as

$$C = \frac{1}{|V|} \sum_{i=1}^{|V|} C_i.$$

Note that the coefficient C_i of node i is defined as

$$C_i = \begin{cases} \frac{2\Delta_i}{d_i(d_i-1)} & (d_i \geq 2) \\ 0 & (d_i = 0,1) \end{cases}, \ \Delta_i = \sum_{j,l} x_{i,j} x_{j,l} x_{l,i},$$

where $x_{i,j} = 1$ indicates the presence of edges between i and j, whereas $x_{i,j} = 0$ indicates the absence of edges.

TABLE 4.1: Summary statistics for two scientific coauthorship networks. $|V|$ and $|E|$ are the total numbers of nodes and edges, respectively, in the final snapshot; T is the number of time steps; $\Delta|V|$ and $\Delta|E|$ are the increments of nodes and edges, respectively, after the initial snapshot; C is the clustering coefficient of the final snapshot; d_{max} is the maximum degree; and b_{max} is the maximum number of common neighbors.

Dataset	$\|V\|$	$\|E\|$	T	$\Delta\|V\|$	$\Delta\|E\|$	C	d_{max}	b_{max}
SMJ	2704	4131	23	1991	3538	0.378	34	15
STA	3607	6808	19	3261	6509	0.320	65	19

It is instructive to investigate more fine-grained statistics. Figures 4.1A and B present the distributions of the numbers of collaborators d in the final snapshots of SMJ and STA, respectively. As the degree distributions in the two datasets exhibit signs of heavy tails, we fit one of the most representative classes of heavy-tailed distributions, namely the power-law distribution $d^{-\gamma_{deg}}$, to these degree distributions by means of Clauset's procedure (Clauset et al., 2009). This procedure first selects the minimum degree d_{min} for which the power-law holds, and then uses a maximum likelihood approach to estimate the power-law exponent γ_{deg}. The estimated power-law exponents for the degree distributions in SMJ and

STA are 2.97 and 3.35, respectively. These values fall within the range of $2 < \gamma_{deg} < 4$, which is a commonly observed range for γ_{deg} in real-world networks (Clauset et al., 2009; Newman, 2005).

However, the situation with the distributions of b_{ij} is less clear. Figures 4.1C and D present the distributions of the numbers of node pairs with b common neighbors in the final snapshots of SMJ and STA, respectively. We also fit the power-law distribution $b^{-\gamma_{cn}}$ to these distributions by means of Clauset's procedure and find that the estimated power-law exponents for the distributions of b in SMJ and STA are 2.99 and 3.22, respectively. However, it appears that the power-law form is not a very good fit for these distributions. The right-most point of b_{ij} in SMJ is due to a single paper with 17 authors. The ranges of b in the two distributions seem to be too narrow to draw any definitive conclusion regarding the tails. To the best of our knowledge, no previous work has studied the distributions of b_{ij}, in either co-authorship networks or any other network types. Because the determination of the distributional form of b_{ij} is not our main goal, we leave this task as future work.

4.1.2 Non-power-law characteristics of PA and transitivity functions

By applying the proposed method to two datasets, we find that the estimated PA and transitivity functions exhibit non-power-law and complex trends (Fig. 4.2).

In both networks, the value of A_d increases on average in d, implying the existence of the PA phenomenon: when an author gains more collaborators, they are more likely to gain a new one. This is consistent with previous results in the literature, in which the phenomenon has been found in collaboration networks in diverse fields (Ferligoj et al., 2015; Kronegger et al., 2012; Milojević, 2010; Newman, 2001a).

The situation with the transitivity functions is more complex. In both SMJ and STA, there is a huge jump when b changes from 0 to 1: B_1/B_0 is approximately 60 in SMJ and almost 100 in STA. These jumps in the

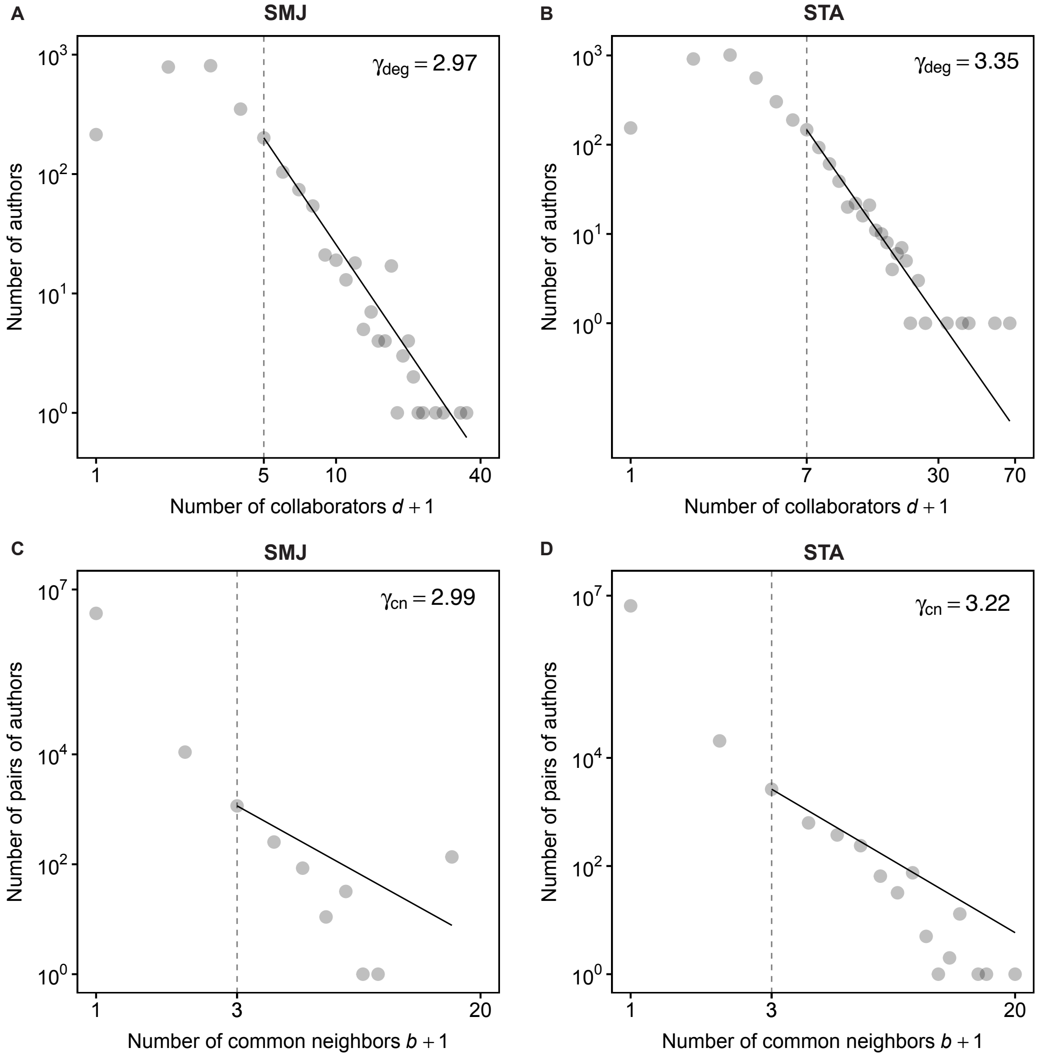


FIGURE 4.1: Distributions of d_i and b_{ij} in final snapshots of two networks. **A**, **B**: degree distributions in the final snapshots of SMJ and STA, respectively. These distributions both display typical heavy-tailed shapes. In each panel, the solid line indicates the fitted power-law distribution, whereas the dotted line indicates where the minimum degree d_{min} is set. **C**, **D**: distributions of the numbers of pairs with b common neighbors in final snapshots of SMJ and STA, respectively. In each panel, the solid line indicates the fitted power-law distribution, whereas the dotted line indicates where the minimum number of common neighbours b_{min} is set. In contrast to the degree distributions, the ranges of these distributions of b_{ij} are too narrow for any signs of heavy tails to emerge.

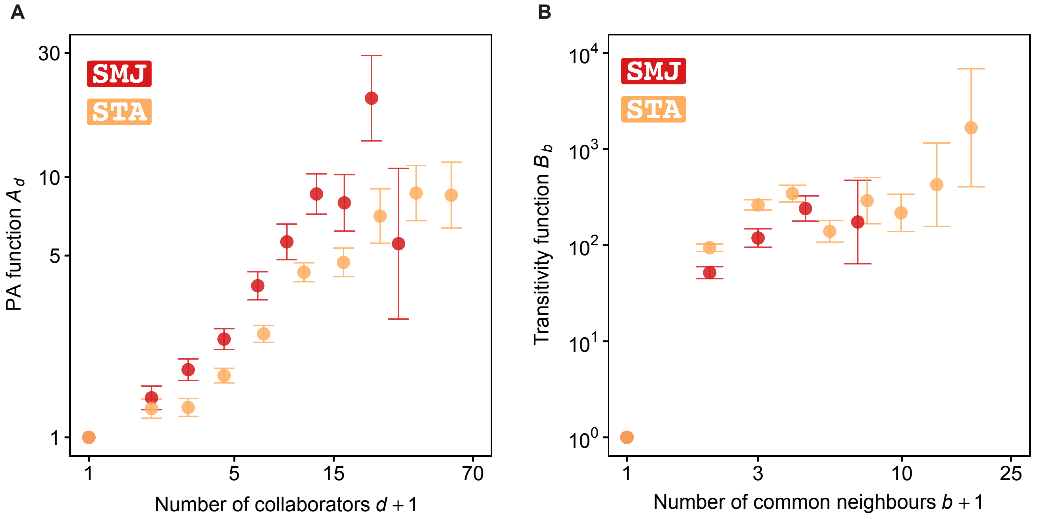


FIGURE 4.2: Non-parametric joint estimation of PA A_d and transitivity B_b functions in SMJ and STA. The vertical bar at each estimated value indicates a $\pm 2\sigma$ confidence interval. **A**: the estimated PA functions increase on average in both networks. This implies the existence of the PA phenomenon: a highly connected author is likely to gain more new collaborations than a lowly connected author. **B**: The transitivity functions substantially deviate from the power-law form. While B_b increases significantly when b changes from 0 to 1 in both networks, after this initial huge jump, B_b remains relatively horizontal in SMJ and only increases slightly in STA. The huge jump at $b = 1$ implies that co-authors of co-authors are at least fifty times more likely to become new co-authors, compared to the case when no mutual co-author exists.

B_b values have been observed in co-authorship networks (Milojević, 2010; Newman, 2001a). However, following this initial jump, B_b remains relatively horizontal in both SMJ and STA, before increasing slightly again in SMJ. This complex departure from the power-law form renders any statement regarding a universal transitivity effect moot. However, the value of B_b at every $b > 0$ is at least fifty times higher than B_0, which suggests that co-authors of co-authors appear to be at least fifty times more likely to become new co-authors, compared to the case when no mutual co-author exists.

A few other jumps in values of A_d and B_b can be seen in the regions of large d and b, respectively. However, the sizes of these jumps are comparable with the large confidence intervals in those regions, where the estimations of A_d and B_b are naturally unstable owing to the fact that there are comparatively fewer data points in those regions. Therefore, it appears to be safe to assume that those jumps convey few interesting insights on the PA and transitivity phenomena.

It is informative to supplement the non-parametric analysis with a parametric analysis, because the theoretical literature offers numerous insights into this context. In this case, we follow standard practice and fit the power-law functional forms $A_d = (d+1)^\alpha$ and $B_b = (b+1)^\beta$ (Jeong et al., 2003; Krapivsky & Redner, 2001; Pham et al., 2015). To determine the PA attachment exponent α and the transitivity attachment exponent β, we substitute these forms into Eq. (3.1), and numerically maximizes the resulting log-likelihood function with respect to α and β. Table 4.2 displays the estimated values of α and β.

TABLE 4.2: Estimated values of PA attachment exponent α and transitivity attachment exponent β in two networks. The values are estimated by the maximum partial likelihood estimation. The interval provided at each estimated value is two sigma.

Network	PA attachment exponent α	Transitivity attachment exponent β
SMJ	0.93 (±0.04)	2.50 (±0.07)
STA	0.84 (±0.03)	3.05 (±0.04)

The PA attachment exponent α in both networks is within the sub-linear region; that is, $0 < \alpha < 1$, which is a frequently observed range in real-world networks (Newman, 2001a; Pham et al., 2015; Ronda-Pupo & Pham, 2018). While this region has been demonstrated to result in a heavy-tailed degree distribution when only PA is at play (Krapivsky & Redner, 2001), no such theoretical result has been observed when PA co-exists with transitivity. However, it is not entirely unreasonable to expect that the sub-linear value of α is responsible for the observed heavy-tailed degree distributions in Figs. 4.1A and B.

The transitivity attachment exponents β are both greater than 1, indicating an exponentially faster growth rate of the transitivity function compared to the PA function. For example, this is evident in STA: while A_{10} is less than 10, B_{10} is already larger than 100. To the best of our knowledge, no theoretical results are available regarding the effect of β on the structure of a growing network, even for the supposedly simpler case when only transitivity exists.

Overall, the results in this section indicate the joint existence of the PA and transitivity phenomena in both networks. Our non-parametric approach reveals that a conventional power-law functional form in a parametric approach may not be optimal for describing the two phenomena. The power-law form fits the estimated A_d reasonably well up to the middle-degree part, but it cannot capture the deviations from the power-law form in the high-degree part. For B_b, the power-law form is even less suitable. We note that our findings hold even if we exclude the paper with 17 authors from SMJ; see Section 4.1.5. We hope that our non-parametric findings can offer hints on more suitable parametric forms for A_d and B_b.

4.1.3 Domination of Transitivity in Both Networks

After obtaining the estimates $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$, we can compute the amounts of contributions of the PA and transitivity phenomena in the growth process of each network by inserting these estimates into Eqs. (3.10) and (3.11). The estimated amounts of contributions $\hat{s}_{\text{PA}}(t)$ and $\hat{s}_{\text{trans}}(t)$ are indicated by the solid lines in Fig. 4.3.

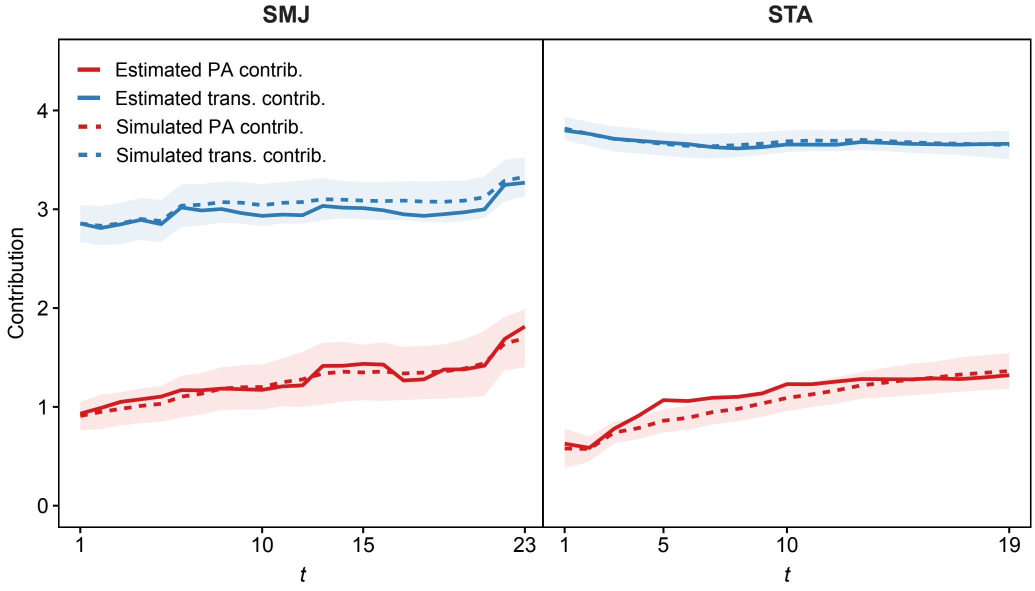


FIGURE 4.3: Estimated and simulated contributions of PA and transitivity at each time step in SMJ and STA. The contribution amount is measured by $\log_2$ fold changes in the model of Eq. (3.1). The solid lines indicate the estimated contributions $\hat{s}_{\text{PA}}(t)$ and $\hat{s}_{\text{trans}}(t)$, calculated by inserting the estimates $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ into Eqs. (3.10) and (3.11), respectively. Each dotted line is the average of the corresponding true contributions of 100 simulated networks, using $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ as the true functions. The band around each depicts the interval of ± two times the population standard deviation of the simulated contributions. The solid and dotted lines agree well with one another, suggesting that $\hat{s}_{\text{PA}}(t)$ and $\hat{s}_{\text{trans}}(t)$ are reliable.

In each network, $\hat{s}_{\text{trans}}(t)$ is greater than $\hat{s}_{\text{PA}}(t)$ for all t. Looking at the left panel in Fig. 4.3, the value of $\hat{s}_{\text{PA}}(t)$ (red solid line) is around 1 at $t = 1$ and it gradually increases up to around 2 at $t = 23$, where the contribution is measured by $\log_2$ fold changes. Thus the contribution of PA in SMJ network is around 2-fold change at the beginning and it becomes around 4-fold change at the end. The value of $\hat{s}_{\text{trans}}(t)$ (blue solid line) is around 3 with slight increase, meaning that the contribution of transitivity in SMJ network is around 8-fold change. Although both contributions are increasing, transitivity has larger impact than PA in the growing mechanism of the SMJ network. Looking at the right panel, we again observe that the contribution of transitivity is much larger than that of PA in the STA network.

It is worth questioning whether these tendencies also hold for the true values $s_{\text{PA}}(t)$ and $s_{\text{trans}}(t)$, or are simply artifacts that arise when we insert $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ into Eqs. (3.10) and (3.11). We demonstrate by means of simulations that, if the true $\boldsymbol{A}$ and $\boldsymbol{B}$ are close to the estimates $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$, $s_{\text{PA}}(t)$ and $s_{\text{trans}}(t)$ are similar to $\hat{s}_{\text{PA}}(t)$ and $\hat{s}_{\text{trans}}(t)$. For each real network, we simulate 100 networks based on Eq. (3.1), using the estimates $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ as true functions. We maintain all of the aspects of the growth process that are not governed by Eq. (3.1) the same as those observed in the real network. This includes using the observed initial graph, and observed numbers of new nodes and edges at each time step in the simulation. Because $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ are the true PA and transitivity functions for each simulated network, we can calculate the true contributions of the PA and transitivity in each simulated network using Eqs. (3.10) and (3.11), respectively. The behaviors of the simulated contributions are very similar to those of the estimated contributions $\hat{s}_{\text{PA}}(t)$ and $\hat{s}_{\text{trans}}(t)$, indicating that the latter are likely to be reliable.

As explained in Section 3.4, one can interpret the contributions $\hat{s}_{\text{PA}}(t)$ and $\hat{s}_{\text{trans}}(t)$ as estimates of the expectations of $\hat{h}_{\text{PA}}(t)$ and $\hat{h}_{\text{trans}}(t)$, which are the sample standard deviations of the PA and transitivity values at the end points of actually observed new edges at time step t. This is expressed as

$$\mathbb{E}\,\hat{h}_{\mathrm{PA}}(t) \approx \hat{s}_{\mathrm{PA}}(t);\ \mathbb{E}\,\hat{h}_{\mathrm{trans}}(t) \approx \hat{s}_{\mathrm{trans}}(t),$$

where the estimates $\hat{s}_{\mathrm{PA}}(t)$ and $\hat{s}_{\mathrm{trans}}(t)$ slightly overestimate the expectations, because

$$\mathbb{E}\,\hat{h}_{\mathrm{trans}}(t) \leq (\mathbb{E}\,\hat{h}_{\mathrm{trans}}(t)^2)^{1/2} \approx \hat{s}_{\mathrm{trans}}(t).$$

Figure 4.4 presents the observed values $\hat{h}_{\mathrm{PA}}(t)$ and $\hat{h}_{\mathrm{trans}}(t)$, the estimates $\hat{s}_{\mathrm{PA}}(t)$ and $\hat{s}_{\mathrm{trans}}(t)$ of their expectations, and the estimates of their standard deviations (see Section 4.1.5). The observed values generally fall within two standard deviations around the estimates of their expectations, implying that Eq. (3.1) is consistent with the observed data.

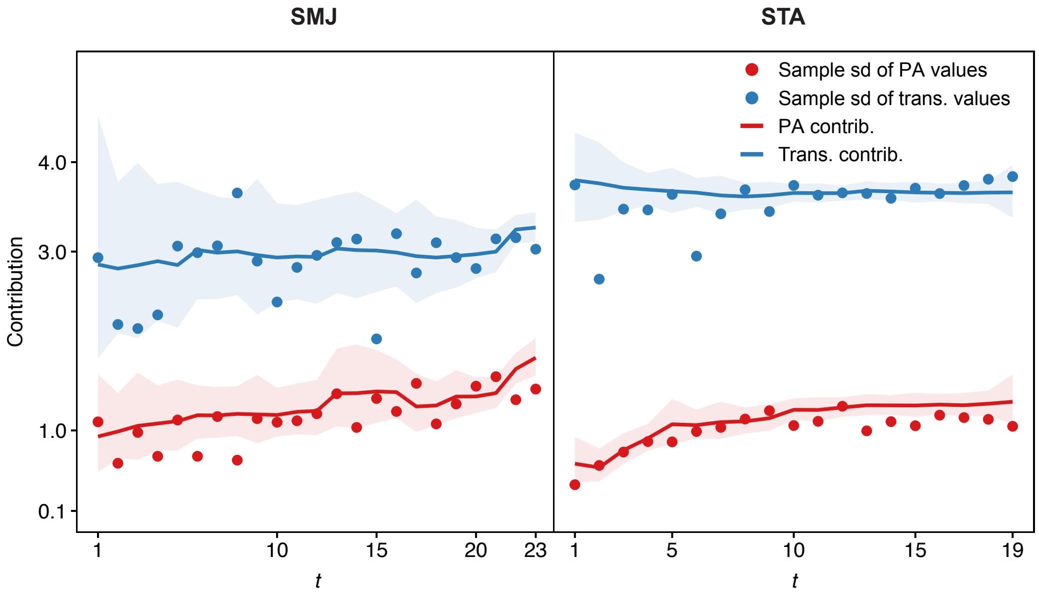


FIGURE 4.4: Sample standard deviations of PA and transitivity values at end points of actually observed new edges, $\hat{h}_{\mathrm{PA}}(t)$ and $\hat{h}_{\mathrm{trans}}(t)$, agree well with their estimated expectations, $\hat{s}_{\mathrm{PA}}(t)$ and $\hat{s}_{\mathrm{trans}}(t)$. This implies that the statistical model is consistent with the observed data. The band around $\hat{s}_{\mathrm{PA}}(t)$ depicts the interval of ± two standard deviations of $\hat{h}_{\mathrm{PA}}(t)$. The band around $\hat{s}_{\mathrm{trans}}(t)$ is similar.

Overall, the data indicate the governing role of transitivity in the growth processes of both networks: the differences in the transitivity values mainly determine where new collaborations are formed. This intuitive result is consistent with previous results, which found that common neighbors are more effective than PA for link prediction in co-authorship networks (Liben-Nowell & Kleinberg, 2007). If the PA dominated, the probability that you will work with a particular scientist in the future will not be significantly affected by whether or not you have worked with the scientists he knows. However, in light of the current result, they may need to be more selective, because the increase of common collaborators may offer greater advantages. Furthermore, the results suggest that, beyond mere comparison of the contribution of PA and transitivity, the following point should be taken into account in PA modeling. That is, the PA reflects nothing of the potential connection between the two. The more extensive the range of data observation, the greater the risk of selecting a person with a large degree, even though she/he is entirely out of scope.

4.1.4 Diagnosis: Time Invariance and Goodness of Fit

Finally, we consider two questions that are critical to our real-world data analysis. The first concerns the validity of the time-invariance assumption of A_d and B_b in two networks: in each network, is the observed data consistent with the assumption that A_d and B_b change little throughout the growth process? The second is whether Eq. (3.1) is a reasonably good model for the networks. Although Fig. 4.4 already hints at an affirmative answer to both questions, we examine each question in finer detail.

Time invariance of PA and transitivity functions

One means of answering the first question is to compare the A_d and B_b in Fig. 4.2 with the A_d and B_b estimated using only a certain portion of the growth process for many different portions. If they are similar, it is reasonable to believe that the observed data is consistent with the assumption that A_d and B_b change little throughout the growth process.

To this end, we create three new networks from each original network. The first new network ("First Half") contains only the first half of the growth process, thereby allowing A_d and B_b to be estimated in this portion. In the second new network ("Initial 0.5"), we set the initial time at the middle of the timeline, effectively enabling the estimation of A_d and B_b in the second half of the growth process. In the third new network ("Initial 0.75"), we set the initial time at the 3/4 point of the timeline. This network allows us to estimate A_d and B_b in the final quarter of the growth process. The estimated A_d and B_b in these three new networks then are compared with the A_d and B_b values obtained from the full growth process (Fig. 4.5). A visual inspection of Fig. 4.5 suggests that both the PA and transitivity functions appear to change little in the growth process of each network, which validates the time invariance assumption.

Goodness of fit

We use a simulation-based approach to investigate the goodness of fit of the model. For each real-world network, we reuse the simulation data used in Fig. 4.3, which consists of 100 simulated networks generated using the estimated A_d and B_b of that network as true functions. We compare several statistics of the simulated networks with the corresponding statistics of the real network. If Eq. (3.1) is a good fit, the observed and simulated statistics must be close. Similar simulation-based approaches have been proposed for inspecting the goodness of fit of exponential random graph models (Hunter et al., 2008) and stochastic actor-based models (Conaldi et al., 2012; Lospinoso, 2012).

To provide an overview, we investigate how well the model can replicate the observed degree curves. In Fig. 4.6, for each real-world network we select ten nodes uniformly at random from the top 1% of all nodes in terms of the number of new edges accumulated during the growth process. For each node, we then plot the evolution line of the observed and simulated degree values. When this line is closer to the equality line, the model captures the observed degree growth of that node better. Although the simulated degree at times tends to be lower than the observed degree for certain nodes, overall, the lines are reasonably close to the identity line, implying that the model captures the degree growth well.

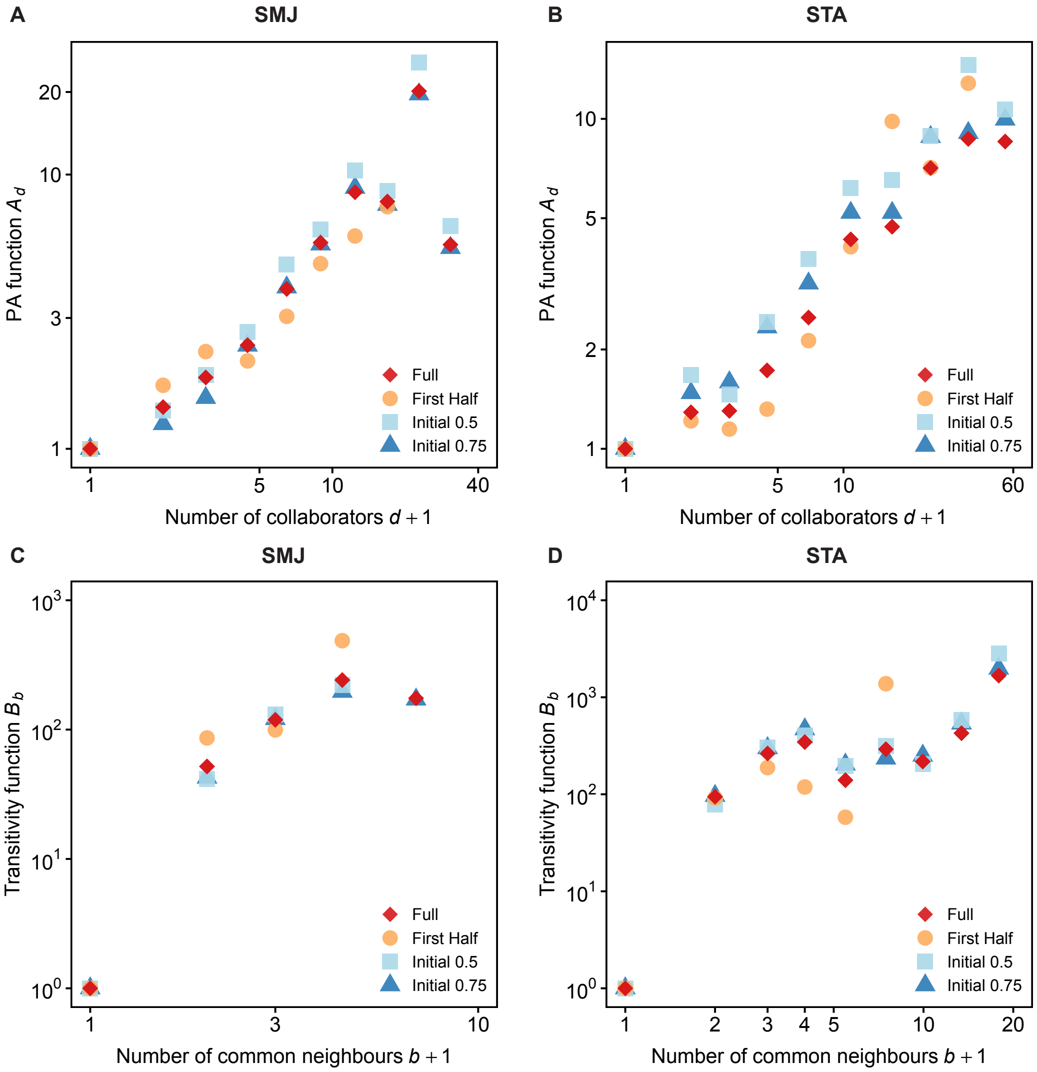


FIGURE 4.5: Time invariance of PA and transitivity functions. **A** and **C**: PA and transitivity functions of SMJ. **B** and **D**: PA and transitivity functions of STA. While "First Half" contains only the first half of the growth process, the initial time is set at the middle and at the 3/4 point of the timeline in "Initial 0.5" and "Initial 0.75", respectively. In each dataset, all four PA /transitivity functions agree well with one another, suggesting that the observed data is consistent with the assumption that the PA and transitivity functions change little during the growth process.

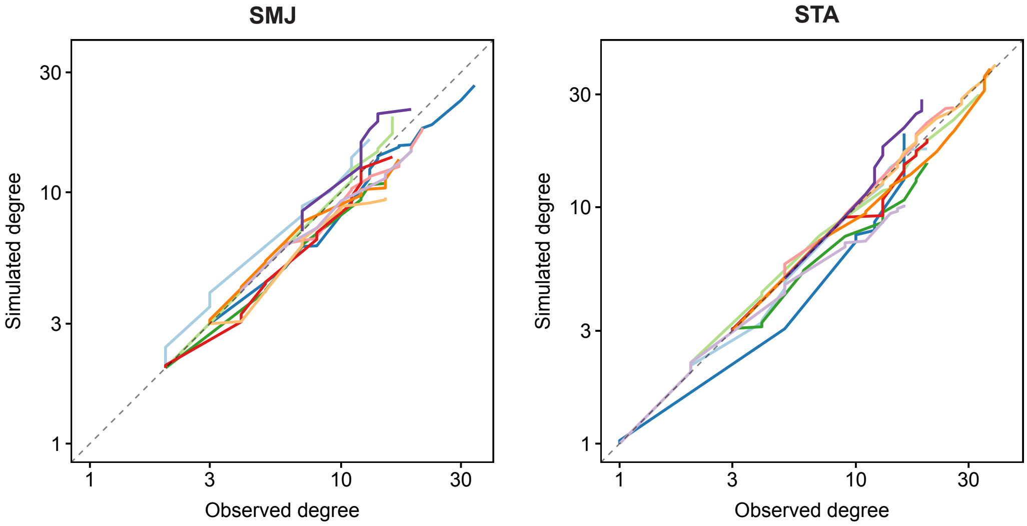


FIGURE 4.6: Pairs of observed and simulated degrees of several high-degree nodes in two networks. The simulation data are the same as those used in Fig. 4.3. In each network, ten nodes are selected uniformly at random from the top 1% of all nodes in terms of the number of new edges accumulated during the growth process. Each line represents the observed degrees and corresponding simulated values at each time step of a node. Each simulated value is averaged over 100 simulations. In each network, the pairs of observed and simulated degrees are reasonably close to the identity line, suggesting that the model fits the degree curves well.

For closer inspection, we analyze how well the model replicates the probability distribution of new edges during the growth process. In particular, consider sampling an edge e uniformly at random from the set of all new edges in the growth process. Suppose that e is between a node pair with degrees D_1 and D_2 ($D_1 \leq D_2$), and the number of their common neighbors is X. The relative frequency, or observed probability, that $D_1 = d_1, D_2 = d_2$, and $X = b$ is $p_{d_1,d_2,b} = \sum_t m_{d_1,d_2,b}(t) / \sum_{d_1=0}^{d_{max}} \sum_{d_2=d_1}^{d_{max}} \sum_{b=0}^{b_{max}} \sum_t m_{d_1,d_2,b}(t)$, in which $m_{d_1,d_2,b}(t)$ is the number of new edges emerging at time t between a node pair with degrees d_1 and d_2, and their number of common neighbors is b. Thus, the probability $p_{d_1,d_2,b}$ summarizes the information regarding the associations of d_1, d_2, and b at the end points of the new edges throughout the growth process.

Our joint estimation of the PA and transitivity is compared with two conventional approaches in which PA (Pham et al., 2015) and transitivity (Newman, 2001a) are estimated in isolation. For each approach, we first estimate the PA/transitivity function in isolation, and then use the estimated function to generate 100 networks to determine how well each existing method replicates $p_{d_1,d_2,b}$. To visualize this probability distribution, which is multi-dimensional, we slice it into many one-dimensional distributions by means of conditioning.

Firstly, we investigate

$$p_{d|b\in\mathcal{B}} := Pr(D_1 + D_2 = d | X \in \mathcal{B}) = \sum_{b\in\mathcal{B}} \sum_{d_1=0}^{d_{max}} p_{d_1,d-d_1,b} / \sum_{b\in\mathcal{B}} \sum_{d_1=0}^{d_{max}} \sum_{d_2=d_1}^{d_{max}} p_{d_1,d_2,b},$$

with the convention that $p_{d_1,d_2,b} = 0$ whenever $d_1 > d_2$ or $d_2 > d_{max}$. This is the conditional probability distribution of $D_1 + D_2$ given the event $X \in \mathcal{B}$. As we know from Fig. 4.1 that the number of node pairs with $b = 0$ or $b = 1$ is vastly greater than the remainder, we consider two probability distributions $p_{d|b\leq 1}$ and $p_{d|b\geq 2}$, and their cumulative probability distributions are illustrated in Fig. 4.7. In all cases, the joint estimation approach best replicates the observed distributions. It is surprising to observe that the transitivity in isolation approach, which does not explicitly leverage any information regarding d, exhibit approximately the same replication performance as the PA in isolation approach, which explicitly leverages this information. This suggests that the dimension of b preserves a fair amount of information regarding d.

Secondly, we consider

$$p_{b|(d_1,d_2)\in\mathcal{D}} := Pr(X = b|(D_1, D_2) \in \mathcal{D}) = \sum_{(d_1,d_2)\in\mathcal{D}} p_{d_1,d_2,b} / \sum_{b=0}^{b_{max}} \sum_{(d_1,d_2)\in\mathcal{D}} p_{d_1,d_2,b},$$

where $\mathcal{D}$ is a non-empty set of unordered pairs. This is the conditional probability distribution of X given the event $(D_1, D_2) \in \mathcal{D}$. Given a pair of nodes with degrees d_1 and d_2, and their number of common neighbors is b, a natural condition is imposed on b: b must be not greater than either d_1 or d_2. Therefore, if one selects $\mathcal{D}$ such that d_1 or d_2 may be too small, the range of b will be severely limited. For this reason, we consider two probability distributions: $p_{b|\max(d_1,d_2)\leq 9}$ and $p_{b|\max(d_1,d_2)\geq 10}$, both of which allow a large range for b. Their cumulative distributions are presented in Fig. 4.8. Once again, the joint estimation approach best replicates the observed cumulative probability distributions in all cases. While the transitivity in isolation approach replicates the observed distributions fairly well in most cases, the PA in isolation approach completely fails to do so in all cases. This implies that, while the dimension of b appears to preserve a fair amount of information regarding d_1 and d_2, the dimensions of d_1 and d_2 maintain little information regarding b.

Overall, the joint estimation approach performs comparatively well. The surprisingly good performance of the transitivity in isolation approach is in agreement with the dominating role of B_b in the growth process of both networks. Combining the results in Fig. 4.6 with those in Figs. 4.7 and 4.8, we can conclude that the joint estimation approach captures both the first-order and second-order information of the networks reasonably well. This good fit is consistent with the fact that the key assumption of the time invariability of A_d and B_b is satisfied in both networks.

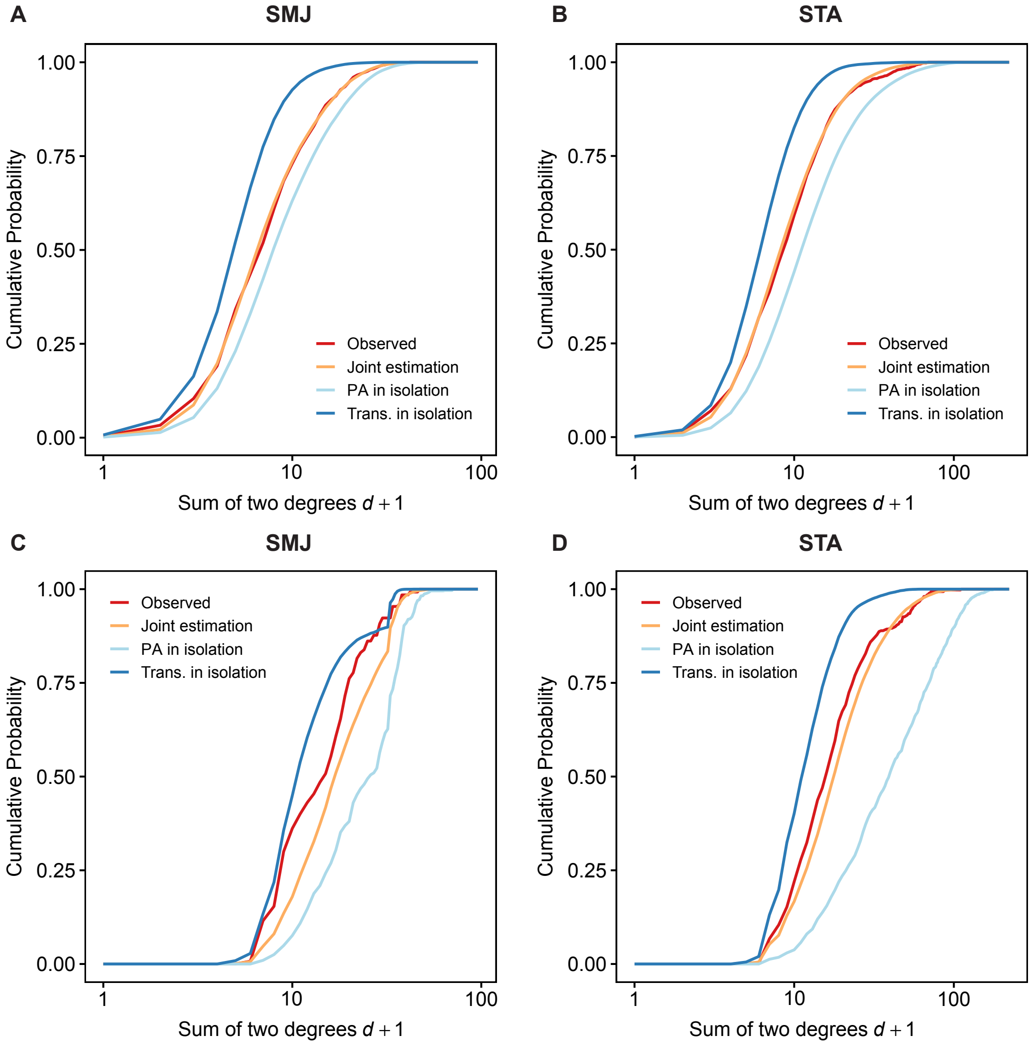


FIGURE 4.7: Observed and simulated cumulative probability distributions $p_{d|b\leq 1}$ and $p_{d|b\geq 2}$ of $d = d_1 + d_2$ in two networks. For each estimation method, we generate 100 networks from the estimation result and report the average values over 100 simulations. **A** and **B**: cumulative probability distributions $p_{d|b\leq 1}$ in SMJ and STA, respectively. **C** and : cumulative probability distributions $p_{d|b\geq 2}$ in SMJ and STA, respectively. In all cases, our joint estimation approach replicates the observed distributions comparatively well.

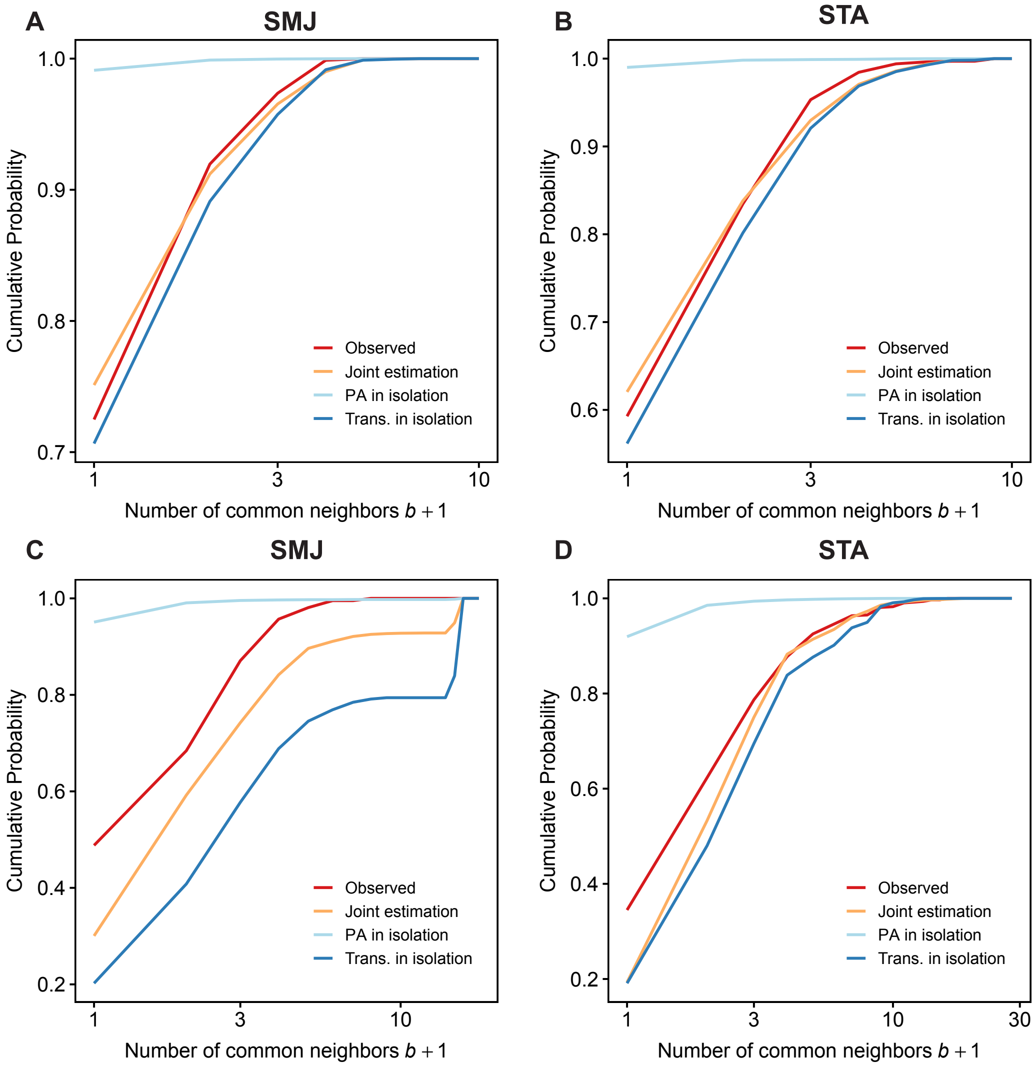


FIGURE 4.8: Observed and simulated cumulative probability distributions $p_{b|\max(d_1,d_2)\leq 9}$ and $p_{b|\max(d_1,d_2)\geq 10}$ in two networks. For each estimation method, we generate 100 networks from the estimation result and report the average values over 100 simulations. **A** and **B**: cumulative probability distributions $p_{b|\max(d_1,d_2)\leq 9}$ in SMJ and STA, respectively. **C** and : cumulative probability distributions of $p_{b|\max(d_1,d_2)\geq 10}$ in SMJ and STA, respectively. In all cases, our joint estimation approach replicates the observed distributions comparatively well.

4.1.5 Further Discussions

Estimation accuracy when β is small

We explain the comparatively large standard errors of β observed when both β and α are small in Fig. 3.2. Figure 4.9 presents the estimated B_b and the number of new edges corresponding to B_b in three randomly chosen networks for two cases: $\alpha = \beta = 0$ and $\alpha = 0$, $\beta = 1.5$. The comparatively large standard errors of β observed when both β and α are small appears to be due to the instability of B_b when b is large. This, in turn, is due to the relatively small number of new edges corresponding to B_b in this region.

Estimation of the standard deviations of $\hat{h}_{\textbf{trans}}(t)$ and $\hat{h}_{\textbf{PA}}(t)$

We have the following closed-form formula for the variance of the sample variance $h_{\text{trans}}(t)^2$:

$$\mathbb{V}h_{\text{trans}}(t)^2 = \frac{1}{m(t)}\,\mathbb{E}(\log_2 B_{b_{ij}} - \mathbb{E}\log_2 B_{b_{ij}})^4 - \frac{(m(t)-3)s_{\text{trans}}(t)^4}{m(t)(m(t)-1)}.$$

The delta method then provides:

$$sd(h_{\text{trans}}(t)) \approx \frac{1}{2}\left(\mathbb{E}\,h_{\text{trans}}(t)^2\right)^{-1/2}\sqrt{\mathbb{V}h_{\text{trans}}(t)^2} \approx \frac{1}{2}\left(s_{\text{trans}}(t)\right)^{-1}\sqrt{\mathbb{V}h_{\text{trans}}(t)^2}.$$

The standard deviation of $\hat{h}_{\text{trans}}(t)$ then can be calculated by inserting $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ into the above formula. The standard deviation of $\hat{h}_{\text{PA}}(t)$ follows the same derivation.

Estimation results in SMJ when excluding outliers

Figure 4.10 presents the estimation results of A_d and B_b in SMJ when excluding one paper with 17 author, compared to those obtained with the full dataset. The results of the two cases are similar.

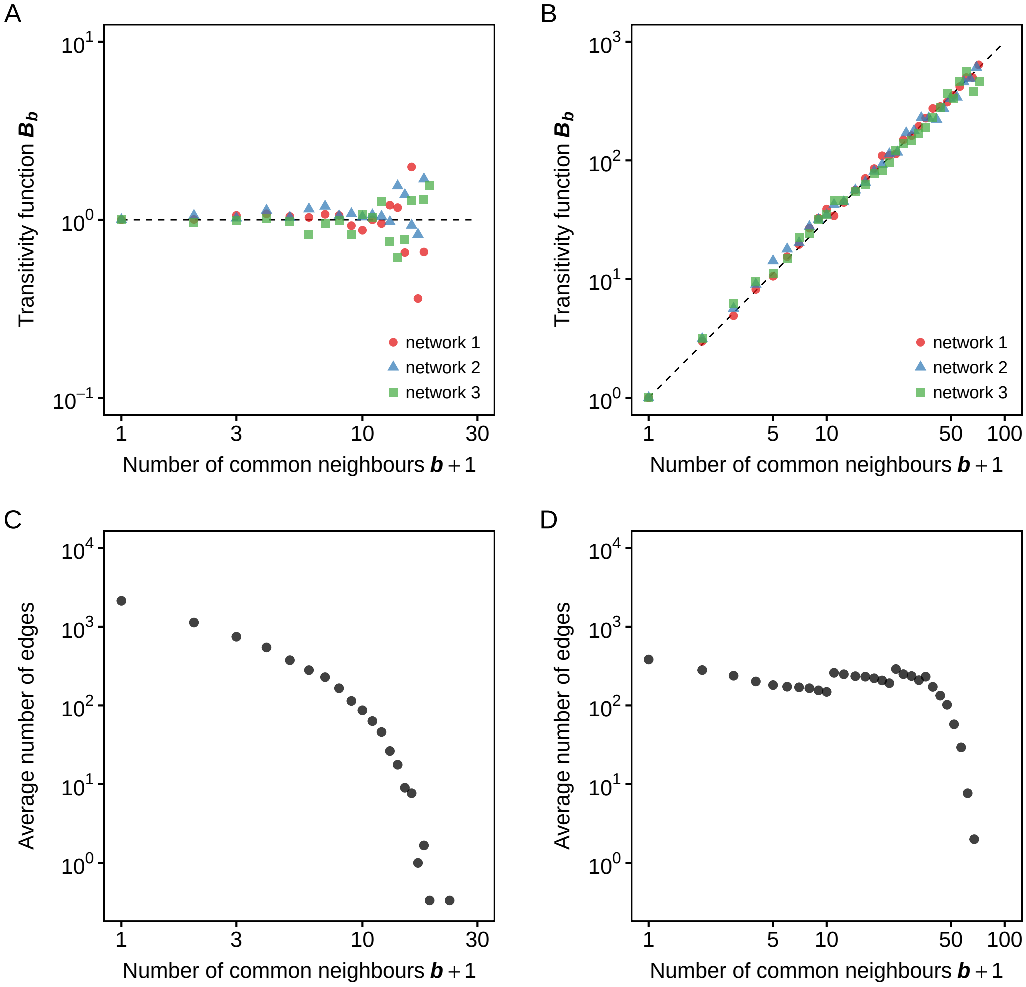


FIGURE 4.9: The comparatively large standard errors of β when α and β are small are ultimately due to a relatively small number of data points corresponding to B_b when b is large. **A**: estimated B_b for three random networks when $\alpha = \beta = 0$. The value of B_b is unstable for large b. **B**: estimated B_b for three random networks when $\alpha = 0$ and $\beta = 1.5$. The value of B_b is stable for large b. **C**: average number of new edges corresponding to b for the networks in panel A. The instability observed in panel A is due to a small number of new edges corresponding to b when b is large. : average number of new edges corresponding to b for the networks in panel B. Here, the average numbers of new edges for large b are relatively higher than those in panel C, leading to stable estimations.

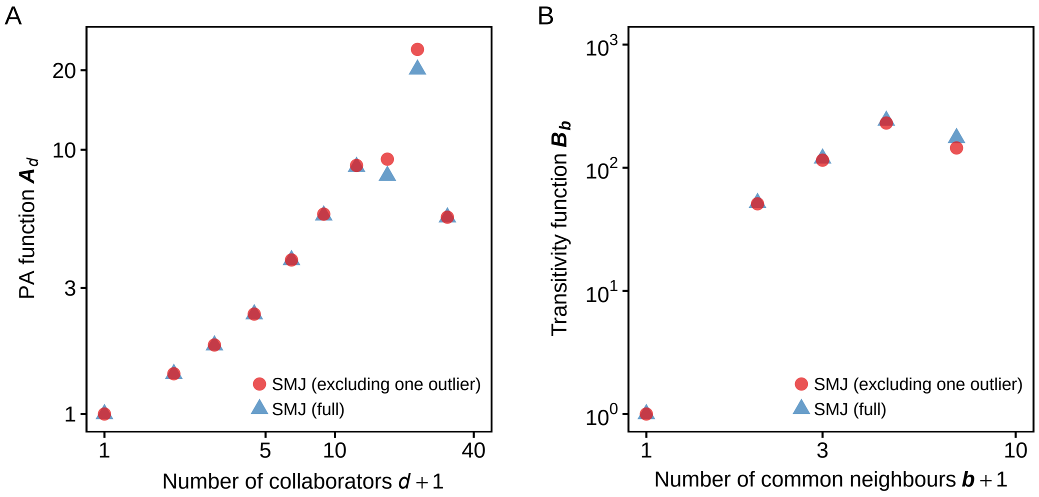


FIGURE 4.10: Estimated PA and transitivity functions in SMJ with and without the outlier. The results in two cases are similar. **A**: estimated PA functions. **B**: estimated transitivity functions.

4.2 Conclusion

In this part, we have provided the proposed statistical network model that incorporates non-parametric PA and transitivity functions, and have derived an efficient MM algorithm for estimating its parameters. Moreover, we presented a method that can quantify the amount of contributions of not only the PA and transitivity, but also many other network growth mechanisms, by exploiting the probabilistic dynamic process induced by the model formula.

We demonstrated that the proposed network model provided a reasonably good fit to two real-world co-authorship networks, and revealed intriguing properties of the PA and transitivity functions in these networks. The PA function increased on average in both networks, implying that the PA effect was at play. Excluding the high-degree part, it followed the conventional power-law form reasonably well. However, the transitivity function exhibited highly non-power-law behavior in the two networks: it jumped substantially after $b = 0$, but remained relatively horizontal or only increased slightly thereafter. This non-conventional form implies that co-authors of co-authors appear to be at least fifty times more likely to become new co-authors, compared to the case when no mutual co-author exists. Furthermore, we found that the transitivity dominated the PA in both networks, suggesting the importance of indirect relations in scientific creative processes.

There are several fascinating directions for further development of the statistical methodology. Firstly, although the proposed model and most other network models in the literature assume that the new edges at each time step are independent, this is hardly the case in real-world collaboration networks, where several co-authorships can emerge simultaneously from one paper with many authors. The independence assumption is thus a limitation of our model. It is important to devise a method to verify the impact of this assumption on the estimation of growth mechanisms. Moreover, efficiently relaxing this assumption may lead to improved models for scientific collaboration networks. These two problems are left for future work.

Secondly, it will be interesting to observe whether one could adapt the time-invariability test developed for stochastic actor-based models (Lospinoso et al., 2011) to our model. Lastly, it is worth extending our model to handle transitivity in directed networks.

On the application front, this work has laid out a potentially fruitful approach for analyzing complex networks, while raising more questions than it answers. For example, does transitivity always dominate PA in co-authorship networks? Which parametric forms are capable of capturing the fine details observed in Fig. 4.2? What are the properties of PA and transitivity in co-authorship networks at the level of institutions or countries? We hope that this research will convince informetricians to include non-parametric modeling of PA and transitivity into their toolbox.

Part II

PA Function in Evolving Hypergraphs

Chapter 5

HyperPA: A Hypergraph Approach for estimating Non-parametric PA Function

Our contributions, descriptions and figures in this chapter are based on our papers (Inoue et al., 2022).

5.1 Introduction

Network modeling, a notable application of graph theory, can reveal static and dynamic natures of interactions between individuals in various real-world complex systems (de Arruda et al., 2018; Riolo & Newman, 2020; Wang et al., 2019). However, in some data domains, there is an information loss in simplifying the interaction of complex systems with graphs: we implicitly break group interactions of three or more individuals into independent pairwise interactions. For example, in scientific co-authorship data, papers may be written by more than two researchers. The co-authorship of such papers is decomposed into pairwise co-authorship when the data is represented by graphs. This inability to preserve higher-order interactions is a serious limitation of graph representations. We can address this problem by replacing graphs with hypergraphs (de Arruda et al., 2020). Using hypergraphs, the co-authorship of each paper can be represented by one hyperedge, regardless of how many authors the paper has. This preserves the collectivity of co-authorship (Lung et al., 2018). This study aims to examine hypergraph growth models that can capture the dynamics of higher-order interactions in temporal data.

A hypergraph consists of a set of nodes and a set of hyperedges. A hyperedge contains an arbitrary number of nodes, whereas an edge in ordinary graphs contains only two nodes. The number of nodes that a hyperedge contains is referred to as the size of the hyperedge. We note that this size can take different values for each hyperedge. For a node in a hypergraph, the number of hyperedges containing it is called its "hyperdegree", whereas the number of edges connected to a node in ordinary graphs is called its "degree". In hypergraph representations of co-authorship data, each node and each hyperedge represents one researcher and the co-authorship of one paper, respectively. The size of a hyperedge indicates the number of co-authors of the corresponding paper, and the hyperdegree of a node corresponds to the number of papers the corresponding researcher has written in the past. Hypergraphs have been applied successfully in a wide variety of domains, including recommender systems (Zheng et al., 2018), bioinformatics (Mithani et al., 2009), classification (Sun et al., 2021), clustering (Kamiński et al., 2019), and document retrieval (Spitz et al., 2020).

Although there have been many attempts in complex network theory to model the growth of interactions in temporal data using graph representations, there is little existing research on hypergraph-based growth models. One of the most well-known growth mechanisms is preferential attachment (PA) (Barabási & Albert, 1999). PA is a "rich-get-richer" mechanism that can provide a compelling explanation of the heavy-tailed degree distributions appeared in many real-world networks. In this mechanism, the probability a node will get new edges at some time-step is proportional to its degree, i.e., the number of edges connected to the node up to that time-step. In case of temporal graph-based models, several models and estimation methods have been proposed for various growth mechanisms, including PA (Inoue et al., 2020b; Overgoor et al., 2019; Pham et al., 2015, 2016; Snijders, 2017).

Of the few existing works on hypergraph-based growth models, most employ a data-independent growth mechanism which is the linear preferential attachment (Do et al., 2020; Lee et al., 2021; Liu et al., 2014). This pre-specification of the growth mechanism risks over-simplifying the potentially complex interactions in real-world datasets. In fact, existing works that employed graph-based models suggested that the PA mechanism in real-world temporal graphs is hardly linear (Pham et al., 2015).

In this chapter, we propose a hypergraph growth model with a data-driven PA mechanism that can be estimated from observed data. Whereas graph-based PA mechanisms are defined on the degree of a node, our hyper-graph based PA mechanism is defined on the hyperdegree of a node. In our PA mechanism, a node with hyperdegree k, i.e., the node is contained in k hyperedges, will belong to a new hyperedge with probability proportional to A_k, the preferential attachment kernel. For example, the linear model is specified as $A_k = k$. The exact form of A_k is estimated from observed data.

The contributions of this part can be summarized as follows:

1. We propose a novel hypergraph-based growth model with a non-parametric PA kernel. What we mean by "non-parametric" is that the tunable parameter is

$$\boldsymbol{A} = [A_1, A_2, A_3, \ldots] \tag{5.1}$$

without specific functional forms. Since the model is invariant to the scale of $\boldsymbol{A}$, we may set $A_1 = 1$ without loss of generality. In most existing works on hypergraph-growth models, the linear PA kernel $A_k = k$ is assumed. Such unfounded pre-specification of the growth mechanism completely ignores the data at hand. In contrast, in our model, the PA kernel A_k is entirely free of assumptions. We stress that our non-parametric PA kernel is more flexible than the one-parameter kernel $A_k = k^\alpha$, which is often employed in graph-based growth models but not used in any existing hypergraph-based models. We provide a method to estimate from the data each value of A_k for each observed hyperdegree k. Specifically, we employ maximum likelihood estimation of $\boldsymbol{A}$ for this task and derive a recursive formula that significantly reduces the computation cost of the likelihood function of our model. An R package `HyperPA` of the proposed method will be available in (Inoue et al., 2022).

2. We provide a new approach to treat a selection bias that arises in modeling the emergence of new hyperedges with new nodes. Since parameter estimation in hypergraph-based growth models has not been considered, there is no existing work on this bias in hypergraph settings. In conventional graph-based growth models, in order to remove this bias, new edges are often removed from calculations of the log-likelihood function. However, a similar approach of removing hyperedges from calculations of the log-likelihood function would discard too much information, since the typical number of hyperedges with new nodes can be high in real-world datasets. In our method we use conditional probabilities in order to treat the selection bias in a principled way. We note that this approach can also be applied to graph-based growth models.

3. We fit our proposed model to 13 real-world datasets that can be divided into five categories: scientific co-authorship datasets, online thread participants datasets, online tagging datasets, national drug code directory datasets, and email datasets. We show that our proposed hypergraph PA model was better in replicating the observed data compared with conventional graph-based models. When one considers replications of the observed distributions of local clustering coefficients, the proposed hypergraph outperformed conventional models in all 13 datasets. When one considers replications of the observed distributions of the number of triplets, the proposed model provided the best fit in seven datasets, including all co-authorship networks. These findings confirm the importance of considering the collectivity of edges in modeling temporal complex data.

The rest of this part is organized as follows: Section 5.2 describes our hypergraph-based approach and presents illustrative results of our proposed model. Section 5.3 provides our estimation methodology and the hypergraph generation algorithm for our growth model. Section 6.1 explores the performance of the proposed method in 13 real-world datasets by comparing it with the conventional method. Finally, Section 6.2 concludes this work and outlines future work.

5.2 Hypergraph PA growth Model

In this section, we first describe our proposed PA growth model for hypergraphs. We then present illustrative results showing the effectiveness of the proposed model.

5.2.1 Proposed Hypergraph Growth Model

We propose a hypergraph version of the PA model based on the GT model (2.1). Instead of defining PA growth for edges, we define the probability of PA growth for hyperedges, i.e., sets of nodes. Let $G_t = (V_t, E_t)$ be the hypergraph at time-step t and $\mathcal{C}_m(V_t)$ be a family of sets whose elements are the sets that satisfy $B \subset V_t, |B| = m$. We define the probability that a node set $B = \{i_1, i_2, \ldots, i_m\} \in \mathcal{C}_m(V_t)$ acquires a hyperedge of size m at time-step t as follows:

$$P_B(t) \propto \prod_{i \in B} A_{k_i(t)}, \tag{5.2}$$

where $k_i(t)$ is the hyperdegree of node i at time-step t, and A_k is the PA value of hyperdegree k. We refer to the above proposed growth model as "Hyper PA". As in Edge PA (2.1), we assume no functional form for the PA function A_k. Do et al. (2020) proposed a generative model with linear PA $P_B \propto k_B$ for hypergraphs in which k_B is defined as the number of hyperedges that contain the set. However, when the size m is large, i.e., $|B| \gg 1$, the value of k_B becomes zero for almost all B, because such node sets seldom have hyperedges. Thus, this is not suitable for estimating the functional form of A_k. Therefore, in our model, we define the PA growth on the hyperdegrees k_i for $i \in B$.

Edge PA and Hyper PA are equivalent only for data where the size of all hyperedges is two. The difference between Hyper PA and Edge

PA emerges when we consider the probability of a hyperedge whose size is greater than two. As an example, let us consider the Hyper PA and Edge PA for a group interaction occurring on the set of three nodes $B = \{i_1, i_2, i_3\}$ at time-step t. In Hyper PA, this interaction is considered as a single hyperedge and the probability of this event is $P_B(t) \propto A_{k_{i_1}(t)} A_{k_{i_2}(t)} A_{k_{i_3}(t)}$. On the other hand, in Edge PA, the joint probability for all pairs of nodes is $P_{i_1,i_2}(t) P_{i_1,i_3}(t) P_{i_2,i_3}(t) \propto (A_{d_{i_1}(t)} A_{d_{i_2}(t)} A_{d_{i_3}(t)})^2$. In addition to the difference between using hyperdegree k_i and degree d_i, the exponent $m - 1$ in Edge PA, which is equal to 2 for the case of $m = 3$ above, makes the event of large m very rare.

The value of A_k in Hyper PA can be estimated from observed data by maximum likelihood estimation. More details of the proposed model, including a treatment of a selection bias that arises when the observed node set B contains some newcomer nodes, and estimation method are described in Sections 5.3.1 and 5.3.2. We can also generate hypergraphs with a given PA function A_k by a procedure provided in Section 5.3.3.

5.2.2 Illustrative Results

This section illustrates that our proposed Hyper PA model is better than the conventional Edge PA model and some other baseline models in terms of goodness-of-fit in two real-world co-authorship temporal networks: STA-coauthor from the statistics field (Ji & Jin, 2016) and HEP-coauthor from the high energy physics field (Inoue et al., 2018; Kunegis, 2013). The details of these datasets are provided in Section 6.1. We first fit the models to these data by estimating the PA function A_k for Hyper PA by our proposed method and A_d for Edge PA by the method in (Pham et al., 2015). We then compare some statistics of simulated graphs generated from the fitted models with those of the real-world data. The closer the simulated statistics of a model are to the real-world statistics, the better the model is in term of goodness-of-fit. To compare the hypergraphs generated by Hyper PA and the graphs generated by Edge PA, the hypergraphs were converted into graphs. The detail of the proposed estimation method and the procedure for generating hypergraph is described in Section 5.3.

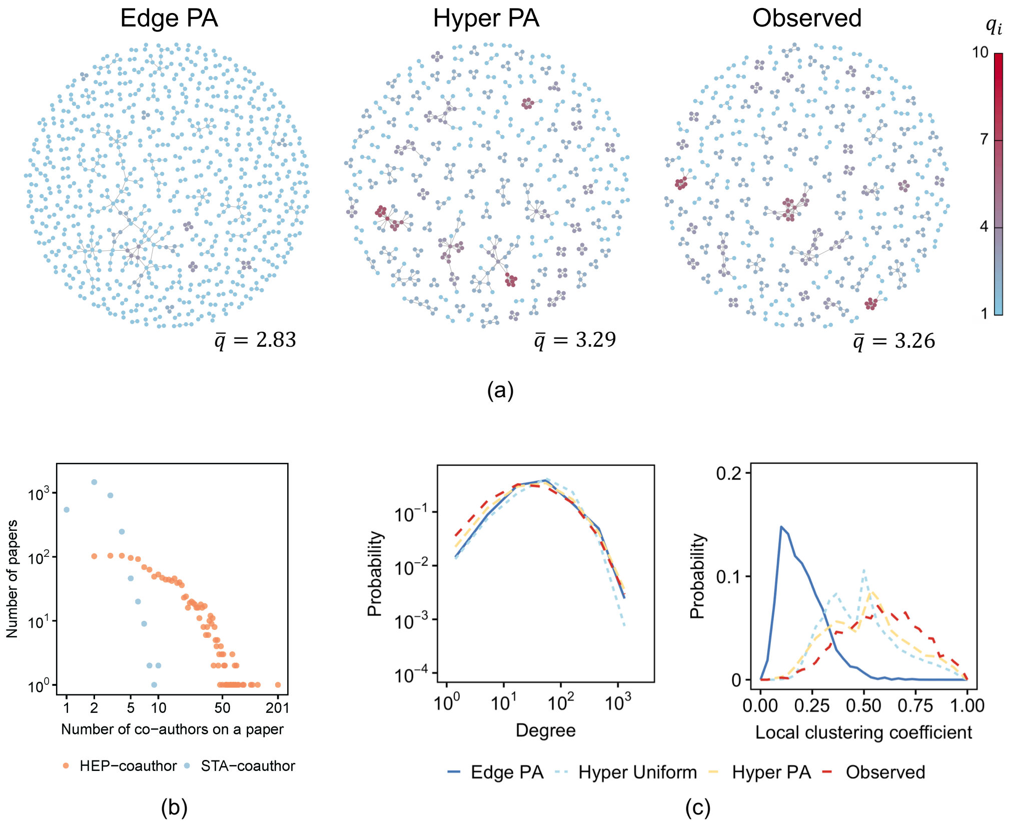


FIGURE 5.1: Hyper PA outperforms conventional models in reproducing first-order and second-order structures in scientific co-authorship data. **(a)**: Observed and simulated graphs of STA-coauthor, a dataset of co-authorships in journals from statistics field. Each graph illustrates the final 8% increments of the temporal graph. The color of each node represents the value q_i of the maximum size of cliques that contain the node. $\bar{q}$ is the average of all q_i in the data. Hyper PA outperformed Edge PA in reproducing both high values of q_i (red nodes) and the average $\bar{q}$. **(b)**: The observed distribution of the numbers of co-authors per paper, i.e., the sizes of hyperedges, of STA-coauthor and that of HEP-coauthor, a dataset of co-authorships in high energy physics. The size of a hyperedge can be enormous, as can be seen from HEP-coauthor. **(c)**: Observed and simulated probability distributions of degrees and local clustering coefficients in HEP-coauthor. The average values over 10 simulations are shown. The generated hypergraphs in Hyper PA and Hyper Uniform were converted into graphs for comparison. Hyper PA outperformed Edge PA and Hyper Uniform in replicating both distributions, thanks to hypergraph-based growth and PA mechanism.

Fig. 5.1(a) visualizes the final portion in the growth of STA-coauthor and those of the simulated temporal graphs generated by Hyper PA and Edge PA. Specifically, we plot the final 8% increments of the temporal graphs, which correspond to the last 260 papers that appear in STA-coauthor. To visualize the collectivity of edges around each node, we colored each node i according to the size q_i of the largest clique that contains i. In the graphs generated by Edge PA, there are fewer red nodes than in the observed data, which means that Edge PA did not capture enough higher-order information and failed to replicate large cliques. On the other hand, Hyper PA generated large cliques similarly to those observed in the real data. This observation can also be supported quantitatively by looking at the average $\bar{q}$ of q_i over the whole graphs in 10 simulations. To the observed value $\bar{q} = 3.26$, Hyper PA gave a close match of $\bar{q} = 3.29$, which is much closer than the value $\bar{q} = 2.83$ given by Edge PA.

The reason why Edge PA failed to replicate the collective nature of edge increments in STA-coauthor is that Edge PA adds each edge independently. The poor fit of Edge PA is also confirmed for second-order structures of graphs such as triangles in not only STA-coauthor but also other real-world co-authorship datasets. See Section 6.1.3 for more results.

As noted earlier, although conventional graph-based models such as Edge PA may be a reasonable modeling choice if the typical size of hyperedges in the data is small, this number can be enormous in some datasets. Fig. 5.1(b) shows the distributions of the numbers of co-authors per paper in STA-coauthor and HEP-coauthor. In hypergraph expression of scientific co-authorship, the number of co-authors of a paper is equal to the size of the corresponding hyperedge. As can be seen in Fig. 5.1(b), HEP-coauthor contains many relatively large hyperedges. The maximum hyperedge size is 201 for the HEP-coauthor and 10 for the STA-coauthor. More detailed data descriptions for all datasets used in this part are provided in Section 6.1.1. For a dataset that has a tendency for edge collectivity as strong as HEP-coauthor, one would expect clear differences between Hyper PA and Edge PA. The following experiment in Fig. 5.1(c) illustrates this point.

In Fig. 5.1(c), we demonstrate that both hypergraph-based growth and

PA mechanism are needed to provide a reasonably good fit to HEP-coauthor. To this end, we add another baseline model, namely Hyper Uniform, that is a special case of Hyper PA in which $A_k = 1$ for every hyperdegree k, i.e., there is no PA effect in Hyper Uniform. Fig. 5.1(c) shows the observed and simulated probability distributions of degrees and local clustering coefficients. The local clustering coefficient, whose mathematical definition is given in Section 6.1.3, is a popular way to express the density of triangles around a node. The distribution of local clustering coefficients can be used to express the degree of collectivity of edges in the data.

From Fig. 5.1(c), the following observations can be made.

1. Hyper PA outperformed both Edge PA and Hyper Uniform in reproducing the overall degree distribution, thanks to both the hypergraph-based growth and the PA mechanism. Although Edge PA captured better the right tail of the degree distribution compared with Hyper Uniform, both Edge PA and Hyper Uniform underestimated the portion of low degrees compared with Hyper PA. This implies that the PA mechanism may be responsible for reproducing the right tail of the degree distribution, whereas the hypergraph-based growth is potentially responsible for replicating the left tail.

2. In replicating the distribution of local clustering coefficients, Hyper PA also outperformed both Edge PA and Hyper Uniform. Edge PA significantly underestimated the local clustering coefficients, which implies that it could not capture the collectivity of edges in the data. This is expected, since Edge PA adds edges independently.

To summarize, both hypergraph growth and PA mechanism are crucial in capturing first-order and second-order structures of the data. Hyper PA employs both ingredients and thus was able to provide good fits to STA-coauthor and HEP-coauthor compared with conventional models. Further experiments are provided in Chapter 6.

5.3 Methodology of Estimation

In this section, we describe the maximum likelihood estimation of the PA function in our model and pseudo codes for generating temporal hypergraphs from our model. In addition to the derivation of the likelihood function, we provide a recursive formula that enables a fast calculation of the likelihood function. We also provide a principled approach based on conditional probabilities for handling a selection bias that arises in modeling the emergence of new hyperedges with newcomer nodes.

5.3.1 Maximum Likelihood Estimation

Likelihood Function

We first derive the likelihood function of $\boldsymbol{A}$ for Hyper PA model. As we described in Section 5.2, our growth model (5.2) is based on the undirected GT model. Therefore, the derivation of the likelihood function in Hyper PA model here is also based on previous works (Inoue et al., 2020b; Pham et al., 2015, 2016) where maximum likelihood estimation of the PA function is derived for the GT model.

We define some notations needed in the exposition. Let $G_t = (V_t, E_t)$ be the hypergraph at time-step t. V_t and E_t are the node set and the hyperedge set, respectively. Let $\{G_t\}_{t=0}^{T}$ be the hypergraph sequence, and $\{h_t\}_{t=1}^{T}$ be the sequence of the number of hyperedges added to the hypergraph at each time-step. We denote the size of each hyperedge at time-step t as $\boldsymbol{m}_t = [m_{t,1}, \ldots, m_{t,h_t}]$ and the number of newcomer nodes that appear with each hyperedge as $\boldsymbol{n}_t = [n_{t,1}, \ldots, n_{t,h_t}]$. For the l-th $(1 \leq l \leq h_t)$ hyperedge at time-step t, its size is given by $m_{t,l}$, and we have $0 \leq n_{t,l} \leq m_{t,l}$; the number of newcomer nodes is $n_{t,l}$ and the number of existing nodes is $m_{t,l} - n_{t,l}$. This hyperedge contains solely existing nodes if $n_{t,l} = 0$, and contains solely newcomer nodes if $n_{t,l} = m_{t,l}$.

Now we consider the probability that some node set $B_{t,l}$ whose size is $m_{t,l}$ acquires a new hyperedge of size $m_{t,l}$. If we assume that $B_{t,l}$ contains only existing nodes, the acquisition probability is:

$$P_{B_{t,l}}(t) = \frac{\prod_{i \in B_{t,l}} A_{k_i(t)}}{\sum_{B' \in \mathcal{C}_{m_{t,l}}(V_t)} \prod_{i' \in B'} A_{k_{i'}(t)}}, \tag{5.3}$$

where $\mathcal{C}_{m_{t,l}}(V_t)$ is the family of sets such that a set B belongs to $\mathcal{C}_{m_{t,l}}(V_t)$ if and only if $B \subset V_t$ and $|B| = m_{t,l}$. The case that $B_{t,l}$ contains some newcomer nodes needs some special care, since there is a selection bias. This problem is treated in Section 5.3.2.

Suppose that the joint distribution of h_t, $\boldsymbol{m}_t$, and $\boldsymbol{n}_t$ is governed by the parameter vector $\boldsymbol{\theta}_t$, and that the initial hypergraph G_0 is determined by $\boldsymbol{\theta}_{\text{init}}$. As in previous works (Inoue et al., 2020b; Pham et al., 2015, 2016), we assume that $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{\text{init}}$ are independent of $\boldsymbol{A}$ in growth process. This assumption is interpreted as follows: the increments of hypergraph (i.e. the number of additional nodes and hyperedges) at each time-step are independent of the PA growth. With this assumption, the probability of the observed data can be written as:

$$\begin{aligned}
&P(G_0, \ldots, G_T) \\
=&\prod_{t=1}^{T} P(G_t|G_{t-1})P(G_0) \\
=&\prod_{t=1}^{T} P(G_t|G_{t-1}, h_t, \boldsymbol{m}_t, \boldsymbol{n}_t, \boldsymbol{A})P(h_t, \boldsymbol{m}_t, \boldsymbol{n}_t|G_{t-1}, \boldsymbol{\theta}_t) \\
&\quad \cdot P(G_0|\boldsymbol{\theta}_{\text{init}}).
\end{aligned}$$

Taking the logarithm of both sides, the log-likelihood function of $\boldsymbol{A}$ can be expressed as:

$$\begin{aligned}
&L(\boldsymbol{A}|G_0, \ldots, G_T) \\
=&\sum_{t=1}^{T} \log P(G_t|G_{t-1}, h_t, \boldsymbol{m}_t, \boldsymbol{n}_t, \boldsymbol{A}) \\
&+ \sum_{t=1}^{T} \log P(h_t, \boldsymbol{m}_t, \boldsymbol{n}_t|G_{t-1}, \boldsymbol{\theta}_t) + \log P(G_0|\boldsymbol{\theta}_{\text{init}}).
\end{aligned}$$

Since on the right-hand side only the first term includes the PA function $\boldsymbol{A}$, we can omit the other terms related to the nuisance parameters $\boldsymbol{\theta}_{\text{init}}$ and $\boldsymbol{\theta}_t$. The log-likelihood function can then be rewritten as follows:

$$
\begin{aligned}
&L(\boldsymbol{A}|G_0,\ldots,G_T) \\
&= \sum_{t=1}^{T} \log P(G_t|G_{t-1}, h_t, \boldsymbol{m}_t, \boldsymbol{n}_t, \boldsymbol{A}) \\
&= \sum_{t=1}^{T}\sum_{l=1}^{h_t} \log P_{B_{t,l}}(t) \qquad (5.4) \\
&= \sum_{t=1}^{T}\sum_{l=1}^{h_t} \log \prod_{i \in B_{t,l}} A_{k_i(t)} \\
&\quad - \sum_{t=1}^{T}\sum_{l=1}^{h_t} \log \left(\sum_{B' \in \mathcal{C}_{m_{t,l}}(V_t)} \prod_{i \in B'} A_{k_i(t)} \right). \qquad (5.5)
\end{aligned}
$$

Note that we substituted (5.3) into (5.4).

The maximum likelihood method estimates the value of $\boldsymbol{A}$ by maximizing $L(\boldsymbol{A}|G_0,\ldots,G_T)$. The parameter vector $\boldsymbol{A}$ in (5.1) actually includes only elements A_k with the observed values of k in the dataset. In addition, to reduce the number of parameters, we employ the "logarithmic binning" (Inoue et al., 2020b; Pham et al., 2015, 2016) of k, where we set $A_{k+1} = A_k$ in groups of k values.

Since (5.5) is computed numerically, we need its efficient evaluation. The term $\sum_{B' \in \mathcal{C}_{m_{t,l}}(V_t)} \prod_{i \in B'} A_{k_i(t)}$ is the normalization of the probability (5.3), which is the summation of the probabilities of every node set in $\mathcal{C}_{m_{t,l}}(V_t)$. When the hyperedge size $m_{t,l}$ is large, the computational cost of this term becomes intractable in a naive calculation. This is because of the combinatorial explosion of the number of possible node sets. Specifically, when the number of nodes in the entire hypergraph at time-step t is $N(t)$, the computational complexity of a naive calculation is $\mathcal{O}\left(\binom{N(t)}{m_{t,l}}\right) = \mathcal{O}\left(\frac{N(t)!}{m_{t,l}!(N(t)-m_{t,l})!}\right)$, which scales exponentially in $N(t)$ if $N(t)$ is much larger than $m_{t,l}$. Next we describe a fast computation which scales linearly in $N(t)$.

A Recursive Formula for Fast Computation of the Normalizing Factor

A fast computation of the normalizing factor

$$S_m(t) = \sum_{B' \in \mathcal{C}_m(V_t)} \prod_{i \in B'} A_{k_i(t)} \tag{5.6}$$

is possible if one can find a way to reduce the numbers of summations needed by exploiting its recursive structures.

Our key observation is that (5.6) is in fact an *elementary symmetric polynomial*, namely, it is the sum of all distinct products of m distinct variables. We define the elementary symmetric polynomial $e_m(x_1, \ldots, x_n)$ $(0 \leq m \leq n)$ with variables $x_1, \ldots, x_n$ as follows. For $m = 0$, $e_0(x_1, \ldots, x_n) = 0$, and for $m > 0$,

$$e_m(x_1, \ldots, x_n) = \sum_{1 \leq i_1 < i_2 < \cdots < i_m \leq n} x_{i_1} x_{i_2} \cdots x_{i_m}.$$

From the definition above, the normalizing factor can be written as an elementary symmetric polynomial of a suitable choice of variables, namely

$$S_m(t) = e_m(A_{k_1(t)}, \ldots, A_{k_{N(t)}(t)}).$$

We also define the m-th power sum as

$$p_m(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i^m,$$

where m and n are positive integers. According to Newton's identities (Baker, 1959), we have:

$$\begin{aligned} & e_m(x_1, \ldots, x_n) \\ =& \frac{1}{m} \sum_{j=1}^{m} (-1)^{j-1} e_{m-j}(x_1, \ldots, x_n) p_j(x_1, \ldots, x_n), \end{aligned}$$

for all positive integers m and n satisfying $m \leq n$. We finally arrive at the key recursive formula:

$$S_m(t) = \frac{1}{m} \sum_{j=1}^{m} (-1)^{j-1} S_{m-j}(t) p_j(A_{k_1(t)}, \ldots, A_{k_{N(t)}(t)}). \tag{5.7}$$

Note that $m \leq N(t)$ always holds in hypergraphs. Our approach is to use this formula recursively to calculate the normalizing factor $S_m(t)$. Since the calculation of the power sums is dominating in (5.7), the time complexity of calculating $S_m(t)$ can be reduced from $\mathcal{O}\left(\frac{N(t)!}{m!(N(t)-m)!}\right)$ to $\mathcal{O}\left(m^2 N(t)\right)$. By utilizing (5.7), the log-likelihood function given in (5.5) can be optimized by standard numerical methods such as quasi-Newton methods or MM algorithms (Pham et al., 2015).

5.3.2 A Selection Bias in Modeling The Emergence of New Hyperedges with Newcomer Nodes

We consider the case that the node set $B_{t,l}$ contains some newcomer nodes. Denote such $B_{t,l}$ simply as B. Naively treating the hyperdegree of newcomer nodes as $k = 0$ causes a selection bias, since in that way such newcomer nodes with hyperdegree $k = 0$ acquire new hyperedges *a priori*. Here we assume for our dataset that newcomer nodes are included in G_t only when they got a hyperedge. This may lead to overestimation of the true value of A_0. We are going to solve this problem by considering conditional probabilities given that newcomer nodes acquire new hyperedges.

Whereas one can use an existing remedy for a similar bias occurring in graph-based models (Pham et al., 2015, 2016), this conventional approach is sub-optimal in hypergraph settings. Specifically, this approach excludes any new hyperedge that contains some newcomer nodes in calculating the log-likelihood in (5.4). Since the proportion of new hyperedges with newcomer nodes can be high in many real-world data (Guimerà et al., 2005), this leads to throwing away too much data and risks destabilizing the estimation of A_k.

Assume that B consists of B_1 and B_2, where B_1 is the set of newcomer nodes and B_2 is the set of existing nodes. Instead of throwing away all the information contained in B as in the existing remedy approach, our approach is to salvage the portion of information contained in the event that a new hyperedge emerges on the set B_2 of existing nodes, given that the hyperedge also contains the set B_1 of newcomer nodes. We will include a term for B_2 in the log-likelihood function, and thus it contributes to the estimation of $\boldsymbol{A}$. However, we do not estimate A_0, because we assume that existing nodes have at least one hyperedge.

This intuition can be formalized as follows. For convenience, we denote $V_{\text{new}} = V_t \setminus V_{t-1}$, $V_{\text{exist}} = V_{t-1}$. With these new notations, note that $B = B_1 \cup B_2$, $B_1 \subset V_{\text{new}}$, and $B_2 \subset V_{\text{exist}}$. Let n and m' be the sizes of B_1 and B_2, respectively, and thus the size of B is $m = n + m'$. Now we consider the conditional probability that B gets a new hyperedge, conditioned on the event that the set of newcomer nodes is equal to some pre-specified set B^* with $B^* \subset V_{\text{new}}$. This conditional probability can be written as:

$$\begin{aligned} & P_{B|B_1=B^*}(t) \\ = & \frac{\prod_{i \in B_2} A_{k_i(t)}}{\sum_{B_2' \in \mathcal{C}_{m'}(V_{\text{exist}})} \prod_{i' \in B_2'} A_{k_{i'}(t)}}, \end{aligned} \tag{5.8}$$

which is essentially equivalent to (5.3) but applied to the B_2 part. In other words, we simply ignore the B_1 part. In calculating (5.4), if the observed node set $B_{t,l}$ contains only existing nodes, one uses (5.3), whereas if it contains some newcomer nodes, one uses (5.8). Note that in (5.8), all calculations occur solely on existing nodes. Therefore, we can remove the selection bias and obtain a stable estimate of $A_k (k > 0)$ at the same time. The calculation of the denominator of (5.8) can also be accelerated by the recursive formula (5.7). We next provide a derivation of (5.8).

Derivation

We here derive the conditional probability (5.8). Let $G_t = (V_t, E_t)$ be the hypergraph at time-step t. V_t and E_t are the node set and the hyperedge set, respectively. For convenience, we denote $V_{\text{new}} = V_t \setminus V_{t-1}$ and $V_{\text{exist}} = V_{t-1}$. Recall that Hyper PA determines the probability that a set B of m nodes will acquire a hyperedge of size m. Let B_1 and B_2 be the sets of the nodes satisfying $B = B_1 \cup B_2$, $B_1 \subset V_{\text{new}}$, $B_2 \subset V_{\text{exist}}$, $|B_1| = n$, and $|B_2| = m - n = m'$. We here decompose (5.2) into the conditional probability given that $B_1 = B^*$ for a pre-specified set $B^* \subset V_{\text{new}}$ and the probability of observing B^*:

$$\begin{aligned}
&P_B(t) \\
=&P_{B_1 \cup B_2}(t) \\
=&P_{B_1 \cup B_2, B_1 = B^*}(t) \\
=&P_{B_1 \cup B_2 | B_1 = B^*}(t) P_{B_1 = B^*}(t).
\end{aligned} \tag{5.9}$$

The term $P_{B_1 \cup B_2 | B_1 = B^*}(t)$ corresponds to the desired conditional probability in (5.8). Note that we denote $B^* \subset V_{\text{new}}$ and not $B^* = V_{\text{new}}$ because we allow the temporal hypergraphs to add multiple hyperedges at each time-step t. With (5.3) and (5.9), we obtain:

$$\begin{aligned}
&P_{B_1 \cup B_2 | B_1 = B^*}(t) \\
=&\frac{P_{B_1 \cup B_2}(t)}{P_{B_1 = B^*}(t)} \\
=&\frac{\frac{(\prod_{i \in B_1} A_{k_i(t)})(\prod_{i \in B_2} A_{k_i(t)})}{\sum_{B' \in \mathcal{C}_m(V_{\text{exist}} \cup V_{\text{new}})} \prod_{i' \in B'} A_{k_{i'}(t)}}}{\frac{\sum_{B'_2 \in \mathcal{C}_{m'}(V_{\text{exist}})} (\prod_{i \in B_1} A_{k_i(t)})(\prod_{i' \in B'_2} A_{k_{i'}(t)})}{\sum_{B' \in \mathcal{C}_m(V_{\text{exist}} \cup V_{\text{new}})} \prod_{i' \in B'} A_{k_{i'}(t)}}} \\
=&\frac{\left(\prod_{i \in B_1} A_{k_i(t)}\right)\left(\prod_{i \in B_2} A_{k_i(t)}\right)}{\left(\prod_{i \in B_1} A_{k_i(t)}\right)\left(\sum_{B'_2 \in \mathcal{C}_{m'}(V_{\text{exist}})} \prod_{i' \in B'_2} A_{k_{i'}(t)}\right)} \\
=&\frac{\prod_{i \in B_2} A_{k_i(t)}}{\sum_{B'_2 \in \mathcal{C}_{m'}(V_{\text{exist}})} \prod_{i' \in B'_2} A_{k_{i'}(t)}},
\end{aligned}$$

thus showing (5.8).

So far we have assumed that $B_1 = B^*$ is pre-specified, but we can change the setting so that B_1 is randomly sampled from V_{new} with $P(B_1) =$

$1/\binom{|V_{\text{new}}|}{n}$. Then $\log P(B_1)$ terms may be added to the log-likelihood function $L(\boldsymbol{A}|G_0, \ldots, G_T)$. However, since $\log P(B_1)$ does not involve $\boldsymbol{A}$, it does not change the maximum likelihood estimation of $\boldsymbol{A}$.

5.3.3 Algorithm for Generating Hypergraphs

In this section, we describe the procedure for generating hypergraphs in simulations. The pseudocode for our proposed hypergraph generator is provided in Algorithm 1.

First, we describe how to determine the input of the generator. By using real-world observations, we can set reasonable values in our simulations; inputs (ii) to (vi) can be given directly as descriptive statistics of a dataset, whereas input (i) needs to be estimated from the data. A_k can be given either as an estimated nonparametric sequence or a function with estimated parameters.

At each time t of the iteration in the procedure, the set of nodes that acquire a new hyperedge (v_i^{exist} at line 5) is sampled from the Hyper PA model with the `HyperPA` procedure, which is described at the bottom of Algorithm 1 as a subroutine. We use the conditional probability given in (5.8) of Hyper PA to separate the effects of the node birth process from the hyperedge acquisition process. In our experiments, the set of newcomer nodes (v_i^{new} at line 4) is simply taken from the history of a dataset; this is not explicitly described in Input though.

Algorithm 1 The proposed Hyper PA generator for temporal hypergraph.

```
Input: (i) preferential attachment: $\boldsymbol{A}$
    (ii) initial hypergraph: $G_0 = (V_0, E_0)$
    (iii) timespan : $T$
    (iv) sequence of the number of new hyperedges: $\{h_t\}_{t=1}^{T}$
    (v) sequence of hyperedge size: $\{\boldsymbol{m}_t\}_{t=1}^{T}$,
        $\boldsymbol{m}_t = [m_{t,1}, \ldots, m_{t,h_t}]$
    (vi) sequence of the number of emerging nodes: $\{\boldsymbol{n}_t\}_{t=1}^{T}$,
        $\boldsymbol{n}_t = [n_{t,1}, \ldots, n_{t,h_t}]$

Output: evolving hypergraph: $\{G_t\}_{t=1}^{T}$, $G_t = (V_t, E_t)$
for time $t$ in $[1, \ldots, T]$ do
    set $V_t \leftarrow V_{t-1}$ and $E_t \leftarrow E_{t-1}$
    for $i$ in $[1, \ldots, h_t]$ do
        $v_i^{\text{new}} \leftarrow$ set $n_{t,i}$ newcomer nodes
        $v_i^{\text{exist}} \leftarrow$ sample $m_{t,i} - n_{t,i}$ nodes from $V_{t-1}$ by HyperPA$(\boldsymbol{A}, G_{t-1}, m_{t,i}, n_{t,i})$
        $e_i \leftarrow$ set a hyperedge containing $v_i^{\text{new}} \cup v_i^{\text{exist}}$
        add $v_i^{\text{new}}$ to $V_t$ and $e_i$ to $E_t$
    end for
    $G_t \leftarrow (V_t, E_t)$
end for
return $\{G_t\}_{t=1}^{T}$

subroutine: HyperPA$(\boldsymbol{A}, G_{t-1}, m_{t,i}, n_{t,i})$
$\{k_j\}_{j=1}^{N(t)} \leftarrow$ calculate the hyperdegrees for all existing nodes at time $t-1$ from hypergraph $G_{t-1}$
$v_i^{\text{exist}} \leftarrow$ sample $m_{t,i} - n_{t,i}$ nodes according to the probability $P_B(t) \propto \prod_{j \in B} A_{k_j(t)}, B \in \mathcal{C}_{m_{t,i}-n_{t,i}}(V_{t-1})$
```

Chapter 6

Real Data Analysis with HyperPA

Our contributions, descriptions and figures in this chapter are based on our papers (Inoue et al., 2022).

6.1 Experiments

In this section, we first describe the real-world datasets and then present the estimation results for the PA function A_k in these datasets. We then perform simulations to evaluate goodness-of-fit of our proposed hypergraph model.

6.1.1 Real-world Datasets

We use 13 real-world datasets as temporal hypergraphs in experiments. The datasets can be divided into five categories.

- **Scientific co-authorship datasets**: Each node is an author and each hyperedge is a set of authors who have written a paper collaboratively. We use the following four datasets: Complex Network Theory (CMP-coauthor) (Pham et al., 2020), High Energy Physics (HEP-coauthor) (Inoue et al., 2018; Kunegis, 2013), Strategic Management Journal (SMJ-coauthor) (Ronda-Pupo & Pham, 2018), and Statistics (STA-coauthor) (Ji & Jin, 2016).

- **Online thread participants datasets**: Each node represents a user answering questions on threads and each hyperedge describes a set

TABLE 6.1: Summary statistics and experiment results for 13 datasets. Summary statistics of the datasets described in Section 6.1.1; T is the number of time-steps, N is the number of nodes, L is the number of edges, H is the number of hyperedges, $\bar{M}$ is the average size of the hyperedges, γ is the power-law exponent of the degree distribution, and C is the clustering coefficient. Here L and H are counting repetitions in duplicate. α is the estimated PA exponent by fitting the proposed hypergraph-based growth model in Section 6.1.2. Each of E^{triplet} is the mean value of the error between the probability distributions of the numbers of triplets in observed and simulated data with Edge PA (EP), Hyper Uniform (HU), and Hyper PA (HP) in Section 6.1.3. Each of E^{local} is the mean value of the error between the distributions of local clustering coefficients averaged over nodes with each degree in observed and simulated data with EP, HU, and HP in Section 6.1.3. For each of E^{triplet} and E^{local}, the values are expressed in percentages, and the best values are in bold (lower is better).

									E^{triplet}			E^{local}		
	T	N	L	H	$\bar{M}$	γ	C	α	EP	HU	HP	EP	HU	HP
CMP-coauthor	57	1498	3095	880	2.85	2.67	0.69	1.44 (±0.08)	3.0	**0.8**	1.0	21.6	14.6	**7.6**
HEP-coauthor	85	6798	293484	1446	14.6	2.23	0.33	1.04 (±0.04)	3.5	4.0	**1.8**	40.5	5.2	**3.0**
SMJ-coauthor	108	2704	4131	2243	2.26	2.35	0.38	1.18 (±0.09)	4.2	1.3	**0.5**	17.4	6.9	**2.6**
STA-coauthor	44	3607	6808	3247	2.36	2.90	0.32	1.19 (±0.05)	3.9	1.4	**0.3**	16.6	4.9	**3.1**
ask-ubuntu-user	50	4520	2296	5000	1.37	2.20	0.06	1.52 (±0.10)	0.6	0.7	**0.2**	6.7	2.7	**0.9**
math-sx-user	50	3972	8711	5000	2.07	2.07	0.10	1.19 (±0.05)	2.6	3.3	**0.7**	11.3	4.0	**0.5**
stack-overflow-user	50	6456	4742	5000	1.66	3.00	0.31	1.39 (±0.10)	1.6	0.8	**0.4**	12.9	3.8	**2.2**
ask-ubuntu-tags	50	1428	16095	5000	2.75	2.01	0.14	1.04 (±0.01)	2.5	7.2	**1.5**	8.5	10.3	**2.5**
math-sx-tags	50	970	11471	5000	2.36	2.07	0.18	0.97 (±0.02)	2.2	6.9	**1.8**	10.6	12.2	**3.7**
stack-overflow-tags	50	3705	18240	5000	2.95	2.10	0.06	1.10 (±0.03)	**3.9**	5.7	6.9	7.8	5.3	**2.0**
NDC-classes	50	632	16201	4995	2.63	1.79	0.58	0.91 (±0.02)	6.0	12.2	**5.6**	49.0	39.4	**19.4**
NDC-substances	50	1425	26395	4996	1.85	1.85	0.47	1.15 (±0.08)	**3.0**	5.4	3.8	38.6	12.3	**2.7**
Eu-email	38	681	19454	5000	2.37	1.82	0.55	0.77 (±0.02)	4.6	4.5	**3.1**	34.3	26.2	**13.3**

of users in a thread in which questions are posted. We use three datasets created from sub-forums of the online Stack Exchange forum: ask-ubuntu-user, math-sx-user, and stack-overflow-user.

- **Online tagging datasets**: Each node is a tag and each hyperedge is a set of tags associated with a question. We use three datasets created from sub-forums of the Stack Exchange forum: ask-ubuntu-tags, math-sx-tags, and stack-overflow-tags.
- **National drug code (NDC) directory datasets**: We use two datasets: NDC-classes and NDC-substances. Each node is a class label of drugs (NDC-classes) or a substance in drugs (NDC-substances) and each hyperedge is a set of class labels of a drug or a set of substances in a drug.
- **Email network datasets**: Each node is an email address and each hyperedge is a set of email addresses of the sender and all recipients contained in an email. There is one dataset in this category: Eu-email.

Except for the four scientific co-authorship datasets, the remaining datasets are from the hypergraph collection of Benson et al. (2018).

Some preprocessing is needed before one can perform model fitting. For each dataset in Benson et al. (2018), we extracted the latest 5000 hyperedges for analysis. In addition, several datasets with too few or too many nodes extracted from Benson et al. (2018) are excluded from the analysis and not listed here. In each dataset, we set the initial state G_0 as the first 50% of each data in terms of the number of edges. In co-authorship datasets, except for STA-coauthor, the original datasets only have temporal graphs and do not contain hyperedges. For this reason, we heuristically reconstructed the hyperedges from the increments of edges at each time-step. Specifically, at each time-step, we repeatedly replaced the new edges that constitute the largest clique with a new hyperedge until there was no more new edges. We tested this procedure on the STA-coauthor, and confirmed that all hyperedges were successfully reconstructed from its graph representation.

Table 6.1 shows some summary statistics of the datasets. It is important to note that HEP-coauthor contains large collaborative research projects such as accelerator physics (Kahn, 2017), and hence the average number of co-authors per paper (i.e., the average size of hyperedges) is larger than the other datasets.

6.1.2 Preferential Attachment in 13 Datasets

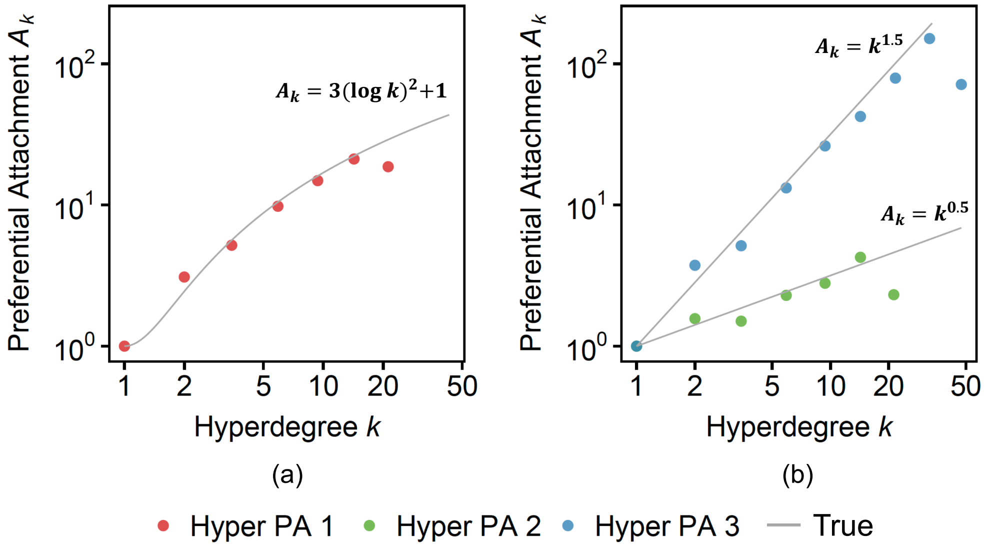


FIGURE 6.1: Our proposed method can estimate the PA function A_k from observed temporal hypergraphs without any assumptions on the functional form of A_k. We generated synthetic hypergraphs from the Hyper PA model, and applied our method to recover the PA function from the simulated data. **(a)**: Hyper PA 1 with $A_k = 3(\log k)^2 + 1$ as the true PA function. **(b)**: Hyper PA 2 and Hyper PA 3 with $A_k = k^{0.5}$ and $A_k = k^{1.5}$, respectively, as the true PA function. In the three functional forms, the method successfully recovered the PA functions.

We first demonstrate that our estimation method works in some hypergraphs generated from the Hyper PA model. We generated one hypergraph (Hyper PA 1) using the PA function $A_k = 3\,(\log k)^2 + 1$, and two other hypergraphs, namely, Hyper PA 2 and Hyper PA 3, using the PA function $A_k = k^\alpha$ with $\alpha = 0.5$ and $\alpha = 1.5$, respectively. The former functional form is also used in previous works (Inoue et al., 2020b; Pham et al., 2015) to verify the nonparametric estimation of the PA function for graph growth models. The latter log-linear form is a widely-used form for the PA function A_d of Edge PA with degree d, as described in Section 5.2. When applying the hypergraph generator of Algorithm 1 in Section 5.3.3, input parameters other than A_k were determined from STA-coauthor in order to generate realistic hypergraphs. Fig. 6.1 shows the estimation results for each of the generated hypergraphs. Without making any assumptions on the functional form of the PA function, each estimation result captured reasonably the shape of the corresponding true PA function.

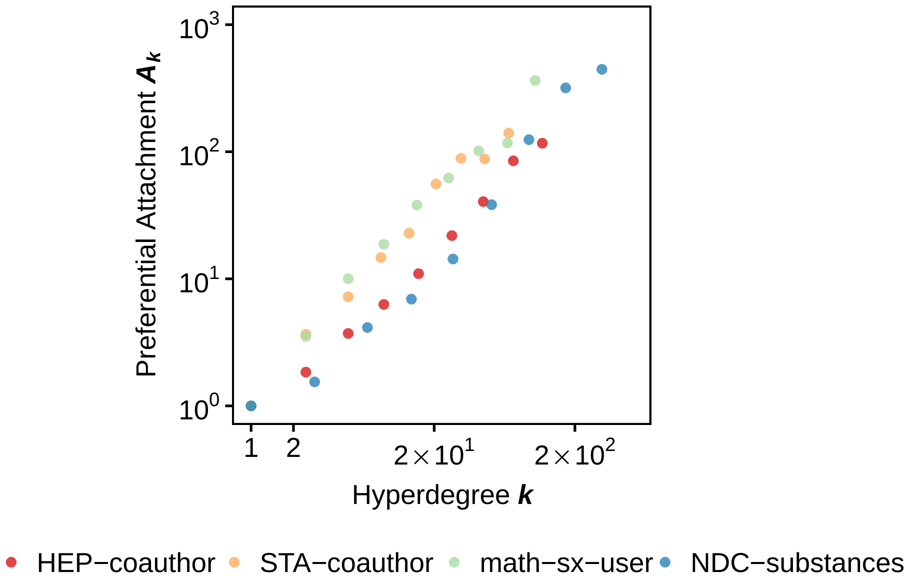
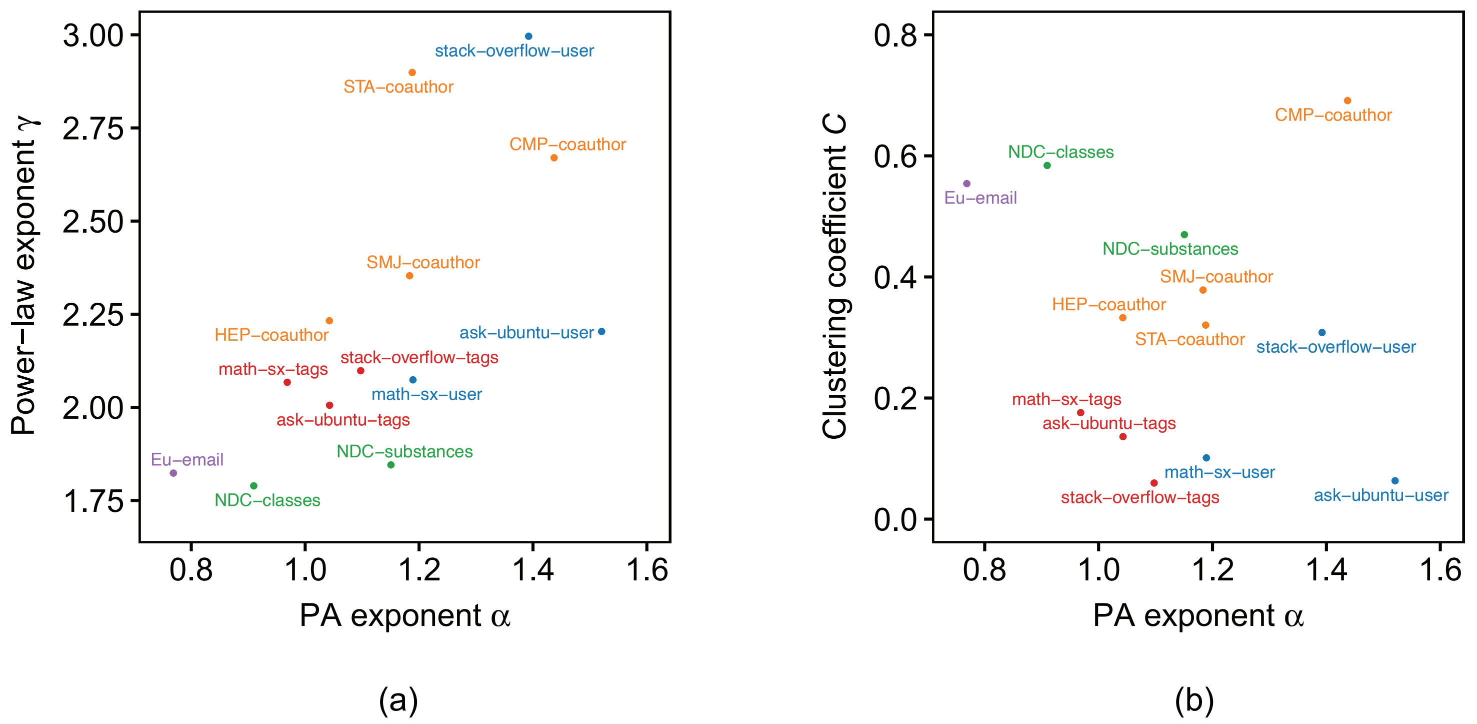


FIGURE 6.2: Nonparametrically estimated PA values of Hyper PA in four real-world datasets: HEP-coauthor, STA-coauthor, math-sx-user, and NDC-substances. The estimated A_k are generally increasing, which implies the existence of preferential attachment in hypergraph growth. PA exponent α calculated with the estimated PA values for all datasets including the above four datasets are given in Table 6.1.

We next estimated the PA function A_k by our proposed method in all datasets. Fig. 6.2 illustrates the nonparametrically estimated values of the PA functions of Hyper PA model in HEP-coauthor, STA-coauthor, math-sx-user, and NDC-substances. The nonparametrically estimated A_k in these datasets increase on average, which indicates the existence of the PA effect. Furthermore, they are substantially linear in log-log scale. Therefore, we also fitted the log-linear form $A_k = k^\alpha$ with hyperdegree k to the estimated A_k values and calculated the exponent α by the least-squares method. The values of PA exponent α of A_k for all datasets are given in Table 6.1. Since the estimated attachment exponents α are greater than 1 in all co-authorship datasets and thread participants datasets, the PA effect is superlinear in those datasets. And the values of α for the tagging datasets and NDC datasets were all in the range of 0.9 to 1.2, and 0.77 for Eu-Email. This result suggests the existence of the PA effect, particularly the strong PA effect in the co-authorship datasets and thread participants datasets. For example, in co-authorship data, the PA effect is that authors who have written more papers in the past are more likely to write new papers in the future.

In the 13 real-world datasets, we found that, while PA successfully captures first-order structures, it alone cannot explain the observed second-order structures. Fig. 6.3(a) shows a remarkably high correlation between the estimated attachment exponent α with the power-law exponent γ in the real-world datasets. There is a theoretical reason for this high correlation. In PA trees with $A_k = k^\alpha$, γ has been shown to be highly correlated with α when $\alpha \leq 1$ (Krapivsky & Redner, 2001). Extrapolating this result to our hypergraph-based growth model, it is reasonable to expect that when the average size of hyperedges is not large, the degree distribution of our model behaves similarly to that of a PA tree. This explains the observed high correlation between α and γ when α is around 1. However, given α, one cannot infer too much about the clustering coefficient C, as can be seen from the high variation of C in Fig. 6.3(b). This implies that PA alone cannot explain second-order structures, which is expected since PA is only a first-order mechanism. It is reasonable to expect that second-order structures, such as the clustering coefficient C, also depend



(a)

(b)

FIGURE 6.3: Estimated PA exponents α, power-law exponents γ, and clustering coefficients C in 13 datasets. (a): Estimated PA exponent α and power-law exponents γ. (b): Estimated PA exponent α clustering coefficient C. In each panel, the data points are colored according to the type of network. Datasets belonging to the same network type show similar trends in relationships between α and γ and between α and C. The PA mechanisms successfully captures first-order structures in the networks, as can be inferred from the high correlation between α and γ. However, PA alone is not enough to explain second-order structures, as can be seen from the low correlation between α and C.

on various higher-order growth factors, such as the distributions of sizes and numbers of hyperedges at each time-step.

6.1.3 Evaluation of Goodness-of-fit for Second-order Structures

In this section we perform additional experiments to compare the proposed Hyper PA model with some baseline models in reproducing second-order structures in the observed data. As in Section 5.2.2, in addition to Edge PA and Hyper PA models, we also consider the Hyper Uniform model. This special case of the Hyper PA model uniformly adds hyperedges, i.e., the PA function in Hyper Uniform is $A_k = 1$ for all hyperdegree k. As already described in Section 5.2.2, we will adopt a simulation-based approach to investigate the goodness-of-fit of the models. Specifically, we will generate networks from each model and compare several important statistics of the generated data to those of the real-world data. In each dataset, Hyper PA incorporates A_k estimated in the previous section, and Edge PA incorporates A_d obtained from a nonparametric estimation method (Pham et al., 2015). Since Edge PA generates graphs, we converted hypergraphs generated by Hyper PA and Hyper Uniform into graphs for comparison.

One of the most important graph properties often found in real-world networks is triangle-rich, which is manifested as a high value of the clustering coefficient (Bianconi et al., 2014; Newman, 2001a). We here examine the distribution of the number of triangles that each node has. We denote the number of triplets of node i as:

$$\Delta_i = \sum_{j,l} x_{i,j} x_{j,l} x_{l,i},$$

where $x_{i,j} = 1$ indicates the presence of edges between i and j, whereas $x_{i,j} = 0$ indicates the absence of edges. Fig. 6.4 shows the observed and simulated cumulative probabilities of the numbers of triplets in representative cases when Hyper PA succeeded and when it failed. When Hyper

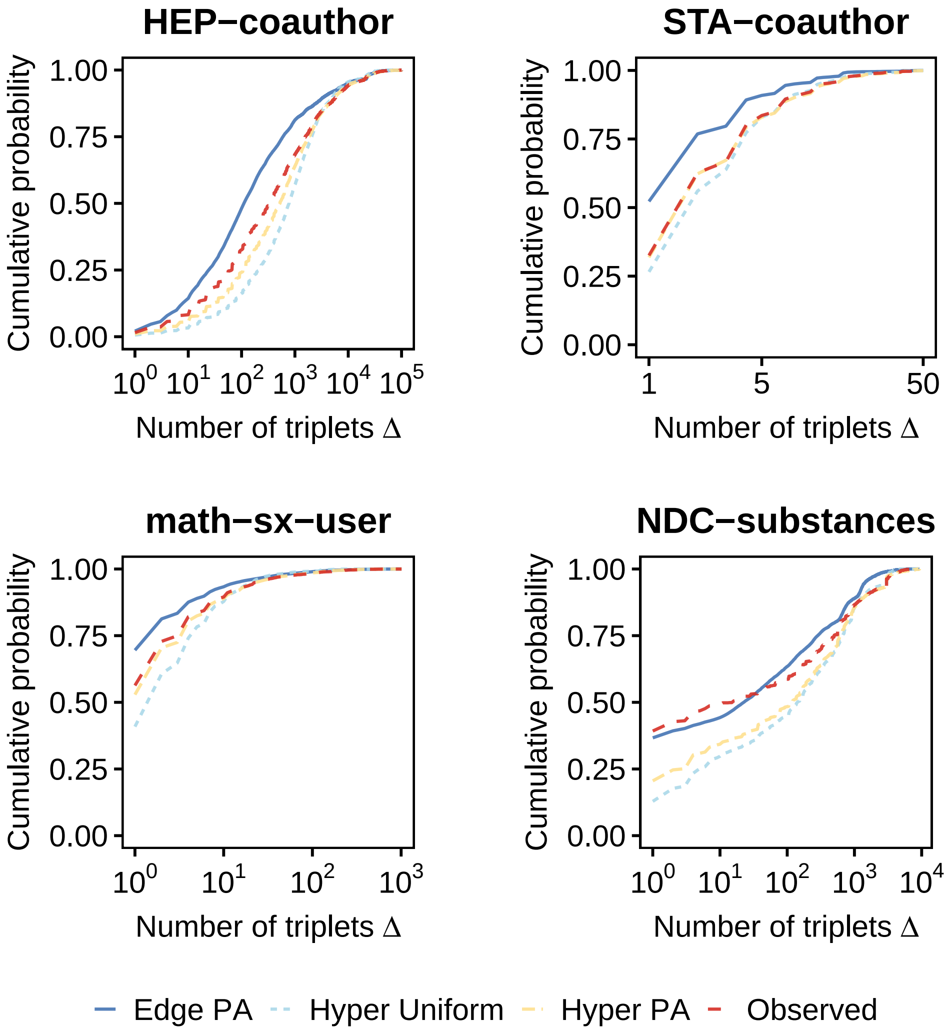


FIGURE 6.4: Observed and simulated cumulative probability distributions of the numbers of triplets in some representative datasets. For each of 13 datasets, we generate 10 graphs by Edge PA, and 10 hypergraphs by Hyper PA and Hyper Uniform, respectively. The generated hypergraphs are converted into graphs for comparison. The average values over 10 simulations are compared with the observed distributions. To illustrate, we show the observed and simulated distributions for four representative datasets: HEP-coauthor, STA-coauthor, math-sx-user, and NDC-substances. The quantitative comparison results for all datasets are given in Table 6.1. In datasets where Hyper PA is the best, Edge PA often over-estimates in the region of small numbers of triplets, as can be seen in HEP-coauthor, STA-coauthor, and math-sx-user. This is expected, since the independence of edges in Edge PA makes it more prone to produce nodes with a small number of triplets. For NDC-substances, Hyper PA and Hyper Uniform failed by underestimating in the region of low number of triplets, while Edge PA performed well.

PA succeeded, Edge PA often over-estimated the region of small number of triplets, which may be caused by the edge independence assumption in Edge PA. When Hyper PA failed, it often under-estimated the region of small number of triplets. Table 6.1 shows a quantitative comparison for all datasets using E^{triplet}, which is calculated as follows:

$$E^{\text{triplet}} = \frac{1}{N_{\text{bin}}} \sum_{n=1}^{N_{\text{bin}}} \left| p_{\text{obs}}(\Delta'_n) - p_{\text{sim}}(\Delta'_n) \right|,$$

where N_{bin} is the number of logarithmic bins, and p_{obs} and p_{sim} are the probability distributions of the numbers of triplets in real-world data and simulation data, respectively. We note that $\Delta'_1, \ldots, \Delta'_{N_{\text{bin}}}$ are the logarithmic binning of Δ, and we describe the result with $N_{\text{bin}} = 10$ in Table 6.1. Hyper PA provided the best fit in 10 datasets, whereas Edge PA and Hyper Uniform prevailed in the remaining three.

For closer inspection, we investigate the density of triangles around nodes, i.e., the local clustering coefficient. The high density of triangles in low-degree nodes is a signature property of many real-world networks. This property is also important since it may make practical tasks such as low-dimensional embedding more difficult (Seshadhri et al., 2020). The local clustering coefficient of node i with degree d_i is

$$C_i = \begin{cases} \frac{2\Delta_i}{d_i(d_i-1)} & (d_i \geq 2) \\ 0 & (d_i = 0, 1). \end{cases}$$

The average of C_i over all nodes in a graph is called the clustering coefficient. We analyze the distribution of C_i averaged over nodes that has the same degree d:

$$C(d) = \frac{1}{N_d} \sum_{i \in V_d} C_i, \tag{6.1}$$

where the set V_d is all nodes with degrees d in a graph, and N_d is the number of nodes in V_d. Fig. 6.5 shows the observed and simulated distributions $C(d)$ for some representative datasets. Hyper PA succeeded

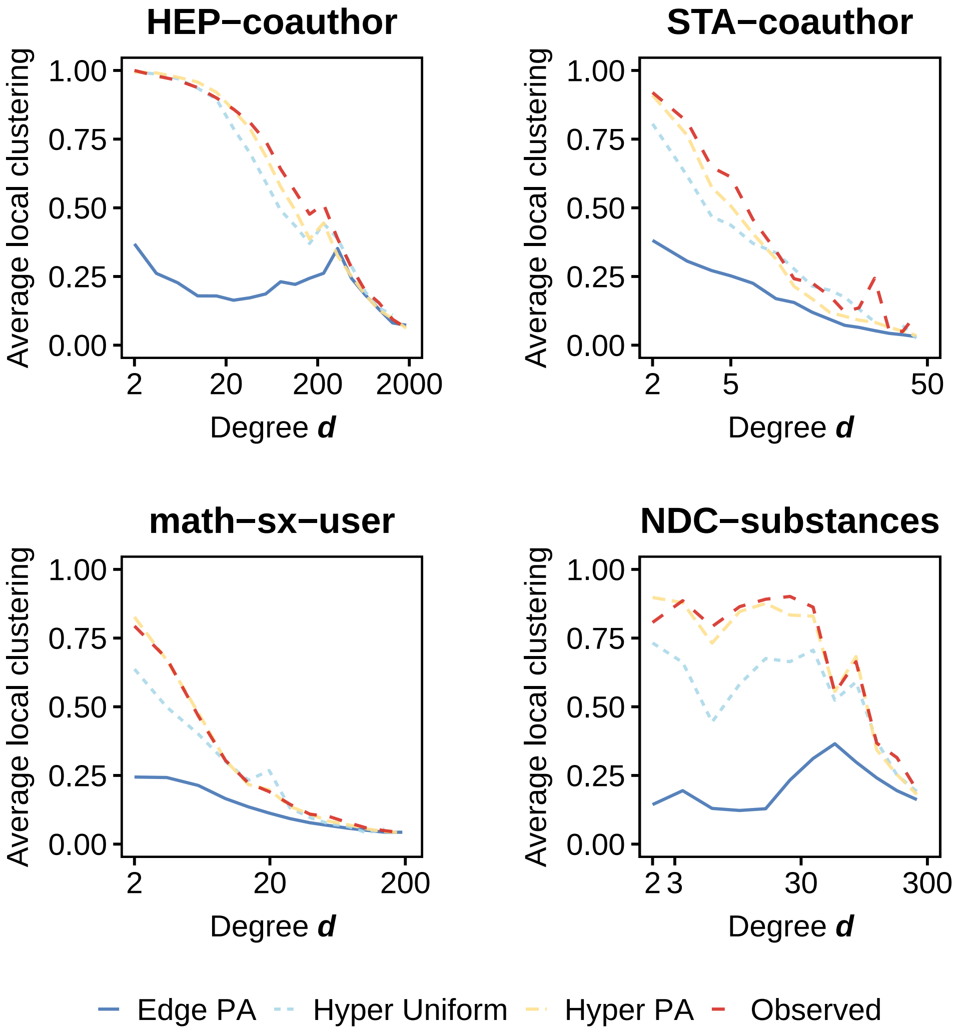


FIGURE 6.5: Observed and simulated local clustering coefficients averaged over nodes with degree d in some representative datasets. See Fig. 6.4 for simulation settings. To illustrate, we show the observed and simulated $C(d)$ for four representative datasets: HEP-coauthor, STA-coauthor, math-sx-user, and NDC-substances. The quantitative comparison results for all datasets are given in Table 6.1. In these four datasets, Hyper PA succeeded in replicating the signature decreasing of $C(d)$ when d increases. Edge PA often underestimated $C(d)$, especially in the region of low d.

in reproducing the signature decreasing of $C(d)$ when d increases, while

Edge PA under-estimated $C(d)$, especially in the region of low d. Table 6.1 shows a quantitative comparison in all the 13 datasets using E^{local}, which is calculated as follows:

$$E^{\text{local}} = \frac{1}{N_{\text{bin}}} \sum_{n=1}^{N_{\text{bin}}} \left| C_{\text{obs}}(d'_n) - C_{\text{sim}}(d'_n) \right| ,$$

where N_{bin} is the number of logarithmic bins, and C_{obs} and C_{sim} are (6.1) of observed and simulated data, respectively. We note that $d'_1, \ldots, d'_{N_{\text{bin}}}$ are the logarithmic binning of d, and we describe the result with $N_{\text{bin}} = 10$ in Table 6.1. Out of the 13 datasets, Hyper PA provided the best fit in all. Hyper Uniform prevailed over Edge PA in 11 datasets, in spite of the fact that Hyper Uniform uses a constant PA function, while Edge PA estimates the PA function from data. Even though the number of parameters in Hyper Uniform is zero, it has a better fit than Edge PA, which uses the estimation results as parameters. This suggests that the good fit of Hyper PA is more than just overfitting. These results highlight the importance of incorporating hyperedge information. Taking into accounts the results in Section 5.2.2, Hyper PA replicates well various first-order and second-order structures in all datasets.

6.1.4 Further Discussions

Some words are needed to bound the scope of our proposed hypergraph growth model. While our model does not allow the deletion of nodes and edges in the temporal network, it is indeed natural for nodes or edges to disappear in some network types. For example, an author may become inactive in co-authorship networks, while in relationship networks a relationship edge may be dissolved after some years. Our model also assumes that the PA function does not change with time. However, in co-authorship networks it is not unreasonable to expect that yearly advancements in communication and transportation technologies, which potentially ease how collaborations are formed and maintained, may make the

PA function change with time. Even for those types of networks, the proposed growth model can still be a viable approximation for the growth of the network in a short time span, e.g., five to ten years, where one can reasonably assume that the disappearances of nodes and edges, as well as any temporal change in the PA function, are negligible.

Our approach can potentially be used in predicting properties of a temporal network in the future. Some common network properties that are of potential interest are local properties such as degree and betweenness centrality of a node or global properties such as the diameter of a network (Yang et al., 2014). In principle, by using a probabilistic generative model such as our proposed hypergraph-based model, one can get information about any network property at a specific time in the future in the form of probability distributions. In order to do this, one uses the fitted model to generate multiple simulated networks at that specific time in the future, and calculates the empirical probability distribution of the property of interest in these simulated networks.

6.2 Conclusion

In this part, we have provided a proposed statistical method for estimating the preferential attachment in temporal hypergraphs. We also derived the conditional probability and the recursive formula that stabilize and accelerate the estimation on the hypergraph model. The analysis of the real-world datasets showed that the PA function of the hypergraph model had a similar form to that of the graph model in previous works. Furthermore, we demonstrated that our hypergraph PA growth model has advantages over conventional graph-based models in that it can better capture the first-order and second-order structures around each node.

Investigating the trade-offs of graph models, such as simplification by pairwise relationships, can provide valuable insight when considering which structures to choose for real-world complex systems: graphs, hypergraphs, or others. Future work includes more scrutiny of growth mechanisms in hypergraph models. In the case of functions using node features such as hyperdegree in our model, the computational complexity of the likelihood function can be similarly reduced by utilizing the proposed recursive formula. For example, the log-likelihood function can be efficiently computed when (5.2) is modified to

$$P_B(t) \propto \prod_{i \in B} A_{k_i(t)} f_i,$$

where f_i is the "fitness" parameter of node i (Pham et al., 2016). However, in the case of features that use dyadic relations, such as common neighbor nodes between a node pair, or features defined for a node set, the recursive formula can not be directly applied. Therefore, when extending the method to higher-order features, it will be necessary to solve the combinatorial computation problem, which hypergraph models often face.

Part III

Conclusion and Future Directions

Chapter 7

Conclusion

7.1 Summary

In this thesis, we have discussed the modeling of PA in complex networks which have local community structures and the statistical inference of the growth functions to address the issues of the conventional graph-based PA models. In this thesis, we have discussed the modeling and statistical inference of the growth functions of PA in complex networks which have local community structures. In Chapter 1, we provided the motivation and overview of this thesis. We then described the background knowledge of the network growth models related to Parts I and II in Chapter 2. In Parts I and II, we considered the following issues:

7.1.1 Issues

(1) The scale-free degree distribution and the high value of the clustering coefficient are often observed simultaneously in real-world complex networks. PA and transitivity, the classic and simple mechanisms, are widely used to explain the formation of the heavy tail of the degree distribution and the high clustering, respectively. Since one of the above simple mechanisms is not well suited to capture both features, many existing studies have attempted to reveal the driving force behind the formation of the two features by considering both PA and transitivity. The estimation of PA and transitivity in existing studies can be classified into the following two categories,

each of which has its own problems. Existing approaches either estimate one mechanism in isolation (Jeong et al., 2003; Newman, 2001a; Pham et al., 2015) or jointly estimate both mechanisms assuming some functional forms (Krivitsky & Handcock, 2019; Ripley et al., 2018). They each have the problems of poor fitting or risks losing the fine details of the two phenomena. Thus, statistically sound methods are needed to answer the questions: Do PA and transitivity co-exist in the growth of real-world complex networks? If they co-exist, how can we compare the effect of the two?

(2) Graphs have been widely utilized to represent pairwise interactions between individuals and their dynamics in various domains. However, in some real-world data, the collectivity of interactions is lost when expressed in graphs. Since group interactions such as co-authorship may contain more than two individuals, in graph expression, each of them is decomposed into multiple edges: the pre-specification of pairwise relationships. Most of the existing growth models for complex networks rely on graph representations and thus fail to capture the feature caused by group interactions in the growth process. Existing hypergraph models of temporal complex networks often employ some data-independent growth mechanism, which is the linear PA in most cases (Do et al., 2020; Lee et al., 2021; Liu et al., 2014). In principle, this pre-specification is undesirable since it completely ignores the data at hand. Thus, modelings which are free from the pre-specifications need to be considered to answer the questions: Is the graph-based growth model most suitable for representing networks with group interactions? Isn't there any drawback to graph-based PA that has still not been apparent?

To address the above issues, we provided the following contributions:

7.1.2 Contributions

(a) We discussed the issue (1) in Part I. We proposed a method for the non-parametric joint estimation of PA and transitivity in complex networks, as opposite to conventional methods that either estimate one mechanism in isolation or jointly estimate both assuming some functional forms. We also derived an efficient MM algorithm that iteratively updates the estimates. We apply our method to two scientific co-authorship networks: the authors in the Strategic Management Journal and the statistics field. Our non-parametric method revealed complex trends of PA and transitivity that would be unavailable under conventional parametric approaches. In both networks, having one common collaborator with another scientist increases at least 60 times the chance that one will collaborate with that scientist. Finally, by quantifying the contribution of each mechanism, we found that transitivity dominates preferential attachment in the two networks. We also developed a publicly available R package FoFaF (Inoue et al., 2020a).

(b) We discussed the issue (2) in Part II. We proposed a hypergraph approach for estimating the function that determines the growth of real-world hypergraphs and generating hypergraphs with the estimated functions. We used PA for our hypergraph growth model and presented a maximum likelihood estimation for the function. We analyzed 13 real-world networks, and the results suggest the existence of PA growth in real-world hypergraphs. We also found the advantages of combining PA with hypergraph growth in terms of the first-order and second-order structures. We also derived a recursive formula and a conditional probability that significantly reduce the computational cost of our model and the selection bias of newcomer nodes in analyzing hypergraphs, respectively. Even the special case of the proposed hypergraph model, which adds hyperedges to uniformly chosen nodes, outperformed the graph PA model in reproducing second-order structures of graphs. We also developed a publicly available R package HyperPA (Inoue et al., 2022).

7.2 Future Directions

Finally, we list the future directions for the modeling and statistical inference of the PA growth models discussed in this thesis.

Time-invariability Test of Growth Mechanisms

The first future direction is verifying the time-invariability of the growth functions. The growth models discussed in both Chapter 3 and Chapter 5 assume time-invariance of the growth function that determines PA and transitivity. In Section 4.1.4, we tested this assumption on real data and confirmed that the influence of time points on the estimation results is small for the two co-author network datasets used in the experiment. However, testing this on networks of various types or networks collected over longer time periods would help us to better understand the nature of growth mechanisms in the real-world. One way to perform this analysis is adapting the time-invariability test also used in stochastic actor-based models (Lospinoso et al., 2011).

Directed Networks

Secondly, there is future work to consider directions or orders in the interactions of network data. For example, in social network field, it is difficult to observe the direction of the co-authorship relationship in co-authorship networks, but the directions can be easily obtained from some domains such as citation networks of scientific papers or friendship networks on Twitter. In this thesis, we did not exploit the directions or orders of the interactions in the complex network data. That is, in the case of graphs, we considered undirected graphs, which do not have the directions on edges. In the case of hypergraphs, we considered undirected hypergraphs, where a set of nodes included in a hyperedge does not have any order. The PA functions of directed graphs have already been discussed in previous works since only two patterns, in-degree and out-degree, need to be considered. However, the estimation of the growth function of the following situations has not yet been discussed. The directed version of transitivity is more complicated than PA since all the direction of the edge among three nodes needs to be considered. In directed hypergraphs, estimating growth functions is more complex because one hyperedge may contain more than three nodes. Since a hyperedge can contain more than three nodes, even PA needs new frameworks for the growth of directed hypergraphs.

Generalizing the Growth Mechanism of Hypergraph Models

The third possible future direction is to consider growth mechanisms other than PA to the hypergraph growth mechanisms discussed in Part II. In our Hyper PA model (5.2), the probability of growth was defined only by the hyperdegree. If a feature θ_i is defined for each node i, its PA function A_θ can be estimated by the same procedure as Hyper PA, and this would be a promising extension. On the other hand, a feature θ'_{ij} defined between two nodes i, j, such as the number of common neighbors b_{ij} of transitivity in Chapter 3, cannot be directly adopted into our model. When using such features that are defined between two or more nodes, it is necessary to consider a new scheme of acceleration for maximum likelihood estimation, such as the recursive formula proposed in Section 5.3.1. Another challenging direction is to consider the hyperedge size m in the growth mechanisms defined for each node. This would require estimating a function of m for each node, which would be an advanced version of the fitness mechanism (Pham et al., 2016), and it would be difficult to estimate them. However, if this becomes possible, it is expected that the model can capture the growth characteristics between the mass and individual, for example, researchers who are good at single-author research and researchers who are good at collaborative research in co-authorship networks.

Bibliography

Baker, G. A. (1959). A New Derivation of Newton's Identities and their Application to the Calculation of the Eigenvalues of a Matrix. *Journal of the Society for Industrial and Applied Mathematics*, *7*(2), 143–148.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. https://doi.org/10.1126/science.286.5439.509

Benson, A. R., Abebe, R., Schaub, M. T., Jadbabaie, A., & Kleinberg, J. (2018). Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences*, *115*(48), E11221–E11230. https://doi.org/10.1073/pnas.1800683115

Bianconi, G., Darst, R. K., Iacovacci, J., & Fortunato, S. (2014). Triadic closure as a basic generating mechanism of communities in complex networks. *Physical Review. E*, *90*, 042806. https://doi.org/10.1103/PhysRevE.90.042806

Bornmann, L. (2017). Is collaboration among scientists related to the citation impact of papers because their quality increases with collaboration? An analysis based on data from f1000prime and normalized citation scores. *Journal of the Association for Information Science and Technology*, *68*(4), 1036–1047. https://doi.org/10.1002/asi.23728

Callaway, D. S., Hopcroft, J. E., Kleinberg, J. M., Newman, M. E. J., & Strogatz, S. H. (2001). Are randomly grown graphs really random? *Physical Review E*, *64*, 041902. https://doi.org/10.1103/PhysRevE.64.041902

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703. https://doi.org/10.1137/070710111

Conaldi, G., Lomi, A., & Tonellato, M. (2012). Dynamic models of affiliation and the network structure of problem solving in an open source software project. *Organizational Research Methods*, *15*(3), 385–412. https://doi.org/10.1177/1094428111430541

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, *52*(7), 558–569. https://doi.org/10.1002/asi.1097

Csárdi, G., Strandburg, K. J., Zalányi, L., Tobochnik, J., & Érdi, P. (2007). Modeling innovation by a kinetic description of the patent citation system. *Physica A: Statistical Mechanics and its Applications*, *374*(2), 783–793. https://doi.org/10.1016/j.physa.2006.08.022

David W. Johnson, K. A. S., Roger T. Johnson. (1991). *Active learning: Cooperation in the college classroom*. Interaction Book Company.

de Arruda, G. F., Petri, G., & Moreno, Y. (2020). Social contagion models on hypergraphs. *Physical Review Research*, *2*, 023032. https://doi.org/10.1103/PhysRevResearch.2.023032

de Arruda, G. F., Rodrigues, F. A., & Moreno, Y. (2018). Fundamentals of spreading processes in single and multilayer complex networks. *Physics Reports*, *756*, 1–59. https://doi.org/https://doi.org/10.1016/j.physrep.2018.06.007

Do, M. T., Yoon, S.-e., Hooi, B., & Shin, K. (2020). Structural patterns and generative models of real-world hypergraphs. *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining* (176–186). Association for Computing Machinery. https://doi.org/10.1145/3394486.3403060

D'Souza, R. M., Borgs, C., Chayes, J. T., Berger, N., & Kleinberg, R. D. (2007). Emergence of tempered preferential attachment from optimization. *Proceedings of the National Academy of Sciences*, *104*(15), 6112–6117. https://www.pnas.org/content/104/15/6112

Dugatkin, L. A. (1997). *Cooperation among animals: An evolutionary perspective*. Oxford University Press.

Ferligoj, A., Kronegger, L., Mali, F., Snijders, T. A. B., & Doreian, P. (2015). Scientific collaboration dynamics in a national scientific system. *Scientometrics*, *104*(3), 985–1012. https://doi.org/10.1007/s11192-015-1585-7

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., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science*, *359*(6379). https://doi.org/10.1126/science.aao0185

Godsil, C., & Royle, G. F. (2001). *Algebraic Graph Theory* (Vol. 207, Graduate Texts in Mathematics). Springer Science & Business Media.

Golosovsky, M., & Solomon, S. (2017). Growing complex network of citations of scientific papers: Modeling and measurements. *Physical Review E*, *95*, 012324. https://link.aps.org/doi/10.1103/PhysRevE.95.012324

Gómez, V., Kappen, H. J., & Kaltenbrunner, A. Modeling the structure and evolution of discussion cascades. In: *Proceedings of the 22nd ACM conference on hypertext and hypermedia*. HT '11. Eindhoven, The Netherlands: ACM, 2011, 181–190. ISBN: 978-1-4503-0256-2. https://doi.org/10.1145/1995966.1995992

Guimerà, R., Uzzi, B., Spiro, J., & Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science*, *308*(5722), 697–702. https://science.sciencemag.org/content/308/5722/697

Hamel, G., Doz, Y. L., & Prahalad, C. K. (1989). Collaborate with your competitors–and win. *Harvard Business Review*, *67*(1). https://hbr.org/1989/01/collaborate-with-your-competitors-and-win

Heider, F. (1946). Attitudes and cognitive organization [PMID: 21010780]. *The Journal of Psychology*, *21*(1), 107–112. https://doi.org/10.1080/00223980.1946.9917275

Holland, P. W., & Leinhardt, S. (1970). A method for detecting structure in sociometric data. *American Journal of Sociology*, *76*(3), 492–513. https://doi.org/10.1086/224954

Holland, P. W., & Leinhardt, S. (1971). Transitivity in structural models of small groups. *Comparative Group Studies*, *2*(2), 107–124. https://doi.org/10.1177/104649647100200201

Holland, P. W., & Leinhardt, S. (1976). Local structure in social networks. *Sociological Methodology*, *7*, 1–45. https://doi.org/10.2307/270703

Holland, P. W., & Leinhardt, S. (1977). A dynamic model for social networks. *The Journal of Mathematical Sociology*, *5*(1), 5–20. https://doi.org/10.1080/0022250X.1977.9989862

Holme, P., & Kim, B. J. (2002). Growing scale-free networks with tunable clustering. *Phys. Rev. E*, *65*, 026107. https://doi.org/10.1103/PhysRevE.65.026107

Hu, X., Rousseau, R., & Chen, J. (2010). In those fields where multiple authorship is the rule, the h-index should be supplemented by role-based h-indices. *Journal of Information Science*, *36*(1), 73–85. https://doi.org/10.1177/0165551509348133

Hunter, D. R., & Lange, K. (2000). Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics*, 60–77. https://doi.org/10.2307/1390613

Hunter, D. R., & Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, *58*, 30–37. https://doi.org/10.1198/0003130042836

Hunter, D. R., Goodreau, S. M., & Handcock, M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association*, *103*(481), 248–258. https://doi.org/10.1198/016214507000000446

Inoue, M., Pham, T., & Shimodaira, H. Transitivity vs Preferential Attachment: Determining the Driving Force Behind the Evolution of Scientific Co-Authorship Networks. In: *Unifying Themes in Complex Systems IX. ICCS 2018. Springer Proceedings in Complexity.* Springer, 2018, 262–271. https://doi.org/10.1007/978-3-319-96661-8_28.

Inoue, M., Pham, T., & Shimodaira, H. (2020a). FoFaF [R package version 0.1.0]. https://github.com/minoue637/FoFaF

Inoue, M., Pham, T., & Shimodaira, H. (2020b). Joint estimation of nonparametric transitivity and preferential attachment functions in scientific co-authorship networks. *Journal of Informetrics*, *14*(3), 101042. https://doi.org/10.1016/j.joi.2020.101042

Inoue, M., Pham, T., & Shimodaira, H. (2022). A hypergraph approach for estimating growth mechanisms of complex networks [Early Access Version]. *IEEE Access*. https://doi.org/10.1109/ACCESS.2022.3143612

Jeong, H., Néda, Z., & Barabási, A.-L. (2003). Measuring preferential attachment in evolving networks. *Europhysics Letters*, *61*(4), 567–572.

Ji, P., & Jin, J. (2016). Coauthorship and citation networks for statisticians. *Ann. Appl. Stat.*, *10*(4), 1779–1812. https://doi.org/10.1214/15-AOAS896

Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: Shifting impact, geography, and stratification in science. *Science*, *322*(5905), 1259–1262. https://doi.org/10.1126/science.1158357

Kahn, M. (2017). Co-authorship as a proxy for collaboration: a cautionary tale. *Science and Public Policy*, *45*(1), 117–123. https://doi.org/10.1093/scipol/scx052

Kamiński, B., Poulin, V., Prałat, P., Szufel, P., & Théberge, F. (2019). Clustering via hypergraph modularity. *PLOS ONE*, *14*(11), 1–15. https://doi.org/10.1371/journal.pone.0224307

Kong, J. S., Sarshar, N., & Roychowdhury, V. P. (2008). Experience versus Talent Shapes the Structure of the Web. *Proceedings of the National Academy of Sciences of the USA*, *37*, 105.

Krapivsky, P. L., & Redner, S. (2001). Organization of growing random networks. *Physical Review. E*, *63*(6), 066123.

Krapivsky, P. L., Rodgers, G. J., & Redner, S. (2001). Degree distributions of growing networks. *Physical Review Letters*, *86*(23), 5401–5404.

Krivitsky, P. N., & Handcock, M. S. (2019). *Tergm: Fit, simulate and diagnose models for network evolution based on exponential-family random graph models* [R package version 3.6.0]. The Statnet Project (https://statnet.org). https://CRAN.R-project.org/package=tergm

Kronegger, L., Mali, F., Ferligoj, A., & Doreian, P. (2012). Collaboration structures in slovenian scientific communities. *Scientometrics*, *90*(2), 631–647. https://doi.org/10.1007/s11192-011-0493-8

Kunegis, J. Konect: The koblenz network collection. In: *Proceedings of the 22nd international conference on world wide web*. WWW '13 Companion. Rio de Janeiro, Brazil: Association for Computing Machinery, 2013, 1343–1350. ISBN: 9781450320382. https://doi.org/10.1145/2487788.2488173.

Kunegis, J., Blattner, M., & Moser, C. Preferential attachment in online networks: Measurement and explanations. In: *Proceedings of the 5th annual acm web science conference*. ACM. 2013, 205–214.

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2015). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology*, *66*(7), 1323–1332. https://doi.org/10.1002/asi.23266

Lee, G., Choe, M., & Shin, K. How do hyperedges overlap in real-world hypergraphs? - patterns, measures, and generators. In: *Proceedings of the web conference 2021*. WWW '21. Ljubljana, Slovenia: Association for Computing Machinery, 2021, 3396–3407. ISBN: 9781450383127. https://doi.org/10.1145/3442381.3450010.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, *58*(7), 1019–1031. https://doi.org/10.1002/asi.20591

Liu, J.-G., Yang, G.-Y., & Hu, Z.-L. (2014). A knowledge generation model via the hypernetwork. *PLOS ONE*, *9*(3), 1–8. https://doi.org/10.1371/journal.pone.0089746

Lospinoso, J. A. (2012). *Statistical models for social network dynamics* (Doctoral dissertation). Oxford University. UK. https://ora.ox.ac.uk/objects/ora:6726

Lospinoso, J. A., Schweinberger, M., Snijders, T. A. B., & Ripley, R. M. (2011). Assessing and accounting for time heterogeneity in stochastic actor oriented models. *Advances in Data Analysis and Classification*, *5*(2), 147–176. https://doi.org/10.1007/s11634-010-0076-1

Lung, R. I., Gaskó, N., & Suciu, M. A. (2018). A hypergraph model for representing scientific output. *Scientometrics*, *117*(3), 1361–1379.

Massen, C., & Jonathan, P. (2007). Preferential attachment during the evolution of a potential energy landscape. *The Journal of Chemical Physics*, *127*, 114306. https://doi.org/10.1063/1.2773721

Medo, M. (2014). Statistical validation of high-dimensional models of growing networks. *Physical Review E*, *89*, 032801. https://doi.org/10.1103/PhysRevE.89.032801

Medo, M., Cimini, G., & Gualdi, S. (2011). Temporal effects in the growth of networks. *Physical Review Letter*, *107*, 238701. https://doi.org/10.1103/PhysRevLett.107.238701

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56

Michoel, T., & Nachtergaele, B. (2012). Alignment and integration of complex networks by hypergraph-based spectral clustering. *Physical Review. E*, *86*, 056111. https://doi.org/10.1103/PhysRevE.86.056111

Milojević, S. (2010). Modes of collaboration in modern science: Beyond power laws and preferential attachment. *Journal of the American Society for Information Science and Technology*, *61*(7), 1410–1423. https://doi.org/10.1002/asi.21331

Mithani, A., Preston, G. M., & Hein, J. (2009). Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison. *Bioinformatics*, *25*(14), 1831–1832. https://doi.org/10.1093/bioinformatics/btp269

Newman, M. E. J. (2001a). Clustering and preferential attachment in growing networks. *Physical Review E*, *64*(2), 025102. https://doi.org/10.1103/PhysRevE.64.025102

Newman, M. E. J. (2001b). Scientific collaboration networks. i. network construction and fundamental results. *Phys. Rev. E*, *64*, 016131. https://doi.org/10.1103/PhysRevE.64.016131

Newman, M. E. J. (2001c). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404–409. https://doi.org/10.1073/pnas.98.2.404

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5200–5205. https://doi.org/10.1073/pnas.0307545100

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*, 323–351. https://doi.org/10.1080/00107510500052444

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560–1563. https://doi.org/10.1126/science.1133755

Overgoor, J., Benson, A., & Ugander, J. Choosing to grow a graph: Modeling network formation as discrete choice. In: *The world wide web conference*. WWW '19. San Francisco, CA, USA: Association for Computing Machinery, 2019, 1409–1420. ISBN: 9781450366748. https://doi.org/10.1145/3308558.3313662.

Pham, T., Sheridan, P., & Shimodaira, H. (2015). PAFit: A Statistical Method for Measuring Preferential Attachment in Temporal Complex Networks. *PLOS ONE*, *10*(9), e0137796. https://doi.org/10.1371/journal.pone.0137796

Pham, T., Sheridan, P., & Shimodaira, H. (2016). Joint estimation of preferential attachment and node fitness in growing complex networks. *Scientific Reports*, *6*, 32558. https://doi.org/10.1038/srep32558

Pham, T., Sheridan, P., & Shimodaira, H. (2020). PAFit: An R package for estimating preferential attachment and node fitness in temporal complex networks. *Journal of Statistical Software*, *92*, 1–30. https://doi.org/10.18637/jss.v092.i03

Price, D. D. S. (1965). Networks of scientific papers. *Science*, *149*(3683), 510–515. https://doi.org/10.1126/science.149.3683.510

Price, D. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306. https://doi.org/10.1002/asi.4630270505

Riolo, M. A., & Newman, M. E. J. (2020). Consistency of community structure in complex networks. *Physical Review. E*, *101*, 052306. https://doi.org/10.1103/PhysRevE.101.052306

Ripley, R. M., Snijders, T. A., Boda, Z., Vörös, A., & Preciado, P. (2018). *Manual for siena version 4.0 (version may 24, 2018)* [http://www.stats.ox.ac.uk/~snijders/siena/]. Oxford: University of Oxford, Department of Statistics; Nuffield College.

Ronda-Pupo, G. A., & Pham, T. (2018). The evolutions of the rich get richer and the fit get richer phenomena in scholarly networks: The case of the strategic management journal. *Scientometrics*, *116*(1), 363–383. https://doi.org/10.1007/s11192-018-2761-3

Seshadhri, C., Sharma, A., Stolman, A., & Goel, A. (2020). The impossibility of low-rank representations for triangle-rich complex networks. *Proceedings of the National Academy of Sciences*, *117*(11), 5631–5637. https://doi.org/10.1073/pnas.1911030117

Sheridan, P., Yagahara, Y., & Shimodaira, H. (2012). Measuring preferential attachment in growing networks with missing-timelines using markov chain monte carlo. *Physica A: Statistical Mechanics and its Applications*, *391*(20), 5031–5040. https://doi.org/https://doi.org/10.1016/j.physa.2012.05.041

Simkin, M. V., & Roychowdhury, V. P. (2007). A mathematical theory of citing. *Journal of the American Society for Information Science and Technology*, *58*(11), 1661–1673. https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20653

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, *42*(3/4), 425–440. https://doi.org/10.1093/biomet/42.3-4.425

Snijders, T. A. (2001). The statistical evaluation of social network dynamics. *Sociological Methodology*, *31*(1), 361–395. https://doi.org/10.1111/0081-1750.00099

Snijders, T. A. (2017). Stochastic actor-oriented models for network dynamics. *Annual Review of Statistics and Its Application*, *4*(1), 343–363. https://doi.org/10.1146/annurev-statistics-060116-054035

Spitz, A., Aumiller, D., Soproni, B., & Gertz, M. A versatile hypergraph model for document collections. In: *32nd international conference on scientific and statistical database management*. SSDBM 2020. Vienna, Austria: Association for Computing Machinery, 2020. ISBN: 9781450388146. https://doi.org/10.1145/3400903.3400919.

Sun, X., Yin, H., Liu, B., Chen, H., Cao, J., Shao, Y., & Viet Hung, N. Q. Heterogeneous hypergraph embedding for graph classification. In: *Proceedings of the 14th acm international conference on web search and data mining*. WSDM '21. Virtual Event, Israel: Association for Computing Machinery, 2021, 725–733. ISBN: 9781450382977. https://doi.org/10.1145/3437963.3441835.

Tahai, A., & Meyer, M. J. (1999). A revealed preference study of management journals' direct influences. *Strategic Management Journal*, *20*(3), 279–296. https://doi.org/10.1002/(SICI)1097-0266(199903)20:3<279::AID-SMJ33>3.0.CO;2-2

Wang, M., Yu, G., & Yu, D. (2008). Measuring the preferential attachment mechanism in citation networks. *Physica A: Statistical Mechanics and its Applications*, *387*(18), 4692–4698. https://doi.org/10.1016/j.physa.2008.03.017

Wang, X., Wang, Z., & Shen, H. (2019). Dynamical analysis of a discrete-time sis epidemic model on complex networks. *Applied Mathematics Letters*, *94*, 292–299. https://doi.org/https://doi.org/10.1016/j.aml.2019.03.011

Watson, A. (1984). *Diplomacy*. London: Routledge.

Yang, Y., Dong, Y., & Chawla, N. V. (2014). Predicting node degree centrality with the node prominence profile. *Scientific Reports*, *4*(1), 7236. https://doi.org/10.1038/srep07236

Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis,F.R.S. *Philosophical Transactions of the Royal*

Society of London B: Biological Sciences, *213*(402–410), 21–87. https://doi.org/10.1098/rstb.1925.0002

Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., & Stanley, H. E. (2017). The science of science: From the perspective of complex systems. *Physics Reports*, *714*, 1–73.

Zheng, X., Luo, Y., Sun, L., Ding, X., & Zhang, J. (2018). A novel social network hybrid recommender system based on hypergraph topologic structure. *World Wide Web*, *21*(4), 985–1013.

Zinilli, A. (2016). Competitive project funding and dynamic complex networks: Evidence from Projects of National Interest (PRIN). *Scientometrics*, *108*(2), 633–652. https://doi.org/10.1007/s11192-016-1976-4