Evolving Scholarly Networks: Experiments, Modeling and Analysis

Fengyu Deng, Lingkun Kong, Jiaqi Liu, Jinghao Zhao, Jialu Wang, Luoyi Fu, Xinbing Wang
Shanghai Jiao Tong University, China

ABSTRACT

Scholarly networks contain massive scholarly information that can be mainly categorized into three elements, i.e., paper, author and topic, which exhibit a co-evolution over time. Understanding scholarly network structure and evolution has important implications in many aspects such as wiser design of scholar recommendation systems, better evaluation of research communities, more accurate prediction of scientific trends and etc. However, due to theoretical and technical difficulties, there have been few studies that provide a systematic understanding of scholarly networks at scale.

We bridge this gap using real scholarly datasets — *Microsoft Academic Graph* [1] with 126 million papers collected from multiple domains. By empirical exploration, we observe novel features that belong exclusively to scholarly networks, such as varying and converging exponents of power-law distributions with time, degree densification in each of the three aforementioned element sets, i.e., paper, author and topic. We also observe interesting evolving patterns like simultaneous co-evolution of all the three sets, faster growth rate of elements with larger size, and etc.

Based on our empirical observations, we propose a new and novel evolving scholarly model that jointly captures both intra and inter correlations of papers, authors and topics during the evolving process. Through both theoretical analysis and empirical evaluations, we demonstrate that our model can accurately reproduce the global and local structures of real scholarly networks.

CCS CONCEPTS

•Information systems →Data mining; •Networks →Network dynamics; •Computing methodologies →Modeling and simulation;

KEYWORDS

Data mining; Scholarly networks; Evolution

1 INTRODUCTION

Recent years have witnessed the rapidly growing scholarly information due to vast research works are undertaken in academia and industry [2]. As a result of advancement in research communities, researchers all over the world steadily produce a large volume of research articles, which provide the technological basis for worldwide dissemination of scientific findings. Therefore, large collections of scholarly data such as publication's name, author names, citations, topics, etc. emerge by researchers' continuous working. All the information, when combined together, leads to the formation of the scholarly network that contains three major elements, i.e., *paper*, *author* and *topic*, as will also be the concern of this paper. As can be seen from the Figure 1, the title of the publication, which we regard as paper, the authors of the publication, which we view as author, and the listed keywords which we also consider as topic are strongly connected to each other in the sense that authors on the same paper exhibit collaborative relations (shown by arrows between *A*s, which mean authors), with that paper (presented by arrows between *A* and *P*) further cited by other literatures (represented by arrows among *P*s, which mean papers) belonging to some specific topics (denoted by arrows between *K*s, which mean keywords, and meanwhile topics, and *A*). All the three elements interact with each other in the prescribed way as time goes by, leading to larger number of new publications that further manifest such correlation among the three elements. Consequently, the scholarly network that contains the three elements is also evolving on the whole, in terms of both the element size and its more complicated connection of inter & intra element sets.

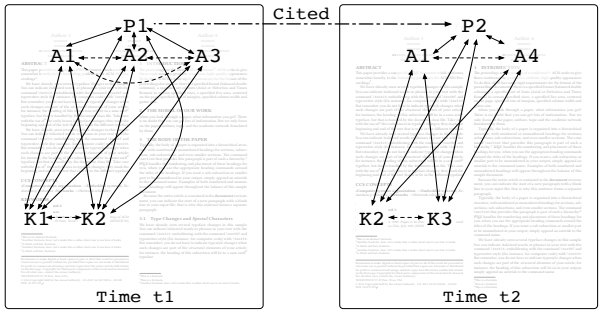


Figure 1: The heterogeneous scholarly network extracted from real papers.

As a matter of fact, studying properties of scholarly networks and getting insight of their evolving mechanism have important implications, as reported by a literature survey [3]. For example, by analyzing the citation relationships extracted from scholarly networks, we can evaluate the impact of a given paper or scholar, which can help to allocate reputations to scientists [4]. Similarly, by analyzing the co-author behaviors among scholars, we can find the distribution of scientific communities, which helps to build the map of sciences. Meanwhile, scholarly network analysis is not only important for academia but also helps in planning and development, i.e., for sociologists to understand researcher interactions [5], for policy makers to address knowledge and resources sharing, and for scientists, businesses, and the general public as a reference.

Despite the importance of scholarly networks in many kinds of applications, there have been few studies at systematically measuring and modeling the evolution of scholarly networks, primarily due to three major challenges. First of all, due to the difficulty in acquisition of real scholarly data that contain complete information of papers, authors and topics, there is a significant lacking of studies that empirically explore the properties of scholarly networks. This is in sharp contrast to traditional social networks, which have been empirically verified to exhibit a series of well-known properties[6]. Secondly, prior related works are mainly restricted to certain parts of extracted from the whole scholarly network (e.g., citation networks, co-author networks or topic networks), with the

corresponding modeling mainly based on the structure of the bipartite graph or even single graph [7–11]. Though facilitating analysis, the lacking of the integrality of scholarly networks might lead to information omitting or even misunderstanding. For instance, an author who is active in co-author networks is likely to contribute no influential papers. Thus, only studying co-author network might result in inaccurate evaluation of this author. Thirdly, due to inappropriate modeling methods, a large amount of previous studies get stuck in giving detailed mathematical proofs to support their model. In contrast, they employ simulating experiments or other tricky evaluating methods to validate their models' brilliance, which in our eyes, is not compelling enough.

Faced with these challenges, we are motivated to give the first comprehensive properties analysis on scholarly networks by launching experiments on real-world datasets which have around 126 million papers. Also, to capture our empirical observations, we propose the first model that is inspired from the structure of tripartite graph to roundly incorporate all information and their correlation in scholarly networks. Last but not least, we mathematically prove the properties of our proposed model, along with further verification through simulations. And here, we illustrate our work and summarize our contributions by three aspects.

Experiments: Our first contribution is to originally explore comprehensive properties in scholarly networks by experiments' results on real-world datasets. Based on scholarly datasets provided by Microsoft [1], which contain about 126 million papers, we use data-mining and other big-data analyzing approaches to observe patterns in the growth of the scholarly network. On one hand, we observe some similar features of scholarly networks to those that have already been discovered in many traditional social networks, such as power-law degree distribution, network's degree densification and etc. On the other hand, in contrast to many prior evolving networks, there also exists several unique features in scholarly networks, like faster growth rate of the elements that have a bigger size, varying and converging exponents in power-law distributions with time, and the simultaneous co-evolution of all elements, and etc. All these evolving features, depending on whether they are established upon single or multiple element sets, can be categorized into three types, i.e., inter-evolution, intra-evolution as well as the co-evolution on the whole. While deferring to Section 3 for more details, we remark that there is no prior work, other than ours, that have studied these properties in scholarly networks.

Modeling: Given empirical observations, our next significant contribution is for the first time establishing a comprehensive modeling of evolving scholarly networks. Converting paper, author and topic into the three major element sets, the proposed model captures both the inter-correlation and intra-correlation of the three sets during the evolving process. Particularly, inter-correlation is characterized through tripartite graph, whose evolving process follows the mode of preferential attachment prevalent in growth of real scholarly networks; Meanwhile, intra-correlation of nodes within each element is described as intra-degree (which we define) power-law distribution, degree densification, and etc.

Analysis: Our third contribution is to offer detailed mathematical proof to consolidate the reasonability of our model. To the best of our knowledge, there is few or even none of previous models targeting scholarly networks gives proper mathematical explanation. In this case, based on the constructing methods of random arrival, preferential attachment, edge copying and the assumption of the affiliation relationship inside elements' set, we successfully obtain the growing rate of nodes' degree, power-law distributions inside or among the elements' sets and the densification of the entire network. Further, we also use empirical evaluation to validate that our model can accurately reproduce real scholarly networks.

The paper is organized as follows. In Section 2, we discuss relevant literatures. In Section 3, We list properties in scholarly networks. We give our generative evolving scholarly model in Section 4 and analyze our model mathematically in Section 5. Section 6 is our simulation and we conclude in Section 7.

2 RELATED WORK

To the best of our knowledge, there are no prior works, other than ours, that fully discussed the evolution of scholarly networks from a global viewpoint, i.e. getting thorough understanding by combining all scholarly information together. However, there are indeed several related works regarding investigations of both evolving and scholarly networks.

Evolving networks have long been a significant research topic [12–14]. And there are a lot of properties discovered from real-world network graphs, such as power-law degree distribution, small-world phenomenon, degree densification and etc.[6, 14–17], which are nobly detected and simulated by a wide spectrum of evolving network models.

However, most of studies mainly focus on social networks, purposing models to reproduce observed features in social networks, including random graph model built by Chakrabarti and Faloutsos [18], preferential attachment model proposed by Barabasi et al. [19], edge-copy model created by R. Kumar et al. [20], and affiliation network model advanced by Silvio Lattanzi et al. [21], while ignoring general understanding of the scholarly network, which is also one of common and important evolving networks in our life.

Besides, some efforts have also been made to launch study on scholarly networks for diverse usage. For instance, Zaihan Yang et al. [7] use a joint topic model to solve scholar ranking and predicting problems. Kajikawa et al. [11] analyze citation networks to create an academic landscape of sustainability science. Jing Li et al. [10] propose a random walk model based on co-author networks for recommending new collaborations. And Lin et al. [8] study topic evolution of scholarly networks. However, due to theoretical and technical difficulties, none of them generates research in a more general perspective – studying entire properties of scholarly networks which include all scholarly information.

Therefore, we propose a novel scholarly model which employs tripartite graph to depict scholarly networks' inter-correlation while uses affiliation networks' structure to portray intra-correlation, can well reproduce properties we observed in real scholarly networks. Besides, by using preferential attachment and edge copying approaches [20, 22], we construct our sophisticated model with concise mathematical proofs.

3 EXPERIMENTS

In this section, we give our experiments based on *Microsoft Academic Graph (MAG)* [1] which is official and authoritative scholarly dataset containing massive scholarly information of publications

such as titles, authors, conferences, fields of study and citations. Around 126 million papers in 19 subjects are included in this database and the published years of them vary from 1800 to 2016. To prove that our experiments are representative and persuasive in scholarly networks, we observe in different fields. And four of them containing about 4.7 million papers are extracted to show the properties, which are: Data Mining, Networks, Literature and Finance. The detailed information is listed in Table 1.

Table 1: Statistics of Scholarly Datasets

Dataset	# of Papers	# of Authors	# of Topics
Data Mining	1042279	1703828	403
Networks	1093537	1391869	774
Literature	679350	939544	446
Finance	1949028	2716094	968

In each field, we already have published time of papers. Then, author's time is defined by the published time of the author's first paper. In the same way, the topic's time is also determined. Based on these four datasets, we mainly study on two kinds of features of the evolving scholarly networks. The first is structure properties, more specifically, degree distribution in our work. The degrees contain both intra-degree in single element and inter-degree between different elements. The second feature is evolving properties over time. We divide evolving properties into three parts. The first is intra relationships in a single scholarly element set such as reference in citation network or coauthorship in co-author network. The second part includes inter correlation of different sets such as topic-paper evolution. Finally, we extract the connections among all elements including papers, authors and topics.

3.1 Degree distributions

Power-law distributed degree is a common feature of social networks, which is also well studied by many existing literatures [15, 23, 24]. But when it coming to scholarly networks, does it still work well? In fact, in scholarly networks, by our experiments result, this feature also exists. And we explore this feature in two cases: intra-degree and inter-degree.

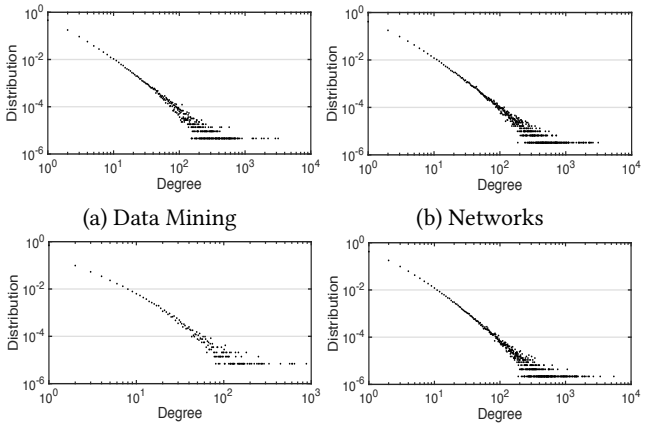
Intra-degree: The first is intra-degree of one single element set, i.e., degree in paper citation network, co-author network or topic networks. We take citation network as an example to present our observations.

Recall that a random variable $x \in \mathbb{Z}^+$ follows a power-law distribution if:

$$\mathbb{P}\{x = k\} = \eta k^{-\varphi},$$
$$log(\mathbb{P}\{x = k\}) = -\varphi log(k) + log(\eta),$$

where φ is an exponent factor of the power-law distribution and η is a constant factor. $\mathbb{P}\{x = k\}$ is referred as the probability when $x = k(k \in \mathbb{Z}^+)$. From above equations, we know that the larger φ is, the greater probability x has to stay in a small value.

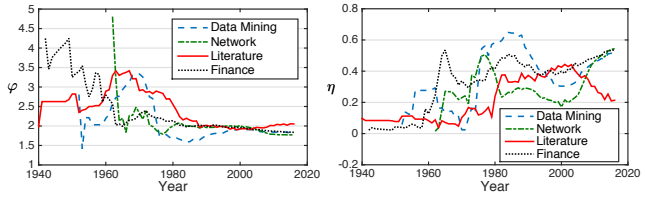
Figure 2 shows the degree distribution of citation networks in four datasets. Obviously, they all follow power-law degree distribution. And we also study on the evolving process of factor φ and η. The result is illustrated in Figure 3. As time grows, in three datasets except Literature, φ keeps the value about 1.77 while η fluctuates with the time and finally reaches 0.53. The φ of Literature is larger, which means that fewer papers have higher citations in this field.



(a) Data Mining (b) Networks

(c) Literature (d) Finance

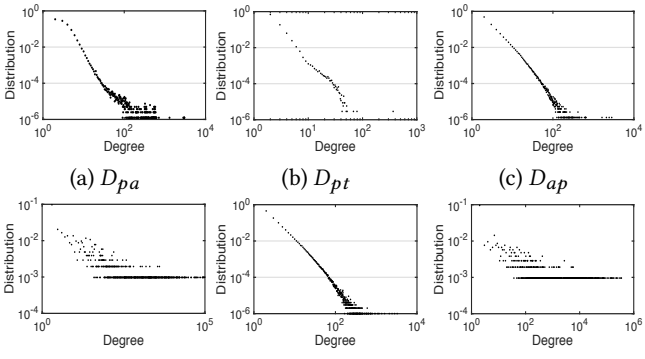
Figure 2: Degree distribution of citation network.



(a) Evolution of φ (b) Evolution of η

Figure 3: Evolution of power-law distribution factors

Inter-degree: The second degree we discuss is inter-degree, i.e., the degree generated by edges between different element sets such as topic degree in topic-paper sub-network which measures the size of this topic, author degree in author-paper sub-network which means how many papers the author has published. In our scholarly network, these sub-networks include two element sets and the links between them. So we have three kinds of sub-network between every two element sets and six degree distributions of these three networks. For convenience, we denote degree distributions of paper, author in paper-author sub-network as D_{pa}, D_{ap}. In the same way, we also define D_{ap}, D_{at}, D_{tp} and D_{ta}.



(a) D_{pa} (b) D_{pt} (c) D_{ap}

(d) D_{at} (e) D_{ta} (f) D_{ta}

Figure 4: Degree distribution between different elements

Figure 4 gives an intuitive degree distribution of these degrees in Finance and they all follow power-law distribution as expected. However, because three elements (paper, author, topic) are not numerical symmetrical in scholarly network, i.e., a paper may only be related to a number of topics while a topic can link to thousands

of papers. Thus the distribution types are different, and we list the final distribution factors in Table 2. Compared with others, the φ of topic's degree is smaller, which means this topic has more chances to include large number of papers and authors.

Table 2: Power-law distribution factors

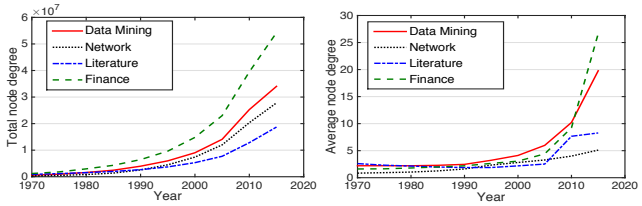
Factors	D_{pa}	D_{pt}	D_{ap}	D_{at}	D_{tp}	D_{ta}
φ	1.67	3.11	2.38	2.32	0.18	0.13
η	1.70	0.34	0.52	0.32	2.32	2.48

3.2 Intra-set evolution

To comprehensively analyze scholarly network, we first study one simple scholarly element in this subsection to reveal the evolving properties, which we refer as intra-set evolution. Then we take author set and the co-author network as an example. Looking into the evolution process of co-author network, we focus on two kinds of degree growths to show the densification [14] property of our scholarly network.

To begin with, we observe the increasing mechanism of nodes' total degree in intra-sets. As shown in Figure 5(a), the nodes' total degree is very small in the last century while the growth rate is very high in the 21st century. We reckon that in the new century, for the benefit of rapid development in science and technology, authors have more opportunities to cooperate with other scholars. Besides, it can be found that the nodes' average degree increases exponentially. The result is shown in Figure 5(b).

As a whole, both kinds of node degree we study comply with an exponential increasing pattern. Detailed theoretical analysis is given in Section 5.



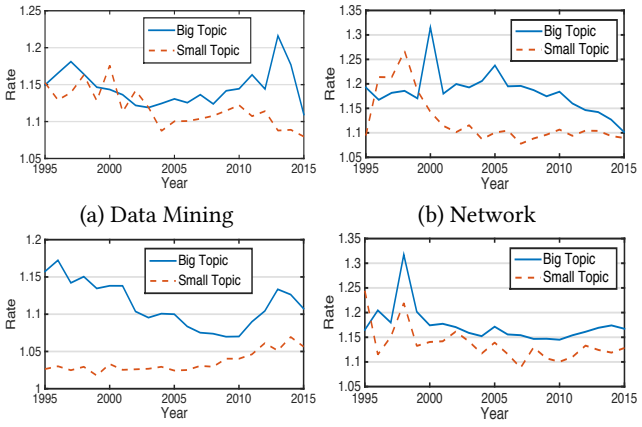
(a) Total node degree (b) Average node degree

Figure 5: Evolution of author's total degree and average degree in co-author network.

3.3 Inter-set evolution

Paper, author, and topic, in fact, are never separated from each other in real scholarly networks. Before we focus on the mechanism of the whole scholarly network, in this subsection, we study the relationship evolving over time between two elements. We refer this as inter-set evolution and use topic size evolution as an example to describe their connections.

According to previous definition, the size of a topic can be measured by how many papers it contains, or how many papers it links to in our scholarly network. In our datasets, we notice that some big topics can link to more than 100 thousand papers while some small topics only contain one hundred papers. Since they all grow up from a topic with small amount of papers, there must exist difference in the process of their growth. What we observe is that topics with larger size grow with faster rate. In our experiments, we divide topics into two groups according to their sizes. Then we calculate the average growth rate of topic size and plot the results in Figure 6. The reasonability of the observation holds because when a topic gets bigger, it will be more likely to draw attention from authors, who, as a result, might contribute new publications under this topic. Consequently, the growth rate stays high.



(a) Data Mining (b) Network

(c) Literature (d) Finance

Figure 6: Evolution of topic size

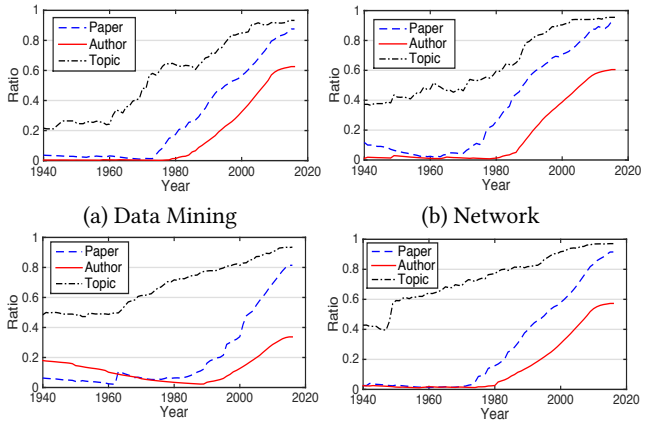
Moreover, the growth rate changes over time. In dataset of Networks, the size growth rate of big topic is decreasing while small topic's remains stable. This indicates that in the next few years, the size of the big topic may increase slowly, which, to some extent, reveals the topic with large size might not still be hot in the foreseeable future.

3.4 Co-evolution of three sets

Apart from intra-set evolution or inter-set evolution, what we are most curious about is the co-evolution of the whole scholarly network including all element sets: paper, author and topic. We are eager to find a novel pattern, which is observed by all elements during the evolving process, to present the evolving structure and property of scholarly networks.

For a promising field, there is no doubt that more authors would like to set foot in this field and launch research on relevant topics, then more papers will be published. Consequently, it will lead to growing citations, stronger coauthorship and more diverse topics in this field. As a result, the connectivity of scholarly networks will become denser. Based on this intuition, we find this pattern from the perspective of connectivity in our experiments on these four datasets.

Giant Component [25] is the largest connected component of a given graph. If a giant component contains C_g nodes and the total nodes in the graph is C_G, then the connectivity of this graph can be measured by the ratio of C_g and C_G: $ratio = C_g/C_G$. We calculate the ratio of each element set at every time slot and the result is illustrated in Figure 7. According to this result, the ratio in each subgraph of the scholarly network has strong connection with other elements and they grow with the same pattern simultaneously. It is a symbol that all element sets co-evolution in the scholarly network. They affect each other and together present the mechanism of evolving scholarly networks.



(a) Data Mining (b) Network

(c) Literature (d) Finance

Figure 7: Evolution of ratio (C_g/C_G) in three sub-graphs of the scholarly network

4 MODELING OF SCHOLARLY NETWORKS

Based on the above observations, we design a novel model to capture these properties and we name this model as: *Evolving Scholarly Model*. In this section, we first introduce the proposed model and then describe the evolving process of it.

4.1 Evolving scholarly model

In our evolving scholarly model, the graph is denoted as $G(P, A, T)$. Then we use tripartite graph to present the inter-correlation between elements. Besides, we also focus on the intra-features of every element. For an intuitive understanding, we illustrate the framework of our evolving scholarly model in Figure 8. It contains:

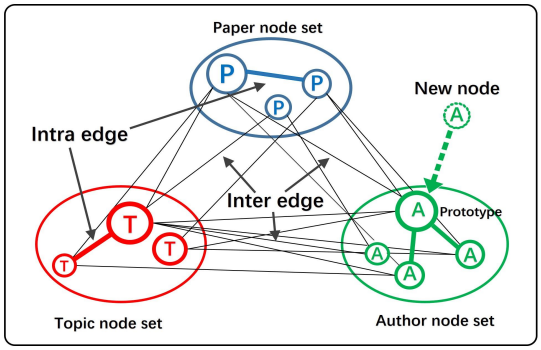


Figure 8: Structure of evolving scholarly model

(1) Three node sets: Paper node set N_p, author node set N_a and topic node set N_t. The node in each node set is marked as n_p, n_a and n_t.

(2) Three inter-edge sets: We denote the edges between every two different node sets as inter-edge sets, and the graph has three inter-edge sets. For example, we refer all edges between paper node and author node as $E_{pa}(E_{ap})$, then an edge $e_{n_p n_a}$ which belongs to E_{pa} means author n_a writes the paper n_p.

(3) Three intra-edge sets: Intra-edge is the edge in the same node set, and our graph has three intra-edge sets, which we refer as E_{pp}, E_{aa}, and E_{tt}. If an edge $e_{n_a^i n_a^j} \in E_{aa}$, then we know that author n_a^i and n_a^j have cooperation.

In Figure 8, nodes are illustrated as colorful circles in each node set while intra edge and inter edge are labeled. And a new author node is trying to preferentially attach himself with some heavily linked authors nodes (distinguished by their sizes) that are already in the author set. With these nodes and edges of the model, we can well extract the structure of scholarly networks. We present notations in Table 3 for later convenience and describe the evolving process of the proposed model in the following subsection.

Table 3: Notations and Definations

Notations	Definations
N_p, N_a, N_t	Node set of Paper, Author and Topic
E_{ij}	Inter-edge set between nodes in N_i and N_j
E_{ii}	Inner-edge set of N_i
α_i	Probability that a new node arrives in N_i
β_{ij}	Probability that an edge added in set E_{jj}
c_{ij}	Number of edges added to set E_{ij} at one time slot
$G(P, A, T)$	Graph of our evolving scholarly model
$B(N_i, N_j)$	Bipartite graph with sets N_i, N_j, and E_{ij}

4.2 Evolving process

While we defer the detailed evolving process of the proposed model to Algorithm 1, we would also like to provide a corresponding brief summary of the process. We first fix parameters including α_i, β_{ij} and c_{ij} where $i \neq j \in \{p, a, t\}$, and then assume that the evolution starts from an initial case that can be modeled as an initial graph, showing that each node in the graph is linked to a number of nodes in other node sets. After initialization, for every time slot, we classified the process into five main steps: **1)** A new node, which can be randomly designated as a paper, an author, or a topic, is added to the graph. For clarity, here we only take the arrival of a new author as example for explanation of the subsequent steps. And the symmetry also holds for paper and topic. **2)** With probability proportional to degree in $B(N_a, N_p)$ and $B(N_a, N_t)$, two author nodes n_a^p and n_a^t are chosen as prototypes for the new node n_a. **3)** c_{ap} neighbors (n_p^1 ,..., $n_p^{c_{ap}}$) of n_a^p in N_p and c_{at} neighbors (n_t^1 ,..., $n_t^{c_{at}}$) of n_a^t in N_t are randomly chosen to have connections with node n_a. **4)** $c_{ap}c_{at}$ edges are added between n_p^1 ,..., $n_p^{c_{ap}}$ and n_t^1 ,..., $n_t^{c_{at}}$. **5)** Edges between every two author nodes are added with probability β_{ap} (β_{at}) if they have a common paper (topic).

For a better intuitive understanding of this evolving process, let us, for instance, consider the arrival of a new author. He is likely to learn from an influential author, who is thus selected as a prototype and influences the new author on choosing research topics. In addition, when conducing the new research, this new author probably has on mind some other author who have many publications as another prototype, from whom he chooses some old papers for references. Obviously, the topics and the papers chosen by the same author are often relevant, indicating that these papers belong to these topics with a high possibility. Last but not least, let us consider a co-author network, where a new author may want to collaborate with others who are also interested in these topics and papers for a joint piece of work. Similarly, when a new topic emerges in the literature, it is usually inspired by some existing topics (prototypes), and when a paper arrives, it is often based on a number of old papers (prototypes).

Algorithm 1: Evolving Process

Simulated time steps: T
Fix probability α_i that a new node arrives in N_i.
Fix parameters $\beta_{ij} \in (0,1)$ and integers $c_{ij} > 0$ where $i \neq j \in \{p, a, t\}$.
Initialization: In initial graph, the node in each set has neighbors with other two node sets. For example, a paper node n_p connects to at least c_{pa} author nodes. Meanwhile, an author node n_a has at least c_{ap} paper neighbors. So the inter-edge set E_{pa} has at least $c_{ap}c_{pa}$ edges in the beginning.
for $1 \leq t \leq T$ **do**

1) Node arrival: According to α_p, α_a, α_t, we decide the type of node to join the graph. In later discussion, we take the arrival of a new author node n_a as example, and the symmetry also holds for paper and topic.

2) Preferentially chosen ProtoType: A node $n_a^p \in N_a$ is chosen as prototype for the new node, with probability proportional to its degree in $B(N_a, N_p)$.In the same way, another node $n_a^t \in N_a$ is chosen as prototype according to its degree in $B(N_a, N_t)$.

3) Edge copying: c_{ap} edges are copied from n_a^p, that is, c_{ap} neighbors of n_a^p, denoted by n_p^1 ,..., $n_p^{c_{ap}}$ in N_p are chosen uniformly at random, and the edges (n_a^p, n_p^1), ..., $(n_a^p, n_p^{c_{ap}})$ are added to the graph. Follow the same method, c_{at} edges (n_a^t, n_t^1), ..., $(n_a^t, n_t^{c_{at}})$ are added to the graph.

4) Indirect evolution: $c_{ap}c_{at}$ edges between nodes p_1 ,..., $p_{c_{pa}} \in N_p$ and t_1 ,..., $t_{c_{pt}} \in N_t$ are added to the graph $B(N_p, N_t)$.

5) Evolution inside: For every two nodes n_a^x and n_a^y $(x \neq y)$, if they have a common author (topic), then with probability $\beta_{ap}(\beta_{at})$, a edge (n_a^x, n_a^y) is added in E_{aa}.

end

5 THEORETICAL ANALYSIS

In this section, we mathematically analyze our model and confirm that our model can well reproduce properties in the real-world scholarly network.

5.1 Growth of node degree

According to our model, we divide the nodes' degree into two types – the first is the *inter-degree*, i.e., the node degree between node sets, related with the growth of E_{ij}, we call it d^{ir}, which represents the inter-correlation in our model, and the second is the *intra-degree*, i.e. the node degree inside node set, related with the growth of E_{ii}, we call it d^{ia}, which reflects the intra-correlation of scholar networks.

Growth of inter-degree: Assuming node n arrives at node set N_i at time t_0 with initial inter-degree $d_i^{ir}(t_0)$, the inter-degree of n at time $t > t_0$ is

$$d_i^{ir}(t) = \left(\frac{t}{t_0}\right)^{\lambda_i} d_i^{ir}(t_0),$$

where $\lambda_i \in (0,1)$ is a constant.

In fact, two implications can be deduced by this result.

(1) The inter-degree $d_i^{ir}(t)$ grows with polynomial rate in time t, following the power $\lambda_i \in (0,1)$.
(2) The two components of vector $d_i^{ir}(t)$ are in the same order. For instance, $d_{pa}^{ir}(t) = \Theta\left(d_{pt}^{ir}(t)\right)$.

The first implication gives the growth rate of node's inter-degree. And the second one implies the similarity of a certain node's degree over two different node sets, which indicates that an influential node (node with large inter-degree) also plays an important role in all the other bipartite relationship network and vice versa. And this complies with properties of our scholarly networks. For example, an author who studied in multiple fields is more likely published more papers and vice versa. The detailed proof is given in Theorem 5.1.

Growth of intra-degree: Again, we set beginning time as t_0 and the intra-degree of node set N_i at time $t > t_0$ is $d_i^{ia}(t)$, then

$$d_i^{ia}(t) = \Theta\left(\max\left\{t^{\frac{1}{\lambda_j}+1}, t^{\frac{1}{\lambda_k}+1}\right\}\right),$$

where $\lambda_j, \lambda k$ represent the constant λ in N_j and N_k which are neighbors of N_i, and the max is the maximum of two formulas.

The equation reveals that, in our model, the intra-degree of a node set actually is related with the inter-degree's growing rate variable λ. As in equation the intra-degree is positively related with the growing with time slot t, we can say the intra-degree also grows with time. The detailed proof is given in Theorem 5.2.

Combining these two results together, it can be easily viewed in our scholarly network graph $G(P, A, T)$ that the degree of the node in graph grows with polynomial rate in time t, and the growth rate differs from inter-degree to intra-degree of the node.

Theorem 5.1. *For graph $G(P, A, T)$ generated after t time slots $(t \geq t_0)$, with the initial condition that a certain node $n \in N_p$ is added to node set N_p at time t_0 with the degree $d^{ir}(t_0)$ from N_p to N_a and N_t, the inter-degree of n at time t satisfies*

$$d_p^{ir}(t) = \left(\frac{t}{t_0}\right)^{\lambda_p} d_p^{ir}(t_0).$$

This result also holds for $n \in N_a$ and $n \in N_t$ with symmetrical expressions.

Proof. At each time slot t, the inter-degree of node $n \in N_p$ in $B(N_p, N_a)$, i.e. $d_{pa}^{ir}(t)$, may increase in two cases:

(1) A new node arrives at N_a and is connected to n, which results in $d_{pa}^{ir}(t) = d_{pa}^{ir}(t-1) + 1$.
(2) A new node arrives at N_t and is connected to n, then n will connect to c_{ta} neighbors of the new node in N_a which results in $d_{pa}^{ir}(t) = d_{pa}^{ir}(t-1) + c_{ta}$.

In edge copying, we choose the prototype node according to its inter-degree, while the endpoint of any edge is chosen with equal probability. Thus, the probability that a new added edge in $B(N_p, N_a)$ points to a certain node n is $\frac{d_{pa}^{ir}(t-1)}{s_{pa}(t-1)}$, where $s_{pa}(t-1)$

denotes the sum number of edges in $B(N_p, N_a)$ at time $t-1$, and we have

$$s_{pa}(t-1) = (\alpha_p c_{pa} + \alpha_a c_{ap} + \alpha_t c_{tp} c_{ta})(t-1).$$

And $s_{pt}(t-1)$ as well as $s_{at}(t-1)$ can be obtained by same method.

Combining the above two cases, we get

$$d^{ir}_{pa}(t) - d^{ir}_{pa}(t-1) = \alpha_a c_{ap} \frac{d^{ir}_{pa}(t-1)}{s_{pa}(t-1)} + \alpha_t c_{tp} c_{ta} \frac{d^{ir}_{pt}(t-1)}{s_{pt}(t-1)},$$

and similarly,

$$d^{ir}_{pt}(t) - d^{ir}_{pt}(t-1) = \alpha_t c_{ta} \frac{d^{ir}_{pt}(t-1)}{s_{pt}(t-1)} + \alpha_a c_{ap} c_{at} \frac{d^{ir}_{pa}(t-1)}{s_{pa}(t-1)}.$$

With the initial condition that

$$d^{ir}_{p}(t) = \left[\left(\frac{t}{t_0}\right)^{\lambda_p} d^{ir}_{pa}(t_0), \left(\frac{t}{t_0}\right)^{\lambda_p} d^{ir}_{pt}(t_0) \right], \tag{1}$$

where

$$\lambda_p = \frac{\sqrt{\Delta} + \alpha_a c_{ap} s_{pt} + \alpha_t c_{tp} s_{pa}}{2 s_{pa} s_{pt}}, \tag{2}$$

here, $s_{pt} = \frac{s_{pt}(t)}{t}$ is a constant, and according to the calculation result,

$$\Delta = (\alpha_t c_{tp} s_{pa} - \alpha_a c_{ap} s_{pt})^2 + 4\alpha_a \alpha_t c_{ap} c_{tp} c_{at} s_{pa} s_{pt}.$$

By same approach we can obtain the expression result of $d^{ir}_{p}(t)$ for nodes in N_a and N_t, thus we complete the proof. □

In fact, the proof of Theorem 5.1 also reflects that in N_p, $d^{ir}_{pa}(t)$ and $d^{ir}_{pt}(t)$ have the same order, as Equation (1) shows $d^{ir}_{pa}(t)$ and $d^{ir}_{pt}(t)$ have same growing function, i.e. $C(\frac{t}{t_0})^{\lambda_p}$, where C is a constant. Symmetrically, this property also holds in N_a and N_t.

Theorem 5.2. *For graph $G(P, A, T)$ generated after t time slots ($t \geq t_0$), with the condition that inter-degree in node set N_a and N_t growing with the power λ_a and λ_t, the intra-degree of $n \in N_p$ at time t satisfies*

$$d^{ia}_{p}(t) = \Theta\left(\max\left\{t^{\frac{1}{\lambda_a}+1}, t^{\frac{1}{\lambda_t}+1}\right\}\right).$$

This result also holds for $n \in N_a$ and $n \in N_t$ with symmetrical expressions.

Proof. The intra-degree in N_p is generated by common neighbors in N_a and N_t independently.

When a certain node $a \in N_a$ has node degree x from N_a to N_p, it has exactly x neighbors in N_p. Thus, the expected intra-degree in N_p added by this node is $2\gamma_{pa}\binom{x}{2}$, where γ_{pa} is the linking probability when two nodes inside node set N_p have a common neighbor node in N_a. And the number of nodes in N_a who have x neighbors in N_p is expected as $|N_a|\mathbb{P}\left\{d^{ir}_{ap}(t) = x\right\}$ where $\mathbb{P}$ denotes the probability that node in N_a having x neighbors in N_p exists and $|N_a|$ denotes the total nodes in N_a. Therefore, the intra-degree generated by nodes with x neighbors in N_a is

$$\text{Contribution}_a(x) = 2\gamma_{pa}\binom{x}{2}|N_a|\mathbb{P}\left\{d^{ir}_{ap}(t) = x\right\}. \tag{3}$$

Considering we add certain number of nodes with a certain probability in the node set, we get $|N_a| = \Theta(t)$. Thus, combining the result of Theorem 5.3, we get the intra-degree $d^{ia}_{pa}(t)$ in node set N_p contributed by node set N_a is

$$\begin{aligned} d^{ia}_{pa}(t) &= \sum_{x=1}^{\max_a} \text{Contribution}_a(x) \\ &= \sum_{x=1}^{\max_a} 2\gamma_{pa}\binom{x}{2}|N_a|\mathbb{P}\left\{d^{ir}_{ap}(t) = x\right\} \\ &= \Theta\left(\sum_{x=1}^{\max_a} x^2 x^{-\frac{1}{\lambda_a}-1} t\right) \\ &= \Theta\left(\sum_{x=1}^{t} x^{-\frac{1}{\lambda_a}+1} t\right), \end{aligned}$$

where $\max_a$ presents the maximum inter-degree from N_a to N_p which satisfies $\max_a = \Theta(t)$. By using the sum of *p-series*, we get

$$\sum_{x=1}^{t} x^{-\frac{1}{\lambda_a}+1} = t^{1-(1-\frac{1}{\lambda_a})}.$$

Therefore, we have $d^{ia}_{pa}(t) = \Theta\left(t^{\frac{1}{\lambda_a}+1}\right)$.

Considering the symmetric contribution of N_t to intra-degree in N_p, we get

$$\begin{aligned} d^{ia}_{p}(t) &= d^{ia}_{pa}(t) + d^{ia}_{pt}(t) \\ &= \Theta\left(\max\left\{t^{\frac{1}{\lambda_a}+1}, t^{\frac{1}{\lambda_t}+1}\right\}\right). \end{aligned}$$

By same approaches, we can also obtain the expression result of d^{ia}_{p} for nodes in N_a and N_t, thus we complete the proof. □

5.2 Power-Law Distribution

Power law distribution is a classical node degree distribution which can be widely found in social network structure. By our design, we can also find power law distribution in our scholarly network model and give proper mathematical proof.

Also, we analyze nodes' power-law distribution in two cases – inter and intra-degree respectively. However, there is little difference from the formal proof in 5.1, that we secondly do not study the intra-degree's case as its detailed distribution expression is hard to obtain, but study the general case when combining inter-degree and intra-degree together, i.e. the total degree of the nodes since when time slot $t \to \infty$.

Distribution of inter-degree: For the node $n \in N_i$ in $G(P, A, T)$ with $t \to \infty$, the inter-degree distribution of it follows

$$\mathbb{P}\left\{d^{ir}_{ij}(t) = x\right\} \propto x^{-\frac{1}{\lambda_i}-1}.$$

And we find that the inter-degree d^{ir} follows the power-law distribution with exponent $-\frac{1}{\lambda_i} - 1$.

Results show our model well capture the power-law distribution of nodes' inter-degree, which are proved in Theorem 5.3 and verified by experimental measurements.

Distribution of total degree: For the node n in $G(P, A, T)$ with $t \to \infty$, the total degree distribution of $n \in N_i$ follows

$$\mathbb{P}\{d_i(t) = x\} \propto x^{-\omega_i},$$

where ω_i is a constant which describes the exponential factor in power-law distribution.

This means our model well simulates the power-law distribution of nodes' total degree. And results are proved in Theorem 5.4.

Theorem 5.3. *For graph $G(P, A, T)$ generated after t time slots, when $t \to \infty$, the inter-degree sequences of $n \in N_p$ in $B(N_p, N_a)$ and $B(N_p, N_t)$ both follow power-law distribution that*

$$\mathbb{P}\left\{d_p^{ir}(t) = x\right\} \propto x^{-\frac{1}{\lambda_p}-1},$$

where x is one node's total degree and $\mathbb{P}$ presents the probability. This result also holds for node $n \in N_a$ and $n \in N_t$ as they share symmetrical expressions.

Proof. We also divide the proof into two parts, i.e. prove the power-law in $d_{pa}^{ir}(t)$ and in $d_{pt}^{ir}(t)$ separately. First of all, we consider the distribution of $d_{pa}^{ir}(t)$ which denotes the degree of node $n \in N_p$ in $B(N_p, N_a)$. According to Equation (1), the cumulative distribution function of $d_{pa}^{ir}(t)$ can be calculated as

$$\begin{aligned}\mathbb{P}\left\{d_{pa}^{ir}(t) < x\right\} &= \mathbb{P}\left\{d_{pa}^{ir}(t_0)\left(\frac{t}{t_0}\right)^{\lambda_p} < x\right\} \\ &= \mathbb{P}\left\{t_0 > \left(\frac{d_{pa}^{ir}(t_0)}{x}\right)^{\frac{1}{\lambda_p}} t\right\} \\ &= 1 - d_{pa}^{ir}(t_0)^{\frac{1}{\lambda_p}} x^{-\frac{1}{\lambda_p}}.\end{aligned}$$

Then, the probability density function of $d_{pa}^{ir}(t)$ can be calculated using $\mathbb{P}\left\{d_{pa}^{ir}(t) = x\right\} = \frac{\partial \mathbb{P}\{d_{pa}^{ir}(t) < x\}}{\partial x}$. Also, it can be expressed as

$$\mathbb{P}\{d_{pa}^{ir}(t) = x\} = \frac{x^{-\frac{1}{\lambda_p}-1}}{\sum_{x=1}^{n} x^{-\frac{1}{\lambda_p}-1}},$$

where $\sum_{x=1}^{n} x^{-\frac{1}{\lambda_p}-1}$ is a constant normalization coefficient. Therefore, we get

$$\mathbb{P}\left\{d_{pa}^{ir}(t) = x\right\} \propto x^{-\frac{1}{\lambda_p}-1},$$

By same approaches, we can also calculate the distribution of $d_{ij}^{ir}(t)$, where $i \neq j \in \{p, a, t\}$ and thus the proof is complete. □

Theorem 5.4. *For graph $G(P, A, T)$ generated after t time slots, when $t \to \infty$, the nodes' total degree sequences of $n \in N_p$ follow power-law distribution that*

$$\mathbb{P}\{d_p(t) = x\} \propto x^{-\omega_p},$$

where x is one node's total degree, $\mathbb{P}$ presents the probability and ω_p is a constant. This result also holds for node $n \in N_a$ and $n \in N_t$ as they share symmetrical expressions.

Proof. The proof uses the result of *Silvio Lattanzi and D. Sivakumar*'s research work. [21]. In their work, the model's bipartite network's structure is similar to our model's bipartite networks' which are disconstructed from $G(P, A, T)$.

And by Theorem 4 and Theorem 8 in their paper, they fully prove the total degree distribution is similar to the inter-degree distribution when time slot $t \to \infty$. Which means the total degree is also power-law distributed.

Therefore, the total degree distribution in our model follows

$$\mathbb{P}\{d_p(t) = x\} \propto x^{-\omega_p},$$

where ω_p is a constant.

Using same methods, we can obtain the distribution for node $n \in N_a$ and $n \in N_t$ and thus complete the proof. □

5.3 Densification

Now we turn to analyze the property of densification, defined as the phenomenon that the network's density of $G(N_i|N_j)$ – generated graph of N_i obtained from $B(N_i, N_j)$, increases with time if $\lambda_j \geqslant \frac{1}{2}$. The detailed proof is presented in Theorem 5.5.

Theorem 5.5. *For graph $G(P, A, T)$ which is generated after t time slots. The ratio of edges to nodes in $G(N_i|N_j)$ is*

$$\frac{|E|}{|V_i|} = \begin{cases}\Theta(1), & 0 < \lambda_j < \frac{1}{2} \\ \Theta(\log t), & \lambda_j = \frac{1}{2} \\ \Theta\left(t^{2-\frac{1}{\lambda_j}}\right), & \frac{1}{2} < \lambda_j < 1.\end{cases}$$

Proof. According to the definition of $G(N_i|N_j)$, each node $v \in N_j$ in $B(N_i, N_j)$ becomes a clique where all neighbors of v are connected with probability γ_{ij}. Therefore, the average number of edges in $G(N_i|N_j)$ is

$$|E| = \sum_{k=1}^{n} n \frac{k^{-\frac{1}{-\lambda_j-1}}}{G_j} \gamma_{ij} C_k^2, \tag{4}$$

where $n\frac{k^{-\frac{1}{-\lambda_j-1}}}{G_j}$ is the average number of nodes with degree k in N_j. Since $|N| = n$, we can use the sum of *p-series* to get final results:

$$\lim_{n\to\infty} \sum_{x=1}^{n} \frac{1}{x^p} = \begin{cases}\Theta(1), & p > 1 \\ \Theta(\log n), & p = 1 \\ \Theta(n^{1-p}), & 0 \leqslant p < 1.\end{cases} \tag{5}$$

Therefore, we complete our proof. □

6 SIMULATION

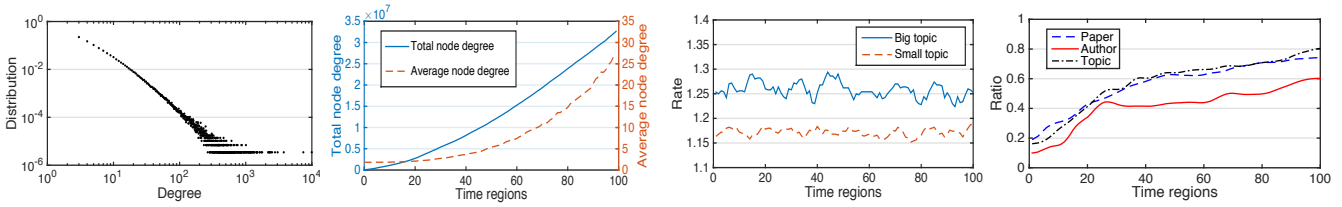
Upon theoretical analysis of our model, in this section, we present simulations to verify that our evolving scholarly model can correctly extract the properties of real scholarly networks.

According to the ratio of paper, author, and topic count in four datasets, we set $\alpha_p = 0.4136$, $\alpha_a = 0.5862$, $\alpha_t = 0.0002$. And other parameters are set as: $c_{ij} = 2$, $\beta_{ij} = 0.2 (i \neq j)$. Then after 2 million time slots, we get 827636 papers, 1171977 authors and 386 topics.

In real scholarly networks, we notice a tendency that during each year, more nodes are added to the network than last year. However, in our simulation, every equal time gap, we add the same number of nodes. To revise the influence of this factor and for the convenience of analysis, we divide the 2 million time slots into 100 time regions unequally. Each time regions can be reckoned as a "year" compared to real datasets. Based on the synthetic dataset, we verify the validity of our model and the result is listed below:

(1) Degree distribution: Figure 9(a) shows the power-law distribution of paper degree in citation sub-network with $\varphi = 1.82$ and $\eta = 0.21$. As for the inter-degree, they all follow the power-law distribution and we give the final factors in table 4.

(2) Intra-set Evolution: As expected, in Figure 9(b), total node degree and average node degree in co-author network both increase exponentially in the evolving process.



(a) Degree distribution *(b) Growth of node degree* *(c) Topic-size evolution* *(d) Factors of power-law distributions*

Figure 9: Property verification of evolving scholarly model

Table 4: Simulation of power-law distribution factors

Factors	D_{pa}	D_{pt}	D_{ap}	D_{at}	D_{tp}	D_{ta}
φ	2.19	3.61	2.30	3.82	0.23	0.26
η	0.28	2.01	0.09	0.65	2.18	2.08

(3) Inter-set Evolution: From Figure 9(c), we again, acknowledge that big topic has the ability to increase its size faster than small topic. In our simulation, the average growth speed rate of big topic is 1.26 while the small topic is 1.17.

(4) Co-evolution of the Three sets: As we can see from Figure 9(d), the connectivity of each element set changes simultaneously in a same pattern. So we can draw a conclusion that paper, author, and topic elements co-evolution with each other in scholarly networks.

In general, the results based on our evolving scholarly model perform well in the simulation process, which verifies that our evolving scholarly model capture the properties of scholarly networks.

7 CONCLUSIONS

In this paper, we have presented the first comprehensive study of the scholarly network which is a structure extracted from massive scholarly data. Using Microsoft's datasets, we provide a first-principled understanding of scholarly networks and their evolution. We observe several interesting phenomena in the structure and evolution of scholarly networks. For example, the inter-correlation including simultaneous co-evolution of all the three sets and faster growth rate of elements' size, as well as intra-correlation such as special power-law distributed degree and degree's densification, manifest themselves in the network structure. Building on these empirical observations, we provide a new and novel evolving scholarly model that jointly captures both intra and inter correlations of papers, authors and topics during the evolving process. Through both theoretical analysis and empirical evaluations, we further validate that our model can accurately reproduce the global and local structures of real scholarly networks.

We believe that our work is one of the first steps in analyzing scholarly information in a general perspective, and that there are several interesting directions for future work to further harness the power of using big scholarly data.

REFERENCES

[1] Microsoft academic graph, 2016. https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

[2] Jason Priem. Scholarship: Beyond the paper. *Nature*, 495(7442):437–440, 2013.

[3] Feng Xia, Wei Wang, Teshome Megersa Bekele, and Huan Liu. Big scholarly data: A survey. *IEEE Transactions on Big Data*, 2017.

[4] Mark Newman. Networks: an introduction. 2010. *United Slates: Oxford University Press Inc., New York*, pages 1–2, 2010.

[5] Wei Wang, Jiaying Liu, Shuo Yu, Chenxin Zhang, Zhenzhen Xu, and Feng Xia. Mining advisor-advisee relationships in scholarly big data: A deep learning approach. In *Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference on*, pages 209–210. IEEE, 2016.

[6] Elena Zheleva, Hossam Sharara, and Lise Getoor. Co-evolution of social and affiliation networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1007–1016. ACM, 2009.

[7] Zaihan Yang, Liangjie Hong, and Brian D Davison. Academic network analysis: A joint topic modeling approach. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 324–333. ACM, 2013.

[8] Yu-Shan Lin. Topic evolution of innovation academic researches. *Journal of Small Business Strategy*, 26(1):25, 2016.

[9] Claudia Colicchia, Alessandro Creazza, and Fernanda Strozzi. Citation network analysis for supporting continuous improvement in higher education. *Studies in Higher Education*, pages 1–17, 2017.

[10] Jing Li, Feng Xia, Wei Wang, Zhen Chen, Nana Yaw Asabere, and Huizhen Jiang. Acrec: a co-authorship based random walk model for academic collaboration recommendation. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 1209–1214. ACM, 2014.

[11] Yuya Kajikawa, Junko Ohno, Yoshiyuki Takeda, Katsumori Matsushima, and Hiroshi Komiyama. Creating an academic landscape of sustainability science: an analysis of the citation network. *Sustainability Science*, 2(2):221, 2007.

[12] Martin Atzmueller, Andreas Ernst, Friedrich Krebs, Christoph Scholz, and Gerd Stumme. On the evolution of social groups during coffee breaks. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 631–636. ACM, 2014.

[13] Yanhong Wu, Naveen Pitipornvivat, Jian Zhao, Sixiao Yang, Guowei Huang, and Huamin Qu. egoslider: Visual analysis of egocentric network evolution. *IEEE transactions on visualization and computer graphics*, 22(1):260–269, 2016.

[14] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM, 2005.

[15] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, volume 29, pages 251–262. ACM, 1999.

[16] Duncan J Watts and Steven H Strogatz. Collective dynamics of fismall-worldfinetworks. *nature*, 393(6684):440–442, 1998.

[17] Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170. ACM, 2000.

[18] Deepayan Chakrabarti and Christos Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM computing surveys (CSUR)*, 38(1):2, 2006.

[19] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[20] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D Sivakumar, Andrew Tomkins, and Eli Upfal. Stochastic models for the web graph. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 57–65. IEEE, 2000.

[21] Silvio Lattanzi and D Sivakumar. Affiliation networks. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 427–434. ACM, 2009.

[22] Albert-László Barabási, Réka Albert, and Hawoong Jeong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1):173–187, 1999.

[23] Lev Muchnik, Sen Pei, Lucas C Parra, Saulo DS Reis, José S Andrade Jr, Shlomo Havlin, and Hernán A Makse. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *arXiv preprint arXiv:1304.4523*, 2013.

[24] Cong Xie, Ling Yan, Wu-Jun Li, and Zhihua Zhang. Distributed power-law graph computing: Theoretical and empirical analysis. In *Advances in Neural Information Processing Systems*, pages 1673–1681, 2014.

[25] Michael Molloy and Bruce Reed. The size of the giant component of a random graph with a given degree sequence. *Combinatorics, probability and computing*, 7(03):295–305, 1998.