IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 30, 2021, accepted December 24, 2021, date of publication January 14, 2022, date of current version April 5, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3143612

A Hypergraph Approach for Estimating Growth Mechanisms of Complex Networks

MASAAKI INOUE[1,2], THONG PHAM[2], AND HIDETOSHI SHIMODAIRA[1,2]
[1]Department of Systems Science, Kyoto University, Kyoto 6068501, Japan
[2]RIKEN Center for Advanced Intelligence Project, Tokyo 1030027, Japan

Corresponding author: Masaaki Inoue (minoue637@gmail.com)

The work of Thong Pham was supported by the Japan Society for the Promotion of Science KAKENHI under Grant JP19K20231.
The work of Hidetoshi Shimodaira was supported by the Japan Society for the Promotion of Science KAKENHI under Grant JP20H04148.

ABSTRACT Temporal datasets that describe complex interactions between individuals over time are increasingly common in various domains. Conventional graph representations of such datasets may lead to information loss since higher-order relationships between more than two individuals must be broken into multiple pairwise relationships in graph representations. In those cases, a hypergraph representation is preferable since it can preserve higher-order relationships by using hyperedges. However, existing hypergraph models of temporal complex networks often employ some data-independent growth mechanism, which is the linear preferential attachment in most cases. In principle, this pre-specification is undesirable since it completely ignores the data at hand. Our work proposes a new hypergraph growth model with a data-driven preferential attachment 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 preferential attachment functions deviates substantially from the linear form. This demonstrates the need of doing away with the linear preferential attachment 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.

INDEX TERMS Co-authorship networks, complex networks, hypergraphs, preferential attachment, selection bias.

I. 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 [1]–[3]. 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 [4]. 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 [5]. 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 associate editor coordinating the review of this manuscript and approving it for publication was Md Asaduzzaman.

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 [6], bioinformatics [7], classification [8], clustering [9], and document retrieval [10].

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) [11]. 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 [12]–[16].

Of the few existing works on hypergraph-based growth models, most employ a data-independent growth mechanism which is the linear preferential attachment [17]–[19]. 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 [13].

In this work, 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 paper can be summarized as follows:

- 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
$$\boldsymbol{A} = [A_1, A_2, A_3, \ldots] \tag{1}$$
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. A publicly available software of the proposed method will be published on Code Ocean.
- 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 principle way. We note that this approach can also be applied to graph-based growth models.
- 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.

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 [20]. 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.

The rest of this paper is organized as follows: Section II describes temporal hypergraphs comparing with temporal graphs. Section III introduces related graph-based PA models and our hypergraph-based approach and presents illustrative results of our proposed model. Section IV provides our estimation methodology and the hypergraph generation algorithm for our growth model. Section V explores the performance of the proposed method in 13 real-world datasets by comparing it with the conventional method. Finally, Section VI concludes this work and outlines future work.

II. TEMPORAL HYPERGRAPH MODELS

In this section, we explain some properties of discrete-time hypergraph-based growth models, comparing it with conventional growth models of ordinary graphs. 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. Although a hypergraph where the set of nodes in a hyperedge does not have any order is called an "undirected hypergraph" [21], in this paper we simply refer to it as a "hypergraph". Similarly, we refer to "undirected graph without self-loops" as "graph".

Fig. 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. 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. 1.

Guimerà *et al.* [22] 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.

Next we 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. 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. 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 [23], [24], few studies have examined whether this independence assumption is appropriate under such large values of m. The datasets used in the experiments in Sections III and V of this paper contain hyperedges whose m are greater than 100. Using these datasets, we will investigate the performances 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 [25]. 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 this study is that it does not require the above independence assumption throughout the growth process. In the next

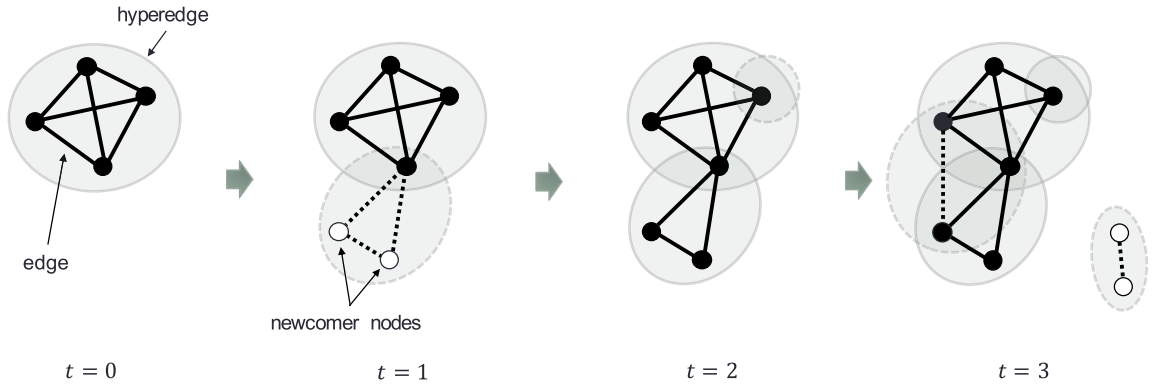


FIGURE 1. Hypergraph expression and graph expression of temporal data. Whereas a graph consists of nodes and edges, a hypergraph consists of nodes and hyperedges. Hyperedges and edges are indicated by ovals and line segments, respectively. Dashed ovals and line segments represent newly added hyperedges and edges at each time-step, respectively. White nodes represent newcomer nodes.

section, we describe our proposed approach to capture the characteristics of temporal hypergraphs.

III. PROPOSED APPROACH

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

A. GRAPH-BASED GROWTH MODELS WITH PA

We describe an undirected graph version of the General Temporal (GT) model with PA [13]. This model is a generalization of various classical models such as Barabási-Albert model [11] and Price's model [26]. In the GT model, the probability that a node pair i, j will acquire an edge at time t is expressed as:

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

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 [27], maximum likelihood estimation methods [28], and methods based on Markov chain Monte Carlo [29]. 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 [30], [31] and maximum likelihood estimation [13].

B. PROPOSED HYPERGRAPH GROWTH MODEL

We propose a hypergraph version of the PA model based on the GT model. 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:

$$P_B(t) \propto \prod_{i \in B} A_{k_i(t)}, \tag{2}$$

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, we assume no functional form for the PA function A_k. Do *et al.* [17] 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 of hyperedges is large, i.e., $|B| \gg 1$, the value of k_B becomes sparse, which 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 $m - 1$, which is equal to 2 for the case of $m = 3$ above, in Edge PA 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 IV-A and IV-B. We can also generate hypergraphs with a given PA function A_k by a procedure provided in Section IV-C.

C. 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 [32] and HEP-coauthor from the high energy physics field [33], [34]. The details of these datasets are provided in Section V. 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 [13]. 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 with the graphs generated by Edge PA, the hypergraphs generated by Hyper PA were converted into graphs. The detail of proposed estimation method and the procedure for generating hypergraph is described in Section IV.

Fig. 2(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 V-C 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. 2(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. 2(b), HEP-coauthor contains many relatively large hyperedges. The maximum hyperedge size is 201 for the HEP-coauthor and 10 in for the STA-coauthor. More detailed data descriptions for all datasets used in this paper are provided in Section V-A. 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. 2(c) illustrates this point.

In Fig. 2(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. 2(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 V-C, 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. 2(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 comparing with conventional models. Further experiments are provided in Section V.

IV. METHOD

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 principle approach based on conditional probabilities for handling a selection bias that

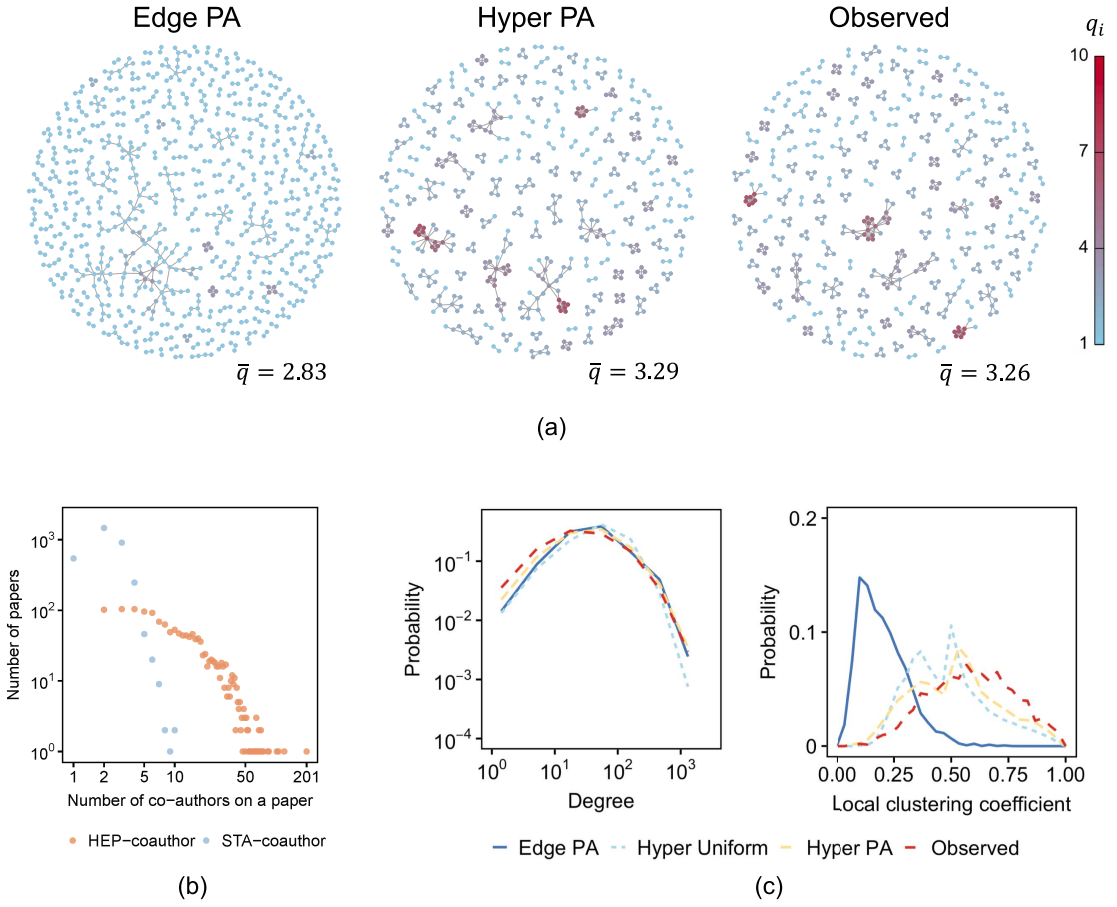


FIGURE 2. Hyper PA outperforms conventional models in reproducing first-order and second-order structures in scientific co-authorship data. (a): Observed and simulated graphs of STA-coauthor, a dataset of co-authorships in journals from statistics field. Each graph illustrates the final 8% increments 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.

arises in modeling the emergence of new hyperedges with newcomer nodes.

A. MAXIMUM LIKELIHOOD ESTIMATION

1) LIKELIHOOD FUNCTION

We first derive the likelihood function of $\boldsymbol{A}$ for Hyper PA model. As we described in Section III, our growth model (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 [13], [14], [16] 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 ($1 \le l \le h_t$) hyperedge at time-step t, its size is given by $m_{t,l}$, and we have $0 \le n_{t,l} \le m_{t,l}$; 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:

$$P_{B_{t,l}}(t) = \frac{\prod_{i \in B_{t,l}} A_{k_i(t)}}{\sum_{B' \in \mathcal{C}_{m_{t,l}}(V_t)} \prod_{i' \in B'} A_{k_{i'}(t)}}, \tag{3}$$

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 IV-B.

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 $\boldsymbol{\theta}_{init}$. As in previous works [13], [14], [16], we assume that $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{init}$ 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:

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

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

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

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 $\boldsymbol{\theta}_{init}$ and $\boldsymbol{\theta}_t$. The log-likelihood function can then be rewritten as follows:

$$\begin{aligned}
&L(\boldsymbol{A}|G_0, \ldots, G_T) \\
&= \sum_{t=1}^{T} \log P(G_t|G_{t-1}, h_t, \boldsymbol{m}_t, \boldsymbol{n}_t, \boldsymbol{A}) \\
&= \sum_{t=1}^{T} \sum_{l=1}^{h_t} \log P_{B_{t,l}}(t) \qquad (4) \\
&= \sum_{t=1}^{T} \sum_{l=1}^{h_t} \log \prod_{i \in B_{t,l}} A_{k_i(t)} \\
&\quad - \sum_{t=1}^{T} \sum_{l=1}^{h_t} \log \left(\sum_{B' \in \mathcal{C}_{m_{t,l}}(V_t)} \prod_{i \in B'} A_{k_i(t)} \right). \qquad (5)
\end{aligned}$$

Note that we substituted (3) into (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 (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" [13], [14], [16] of k, where we set $A_{k+1} = A_k$ in groups of k values.

Since (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 (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 $N(t)$.

2) A RECURSIVE FORMULA FOR FAST COMPUTATION OF THE NORMALIZING FACTOR

A fast computation of the normalizing factor

$$S_m(t) = \sum_{B' \in \mathcal{C}_m(V_t)} \prod_{i \in B'} A_{k_i(t)} \qquad (6)$$

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 (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 $e_m(x_1, \ldots, x_n)$ $(0 \le m \le n)$ with variables $x_1, \ldots, x_n$ as follows. For $m = 0$, $e_0(x_1, \ldots, x_n) = 0$, and for $m > 0$,

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

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 [35], we have:

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

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

$$S_m(t) = \frac{1}{m} \sum_{j=1}^{m} (-1)^{j-1} S_{m-j}(t) p_j(A_{k_1(t)}, \ldots, A_{k_{N(t)}(t)}). \qquad (7)$$

Note that $m \le N(t)$ 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 dominating in (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 $\mathcal{O}\left(m^2N(t)\right)$. By utilizing (7), the log-likelihood function given in (5) can be optimized by standard numerical methods such as quasi-Newton methods or MM algorithms [13].

B. 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 of a similar bias occurred in graph-based models [13], [14], 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 (4). Since the proportion of new hyperedges with newcomer nodes can be high in many real-world data [22], 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, 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 $V_{new} = V_t \setminus V_{t-1}$, $V_{exist} = V_{t-1}$. With these new notations, note that $B = B_1 \cup B_2$, $B_1 \subset V_{new}$, and $B_2 \subset V_{exist}$. 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 $B^* \subset V_{new}$. This conditional probability can be written as:

$$P_{B|B_1=B^*}(t) = \frac{\prod_{i\in B_2} A_{k_i(t)}}{\sum_{B_2'\in\mathcal{C}_{m'}(V_{exist})}\prod_{i'\in B_2'} A_{k_{i'}(t)}}, \tag{8}$$

which is essentially equivalent to (3) but applied to B_2 part. In other words, we simply ignore B_1 part. In calculating (4), if the observed node set $B_{t,l}$ contains only existing nodes, one uses (3), whereas if it contains some newcomer nodes, one uses (8). Note that in (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 (8) can also be accelerated by the recursive formula (7). See Appendix for a derivation of (8).

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: $\{\boldsymbol{m}_t\}_{t=1}^T$, $\boldsymbol{m}_t = \left[m_{t,1}, \ldots, m_{t,h_t}\right]$
(vi) sequence of the number of emerging nodes: $\{\boldsymbol{n}_t\}_{t=1}^T$, $\boldsymbol{n}_t = \left[n_{t,1}, \ldots, n_{t,h_t}\right]$

Output: evolving hypergraph: $\{G_t\}_{t=1}^T$, $G_t = (V_t, E_t)$

1: **for** time t in $[1, \ldots, T]$ **do**
2: set $V_t \leftarrow V_{t-1}$ and $E_t \leftarrow E_{t-1}$
3: **for** i in $[1, \ldots, h_t]$ **do**
4: $v_i^{new} \leftarrow$ set $n_{t,i}$ newcomer nodes
5: $v_i^{exist} \leftarrow$ sample $m_{t,i} - n_{t,i}$ nodes from V_{t-1} by HyperPA($\boldsymbol{A}, G_{t-1}, m_{t,i}, n_{t,i}$)
6: $e_i \leftarrow$ set a hyperedge containing $v_i^{new} \cup v_i^{exist}$
7: add v_i^{new} to V_t and e_i to E_t
8: **end for**
9: $G_t \leftarrow (V_t, E_t)$
10: **end for**
11: **return** $\{G_t\}_{t=1}^T$

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

C. 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 iteration of t 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 (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.

V. 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.

A. 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) [36], High Energy Physics (HEP-coauthor) [33], [34], Strategic Management Journal (SMJ-coauthor) [37], and Statistics (STA-coauthor) [32].
- **Online thread participants datasets**: Each node represents a user answering questions on threads and each hyperedge describes a set 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.* [38].

Some preprocessing is needed before one can perform model fitting. For each dataset in Benson *et al.* [38], we extracted the latest 5000 hyperedges for analysis. In addition, several datasets with too few or too many nodes extracted from Benson *et al.* [38] 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

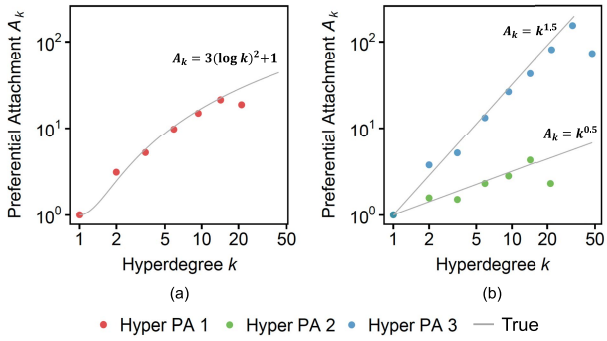


FIGURE 3. Our proposed method can estimate the PA function A_k from observed temporal hypergraphs without any assumptions on the functional form of A_k. We generated synthetic hypergraphs from the Hyper PA model, and applied our method to recover the PA function from the simulated data. (a): Hyper PA 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.

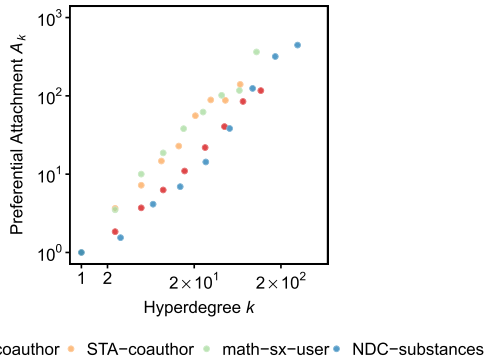


FIGURE 4. Nonparametrically estimated PA values of Hyper PA in four real-world datasets: HEP-coauthor, STA-coauthor, math-sx-user, and NDC-substances. The estimated A_k are generally increasing, which implies the existence of preferential attachment in hypergraph growth. PA exponent α calculated with the estimated PA values for all datasets including the above four datasets are given in Table 1.

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 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 [39], and hence the average number of co-authors per paper (i.e., the average size of hyperedges) is larger than the other datasets.

B. PREFERENTIAL ATTACHMENT IN 13 DATASETS

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

TABLE 1. Summary statistics and experiment results for 13 datasets. Summary statistics of the datasets described in Section V-A; 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 V-B. 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 V-C. 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 V-C. 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**

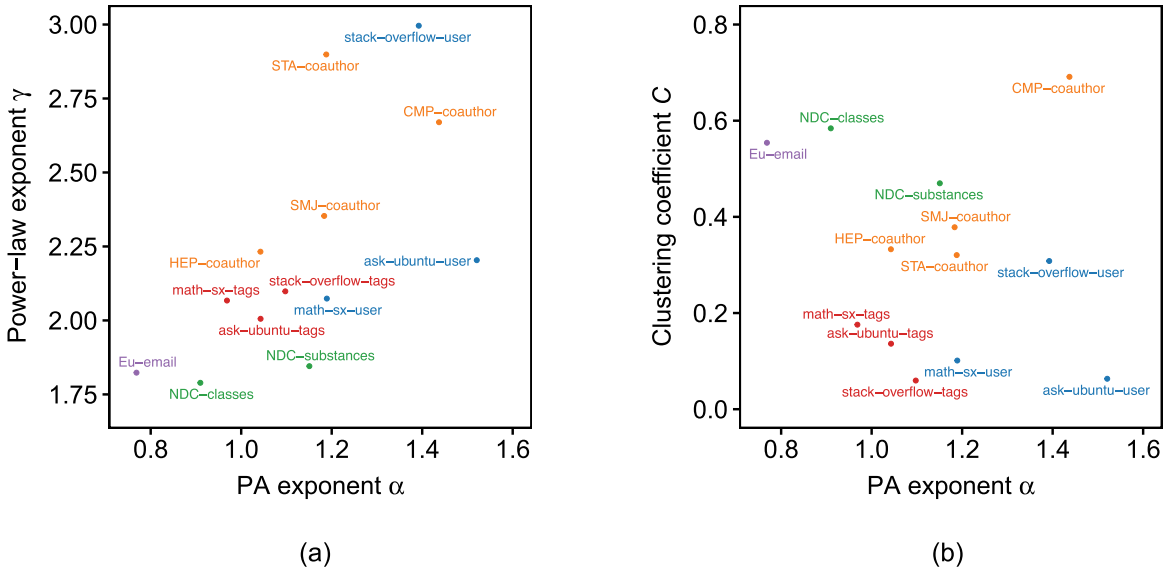


FIGURE 5. Estimated PA exponents α, power-law exponents γ, and clustering coefficients C in 13 datasets. (a): Estimated PA exponent α and power-law exponents γ. (b): Estimated PA exponent α clustering coefficient C. In each panel, the data points are colored according to the type of network. Datasets 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.

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 [13], [16] 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 III. When applying the hypergraph generator of Algorithm 1 in Section IV-C, input parameters other than A_k were determined from STA-coauthor in order to generate realistic hypergraphs. Fig. 3 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.

We next estimated the PA function A_k by our proposed method in all datasets. Fig. 4 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 preferential attachment. 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 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. 5(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 $\alpha \leq 1$ [40]. 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. 5(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 on various higher-order growth factors, such as the distributions of sizes and numbers of hyperedges at each time-step.

C. 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 III-C, 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 III-C, 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

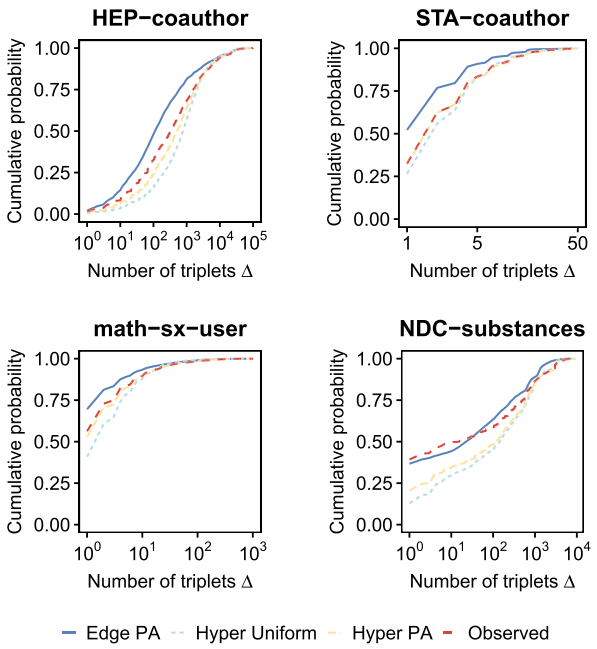


FIGURE 6. Observed and simulated cumulative probability distributions of the numbers of triplets in some representative datasets. For each of 13 datasets, we generate 10 graphs by Edge PA, and 10 hypergraphs by Hyper PA and Hyper Uniform, respectively. The generated hypergraphs are converted into graphs 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 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 under-estimating in the region of low number of triplets, while Edge PA performed well.

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 [13]. 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 [30], [41]. We here examine the distribution of the number of triangles that a 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 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 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

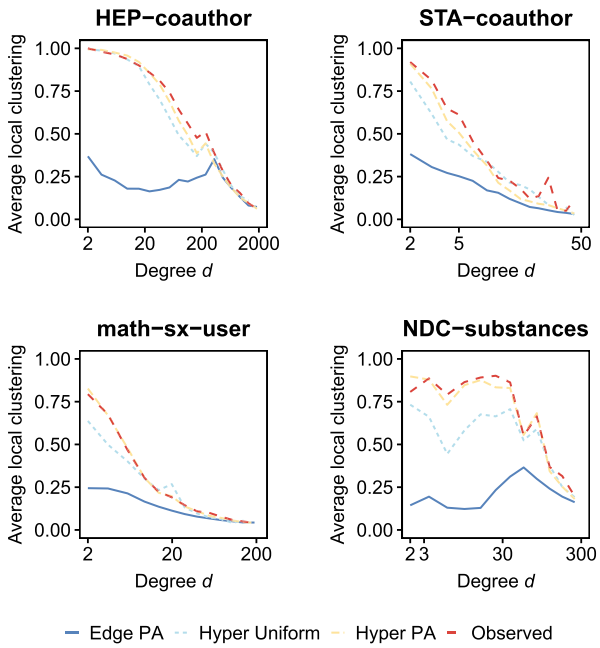


FIGURE 7. Observed and simulated local clustering coefficients averaged over nodes with degree d in some representative datasets. See Fig. 6 for simulation settings. To illustrate, we show the observed and simulated $C(d)$ for four representative datasets: HEP-coauthor, STA-coauthor, math-sx-user, and NDC-substances. The quantitative comparison results for all datasets are given in Table 1. In these four datasets, Hyper PA succeeded in replicating the signature decreasing of $C(d)$ when d increases. Edge PA often under-estimated $C(d)$, especially in the region of low d.

of triplets. Table 1 shows a quantitative comparison for all datasets using $E^{triplet}$, which is calculated as follows:

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

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 $\Delta'_1, \ldots, \Delta'_{N_{\rm bin}}$ are the logarithmic binning of Δ, and we describe the result with $N_{bin} = 10$ in Table 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 [42]. 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:

$$C(d) = \frac{1}{N_d} \sum_{i \in V_d} C_i, \tag{9}$$

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. 7 shows the observed and simulated distributions $C(d)$ for some representative datasets. Hyper PA succeeded 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 1 shows a quantitative comparison in all the 13 datasets using E^{local}, which is calculated as follows:

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

where N_{bin} is the number of logarithmic bins, and C_{obs} and C_{sim} are (9) of observed and simulated data, respectively. We note that $d'_1, \ldots, d'_{N_{\rm bin}}$ are the logarithmic binning of d, and we describe the result with $N_{bin} = 10$ in Table 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 linear PA function, while Edge PA estimates the PA function from data. This highlights the importance of incorporating hyperedge information. Taking into accounts the results in Section III-C, Hyper PA replicates well various first-order and second-order structures in all datasets.

VI. CONCLUSION

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. In this paper, we have proposed a 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.

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 (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 [14]. 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.

APPENDIX A DERIVATION OF THE CONDITIONAL PROBABILITY TO EXCLUDE EFFECTS OF NEWCOMER NODES

We here derive the conditional probability (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 $V_{new} = V_t \setminus V_{t-1}$ and $V_{exist} = V_{t-1}$. 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 $B = B_1 \cup B_2$, $B_1 \subset V_{new}$, $B_2 \subset V_{exist}$, $|B_1| = n$, and $|B_2| = m - n = m'$. We here decompose (2) into the conditional probability given that $B_1 = B^*$ for a pre-specified set $B^* \subset V_{new}$ and the probability of observing B^*:

$$\begin{aligned} P_B(t) &= P_{B_1 \cup B_2}(t) \\ &= P_{B_1 \cup B_2, B_1 = B^*}(t) \\ &= P_{B_1 \cup B_2 | B_1 = B^*}(t) P_{B_1 = B^*}(t). \end{aligned} \tag{10}$$

The term $P_{B_1 \cup B_2 | B_1 = B^*}(t)$ corresponds to the desired conditional probability in (8). Note that we denote $B^* \subset V_{new}$ and not $B^* = V_{new}$ because we allow the temporal hypergraphs to add multiple hyperedges at each time-step t. With (3) and (10), we obtain:

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

thus showing (8).

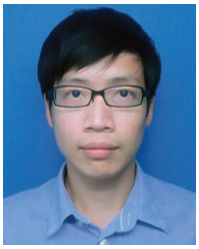
So far we have assumed that $B_1 = B^*$ is pre-specified, but we can change the setting so that B_1 is randomly sampled from V_{new} with $P(B_1) = 1/\binom{|V_{\text{new}}|}{n}$. Then $\log P(B_1)$ terms may be added to the log-likelihood function $L(\boldsymbol{A}|G_0, \ldots, G_T)$. However, since $\log P(B_1)$ does not involve $\boldsymbol{A}$, it does not change the maximum likelihood estimation of $\boldsymbol{A}$.

REFERENCES

[1] M. A. Riolo and M. E. J. Newman, "Consistency of community structure in complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 101, no. 5, May 2020, Art. no. 052306.
[2] X. Wang, Z. Wang, and H. Shen, "Dynamical analysis of a discrete-time SIS epidemic model on complex networks," *Appl. Math. Lett.*, vol. 94, pp. 292–299, Aug. 2019.
[3] G. F. de Arruda, F. A. Rodrigues, and Y. Moreno, "Fundamentals of spreading processes in single and multilayer complex networks," *Phys. Rep.*, vol. 756, pp. 1–59, Oct. 2018.
[4] G. F. de Arruda, G. Petri, and Y. Moreno, "Social contagion models on hypergraphs," *Phys. Rev. Res.*, vol. 2, no. 2, Apr. 2020, Art. no. 023032, doi: 10.1103/PhysRevResearch.2.023032.
[5] R. I. Lung, N. Gaskó, and M. A. Suciu, "A hypergraph model for representing scientific output," *Scientometrics*, vol. 117, no. 3, pp. 1361–1379, Dec. 2018.
[6] X. Zheng, Y. Luo, L. Sun, X. Ding, and J. Zhang, "A novel social network hybrid recommender system based on hypergraph topologic structure," *World Wide Web*, vol. 21, no. 4, pp. 985–1013, 2018.
[7] A. Mithani, G. M. Preston, and J. Hein, "Rahnuma: Hypergraph-based tool for metabolic pathway prediction and network comparison," *Bioinformatics*, vol. 25, no. 14, pp. 1831–1832, Jul. 2009, doi: 10.1093/bioinformatics/btp269.
[8] X. Sun, H. Yin, B. Liu, H. Chen, J. Cao, Y. Shao, and N. Q. V. Hung, "Heterogeneous hypergraph embedding for graph classification," in *Proc. 14th ACM Int. Conf. Web Search Data Mining (WSDM)*. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 725–733, doi: 10.1145/3437963.3441835.
[9] B. Kamiński, V. Poulin, P. Prałat, P. Szufel, and F. Théberge, "Clustering via hypergraph modularity," *PLoS ONE*, vol. 14, no. 11, pp. 1–15, 2019, doi: 10.1371/journal.pone.0224307.
[10] A. Spitz, D. Aumiller, B. Soproni, and M. Gertz, "A versatile hypergraph model for document collections," in *Proc. 32nd Int. Conf. Sci. Stat. Database Manage. (SSDBM)*. New York, NY, USA: Association for Computing Machinery, Jul. 2020, pp. 1–12, doi: 10.1145/3400903.3400919.
[11] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Sep. 1999. [Online]. Available: http://science.sciencemag.org/content/286/5439/509
[12] T. A. B. Snijders, "Stochastic actor-oriented models for network dynamics," *Annu. Rev. Statist. Appl.*, vol. 4, no. 1, pp. 343–363, Mar. 2017, doi: 10.1146/annurev-statistics-060116-054035.
[13] T. Pham, P. Sheridan, and H. Shimodaira, "PAFit: A statistical method for measuring preferential attachment in temporal complex networks," *PLoS ONE*, vol. 10, no. 9, Sep. 2015, Art. no. e0137796.
[14] T. Pham, P. Sheridan, and H. Shimodaira, "Joint estimation of preferential attachment and node fitness in growing complex networks," *Sci. Rep.*, vol. 6, no. 1, 2016, Art. no. 32558, doi: 10.1038/srep32558.
[15] J. Overgoor, A. Benson, and J. Ugander, "Choosing to grow a graph: Modeling network formation as discrete choice," in *Proc. World Wide Web Conf. (WWW)*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1409–1420, doi: 10.1145/3308558.3313662.
[16] M. Inoue, T. Pham, and H. Shimodaira, "Joint estimation of non-parametric transitivity and preferential attachment functions in scientific co-authorship networks," *J. Informetrics*, vol. 14, no. 3, Aug. 2020, Art. no. 101042. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S175115771930269X
[17] M. T. Do, S.-E. Yoon, B. Hooi, and K. Shin, *Structural Patterns and Generative Models of Real-World Hypergraphs*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 176–186, doi: 10.1145/3394486.3403060.
[18] G. Lee, M. Choe, and K. Shin, "How do hyperedges overlap in real-world hypergraphs?—Patterns, measures, and generators," in *Proc. Web Conf. WWW*. New York, NY, USA: Association for Computing Machinery, Apr. 2021, pp. 3396–3407, doi: 10.1145/3442381.3450010.
[19] J.-G. Liu, G.-Y. Yang, and Z.-L. Hu, "A knowledge generation model via the hypernetwork," *PLoS ONE*, vol. 9, no. 3, pp. 1–8, Mar. 2014, doi: 10.1371/journal.pone.0089746.
[20] Y. Yang, Y. Dong, and N. V. Chawla, "Predicting node degree centrality with the node prominence profile," *Sci. Rep.*, vol. 4, no. 1, May 2015, Art. no. 7236, doi: 10.1038/srep07236.
[21] T. Michoel and B. Nachtergaele, "Alignment and integration of complex networks by hypergraph-based spectral clustering," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 86, no. 5, Nov. 2012, Art. no. 056111, doi: 10.1103/PhysRevE.86.056111.
[22] R. Guimerà, B. Uzzi, J. Spiro, and L. A. N. Amaral, "Team assembly mechanisms determine collaboration network structure and team performance," *Science*, vol. 308, no. 5722, pp. 697–702, 2005. [Online]. Available: https://science.sciencemag.org/content/308/5722/697

[23] X. Hu, R. Rousseau, and J. Chen, "In those fields where multiple authorship is the rule, the h-index should be supplemented by role-based h-indices," *J. Inf. Sci.*, vol. 36, no. 1, pp. 73–85, Feb. 2010, doi: 10.1177/0165551509348133.
[24] B. Cronin, "Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices?" *J. Amer. Soc. Inf. Sci. Technol.*, vol. 52, no. 7, pp. 558–569, 2001. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.1097
[25] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, A. Vespignani, L. Waltman, D. Wang, and A.-L. Barabási, "Science of science," *Science*, vol. 359, no. 6379, pp. 1–9, 2018. [Online]. Available: http://science.sciencemag.org/content/359/6379/eaao0185
[26] D. De Solla Price, "A general theory of bibliometric and other cumulative advantage processes," *J. Amer. Soc. Inf. Sci.*, vol. 27, no. 5, pp. 292–306, Sep. 1976. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.4630270505
[27] J. Kunegis, M. Blattner, and C. Moser, "Preferential attachment in online networks: Measurement and explanations," in *Proc. 5th Annu. ACM Web Sci. Conf.*, 2013, pp. 205–214.
[28] V. Gómez, H. J. Kappen, and A. Kaltenbrunner, "Modeling the structure and evolution of discussion cascades," in *Proc. 22nd ACM Conf. Hypertext Hypermedia*, 2011, pp. 181–190.
[29] P. Sheridan, Y. Yagahara, and H. Shimodaira, "Measuring preferential attachment in growing networks with missing-timelines using Markov chain Monte Carlo," *Phys. A, Stat. Mech. Appl.*, vol. 391, no. 20, pp. 5031–5040, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378437112004128
[30] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 64, no. 2, 2001, Art. no. 025102.
[31] H. Jeong, Z. Néda, and A.-L. Barabási, "Measuring preferential attachment in evolving networks," *Europhys. Lett.*, vol. 61, no. 4, p. 567, 2003.
[32] P. Ji and J. Jin, "Coauthorship and citation networks for statisticians," *Ann. Appl. Statist.*, vol. 10, no. 4, pp. 1779–1812, Dec. 2016, doi: 10.1214/15-AOAS896.
[33] M. Inoue, T. Pham, and H. Shimodaira, "Transitivity vs preferential attachment: Determining the driving force behind the evolution of scientific co-authorship networks," in *Unifying Themes in Complex Systems IX* (Springer Proceedings in Complexity). Cham, Switzerland: Springer, 2018, pp. 262–271.
[34] J. Kunegis, "KONECT: The Koblenz network collection," in *Proc. 22nd Int. Conf. World Wide Web (WWW Companion)*. New York, NY, USA: Association for Computing Machinery, 2013, pp. 1343–1350, doi: 10.1145/2487788.2488173.
[35] G. A. Baker, Jr., "A new derivation of Newton's identities and their application to the calculation of the eigenvalues of a matrix," *J. Soc. Ind. Appl. Math.*, vol. 7, no. 2, pp. 143–148, Jun. 1959.
[36] T. Pham, P. Sheridan, and H. Shimodaira, "PAFit: An r package for the non-parametric estimation of preferential attachment and node fitness in temporal complex networks," *J. Stat. Softw.*, vol. 92, no. 3, pp. 1–30, 2020. [Online]. Available: https://www.jstatsoft.org/v092/i03
[37] G. A. Ronda-Pupo and T. Pham, "The evolutions of the rich get richer and the fit get richer phenomena in scholarly networks: The case of the strategic management journal," *Scientometrics*, vol. 116, no. 1, pp. 363–383, Jul. 2018, doi: 10.1007/s11192-018-2761-3.
[38] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg, "Simplicial closure and higher-order link prediction," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 48, pp. E11221–E11230, Nov. 2018. [Online]. Available: https://www.pnas.org/content/115/48/E11221
[39] M. Kahn, "Co-authorship as a proxy for collaboration: A cautionary tale," *Sci. Public Policy*, vol. 45, no. 1, pp. 117–123, Feb. 2018, doi: 10.1093/scipol/scx052.
[40] P. L. Krapivsky and S. Redner, "Organization of growing random networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 63, no. 6, May 2001, Art. no. 066123, doi: 10.1103/PhysRevE.63.066123.
[41] G. Bianconi, R. K. Darst, J. Iacovacci, and S. Fortunato, "Triadic closure as a basic generating mechanism of communities in complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 90, no. 4, Oct. 2014, Art. no. 042806, doi: 10.1103/PhysRevE.90.042806.
[42] C. Seshadhri, A. Sharma, A. Stolman, and A. Goel, "The impossibility of low-rank representations for triangle-rich complex networks," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 11, pp. 5631–5637, Mar. 2020.

MASAAKI INOUE received the B.E. degree in engineering and the master's degree in informatics from Kyoto University, Kyoto, Japan, in 2017 and 2019, respectively, where he is currently pursuing the Ph.D. degree with the Department of Systems Science, Graduate School of Informatics. He is also a Junior Research Associate with the RIKEN Center for Advanced Intelligence Project. His research interests include network science, social network analysis, and data mining.



THONG PHAM received the B.S. degree in computer science from Ritsumeikan University, Kyoto, Japan, in 2012, the M.S. degree in mathematical science from the Tokyo Institute of Technology, Tokyo, Japan, in 2014, and the Ph.D. degree in statistics from Osaka University, Osaka, Japan, in 2017.

From 2016 to 2017, he was a Research Fellowship for Young Scientists from the Japan Society for the Promotion of Science. Since 2017, he has been a Postdoctoral Researcher at the RIKEN Center for Advanced Intelligence Project. His research interests include network science, statistics, and machine learning.

HIDETOSHI SHIMODAIRA was born in Tokyo, Japan, in 1967. He received the bachelor's, master's, and Ph.D. degrees from the Department of Mathematical Engineering and Information Physics, The University of Tokyo, in 1990, 1992, and 1995, respectively.

From 1995 to 1996, he was a Research Fellowship for Young Scientists with the Japan Society for the Promotion of Science. From 1995 to 1996, he was also a Visiting Scholar with the Department of Genetics, University of Washington. From 1999 to 2001, he was a Visiting Scholar with the Department of Statistics, Stanford University. From 1996 to 2002, he was an Assistant Professor with the Department of Prediction and Control, Institute of Statistical Mathematics. He was a Senior Lecturer, from 2002 to 2005, and an Associate Professor, from 2005 to 2012, with the Department of Mathematical and Computing Sciences, Tokyo Institute of Technology. From 2012 to 2017, he was a Professor with the Division of Mathematical Science, Osaka University. Since 2016, he has been a Team Leader (concurrent position) with the RIKEN Center for Advanced Intelligence Project. Since 2017, he has also been a Professor with the Department of Systems Science, Graduate School of Informatics, Kyoto University. His research interests include statistics and machine learning.