Big Data Analysis on Cross-domain Scholarly Data

Lingkun Kong, Bo Wang, Jiaqi Liu, Luoyi Fu, Xinbing Wang
Shanghai Jiao Tong University, China

***Abstract*—Interdisciplinary collaborations have generated huge impact to society. And the rise of collaborations among different scientific domains inevitably becomes the trend in scientific research. The cross-domain scholarly data is the vehicle which properly portray features of these interdisciplinary collaborations.**

Analysing the cross-domain scholarly data helps researchers better sense the abstract features as well the patterns of interdisciplinary collaborations, which has important implications in many aspects, such as wiser design of scholar recommendation systems, better evaluation of research communities, more accurate prediction of scientific domain developing trends and etc. However, due to theoretical and technical difficulties, there have been few studies that provide a systematic and practical understanding of cross-domain scholarly data at scale. Particularly, many papers which study the cross-domain recommendation systems make modeling assumptions without solid experimental observations in real-world scholarly data.

We bridge this gap using real scholarly datasets – *Microsoft Academic Graph* [?] with 126 million papers collected from around 50 thousand domains. By empirical exploration, we observe novel features that belong exclusively to cross-domain scholarly data, such as four power-law distributions in the relationship between scientific papers and papers' domains, the *peak influence* cross-domain collaborations add to papers' quality, i.e., citation. We also observe interesting evolving patterns among different domains' co-paper relationship, and further make case study in the domain of "Data-mining" after adding time information. Moreover, we dig into the papers' citation structure in the perspective of cross-domain distribution, and more accurately study papers' cross-domain performance or influence by giving the distance between different domains to quatify the domains' closeness relationship.

Based on our empirical observations, we also make efforts in proposing novel models that can well reproduce the properties or patterns we discovered. To illustrate, we design a model of cross-domain power-law (*MCP*) to captures the power-law distributions in cross-domain data. And we reproduce the *peak influence* by the help of guassian distribution. Moreover, through both theoretical analysis and empirical evaluations, we demonstrate that our models can accurately reproduce the features as well the patterns we probe in real-world dataset.

I. Introduction

Studying properties of scholarly networks and getting insight of the scholarly data have important implications. Despite the importance of scholarly networks in many kinds of applications, there have been few studies at observation of relationship among different scientific domains, as well the affiliation domains add to paper which changes papers influential factor, due to the paucity of big data and the difficulty of big data analysis.

In this paper, we bridge this gap by implementing elaborate data-mining methods on big scholarly data.

First of all, we properly study a database – *Microsoft Academic Graph* (MAG) of large-scale which thoroughly explores the multi-hierarchy of different domains and extract the scholarly data about cross-domain properties.

Based on the massive data, on the one hand, we probe the subordination among domains, and further study the boundaries among domains as well the evolving pattern of domains relationship. On other hand, we explore the paper's cross-domain performance, which includes its membership relation based on domains paper belong to and its cross-domain citation distribution.

In the domain-oriented perspective, we firstly observe the domain's number power-law distribution with its papers number as well its subdomain number. And we also study the closeness among different domains, i.e. using co-paper's ratio to describe the relationship between different domain. Moreover, we add time information to the relationship we try to explore, by mining paper's publication date, and thus find interesting evolving pattern in domains' relationship.

From the paper-oriented or literature-oriented perspective, we explore the paper's cross-domain performance, which includes its membership relation based on domains paper belong to and its cross-domain citation distribution. We firstly get the power-law distributed paper number with the number of domains paper belongs to. And then, we based on paper's membership relation with domains it belongs to, study paper's citation distribution. And surprisingly we get a "peak" distribution – the paper's citation number is likely to get a maximum value when paper's domain number comes to a certain amount. Further, we dig into the paper's citation, dividing these citation into four parts according to the network structure of paper's membership hierarchy, and thus know the decreasing citing possiblity with the increasing cross-domain distance, which complies with our intuitive thinking.

And in this part, we implement elaborate visualization methods to clearly present our observation.

Further, based on our observation on real world database, we purpose theoretical model to explain and simulate our observing result. And we also explore the substaintial reason behind the data pattern.

In general, this work helps to judge or explain the relation and the boundary between different domains. For instance, we can explore the similarity and distinction of Literature and Mathematic, two domains largely different from each other in common sense. Moreover, when adding publication date as time slot, we can explore the evolving pattern of the domains relationship. Besides, according to the performance of literatures cross-domain studying, the literatures depth and breadth can be well measured by properties of domains which they belong to and their cross-domain citation distribution, which affords researchers more accurate browsing results when they want to wade into a new scientific field.

II. Related Work

Related work

III. Observation And Visualization

In this part, by the help of data-mining methods, we properly explore the cross-domain scholarly data in real world big scholarly data, and discover several stimulating results. Further, we implement visualization to our discoveries to make our result easy to view.

We make our observation both on domain-oriented aspect and paper-oriented or literature-oriented aspect. In the domain's perspective, we observe firstly the domain's number power-law distribution with its papers number as well its subdomain number. And we also study the closeness among different domains, i.e. using co-paper's ratio to describe the relationship between different domain. Moreover, we add time information to the relationship we try to explore, by mining paper's publication date, and thus find interesting evolving pattern in domains' relationship.

From the paper's perspective, we explore the paper's cross-domain performance, which includes its membership relation based on domains paper belong to and its cross-domain citation distribution. We firstly get the power-law distributed paper number with the number of domains paper belongs to. And then, we based on paper's membership relation with domains it belongs to, study paper's citation distribution. And surprisingly we get a "peak" distribution – the paper's citation number is likely to get a maximum value when paper's domain number comes to a certain amount. Further, we dig into the paper's citation, dividing these citation into four parts according to the network structure of paper's membership hierarchy, and thus know the decreasing citing possiblity with the increasing cross-domain distance, which complies with our intuitive thinking. And we also studies successor phenomenon ...

A. Brief Introduction and Study For MAG

We give our experiments based on *Microsoft Academic Graph* (MAG) which is an 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.

As we study the cross-domain performance of scholarly data, we mainly launch research on the scholarly data related with fields of study, i.e. domains in MAG. And the fields in MAG can be divided into four layers and we call them from L0, L1, L2, L3, where layer with lower number, which represents bigger scientific domain contains the layer with higher number, i.e. smaller scientific domain of study. E.g., we get a domain labeled with L0 layer in MAG called "Computer Science", which contains several domains labeled by L1 layers according to the MAG's hierarchy table, including "Artificial Intelligence", "Database", "Data-mining", "Computer Hardware", and etc. And "Data-mining" can contains several domains with lower layer label such as " Big data" in L2 Layer and "K-optimal pattern discovery" in L3 layer. Moreover, one thing is supposed to be noticed is that the hierarchy in MAG is heterogeneous which means the domain of L1 layer can directly contains or relates domains of L2 and L3 layers.

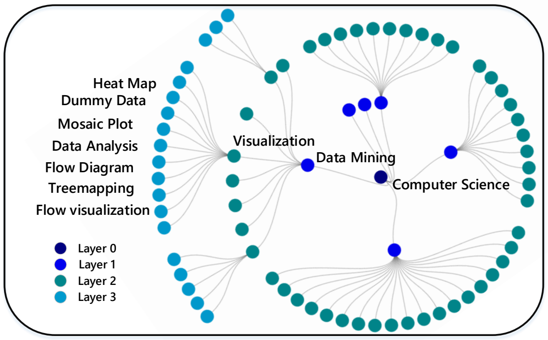


Fig. 1. Domain Hierarchy Structue in MAG

In figure 1, hierarchy example of MAG dataset can be viewed. This figure illustrates a part of the hierarchy struct of the domain "Computer Science". We pick up several nodes and mark their names besides the nodes. As can be seen from the figure, the lower the layer is, the more specifically the domain is.

B. Domain-oriented Exploration

Power-law distributed degree is a common feature of social networks, which is also well studied by many existing literatures. And XXX's work also find the power-law distribution also exists in scholarly network – using network structure to present the scholarly data. Stimulated by this result, when we studying domains' information in scholarly data, we also get power-law distribution. In figure 2, it is clearly can be viewed that there exists two kind of power-law distribution.

First is power-law distributed domains number with papers' count in these domains, as shown in the figure 2.a. And second

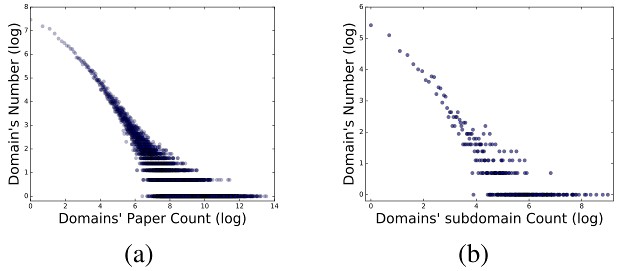


Fig. 2. Domain-oriented power-law distribution

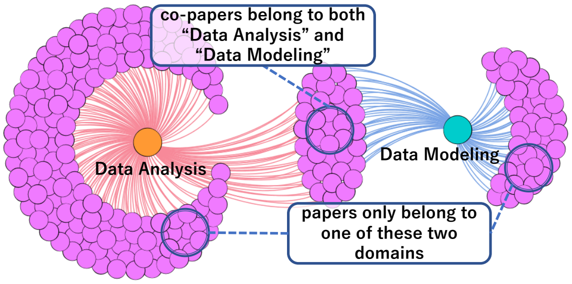


Fig. 3. The paper number distribution among target domains. Where are two domains in L3 layer, called "Data Analysis" and "Data Modeling".

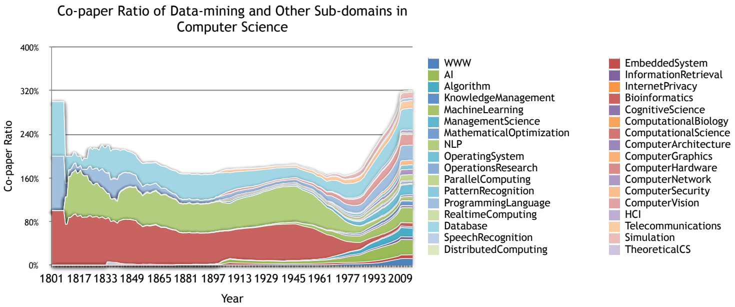


Fig. 4. Domain relationship in "Computer Science"

is power-law distribution bwtween domains number with their sub-domains' count, which is drawn in figure 2.b.

We intuitively use co-paper's number among different domains to describe the correlation of domains, i.e. the closeness of relationships. Though behind this papers, a more complex networking topology, i.e. the paper's reference and citation network might include useful information for judging closeness between different domains, we originally use the number of co-papers to verify the correlation of those domains as co-paper is the bridge which levels up the gap among different domains and the number of co-paper, the basic and essential feature of co-papers can linearly evaluate the closeness among domains.

First of all, we simply visualize this relationship in figure 3. And we can clearly find that different domains in fact have many co-papers. And therefore we can use the co-paper information to quantify the relationship between different domains.

Besides, we can add the time information by the data in MAG's paper publication time table to further explore the evolving pattern of the relationship among different domains.

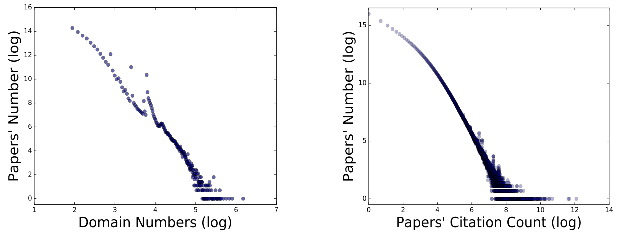
What we do is to label every paper of two domains by their publication date, and thus co-papers are also labeled by time info. After that, we calculate the evolving co-paper ratio of one specific domain vs other different domains. And by that, we can find the changing relationship among one domain and other different domains. For instance, in figure 4, we study in the "computer science" domain, all L1 domains' relationship with "data-mining" domain.

In this figure, we can view the evolving relationship – the fluctuating correlation in different years among Data Mining and other different computer science related domains, such as AI, Algorithm, NLP and etc. which are listed in the figure's legend. By this figure, the affiliation of "Data Mining" domain is revealed, i.e. we can see the switching closeness of data mining to other domains. For instance, in the early stage, "Data Mining" is highly related with "Bio-information" – "Data Mining" has almost 90% co-papers with "Bio-information" which refers high correlation while in current years, the co-papers' ratio of "Bio-information" has been decreased to a very low level, which indicates the degeneration of relationship between "Data Mining" and "Bio-information". Moreover, it can be viewed that recently the "Data Mining" domain has always preferred to combine knowledge in publication from "Artificial Intelligence" and "Machine Learning" domain as their co-paper ratio is rapidly growing.

C. *Paper-oriented Exploration*

We also explore several features of paper's cross-domain performance, including papers' domain distribution, and paper's cross-domain citation distribution. First of all, we find that the paper's number is power-law distributed depending on paper's domain number.

In figure 5, the left graph presents relation bwteen Art's paper domains count and paper number. Though the domain count is not large, it seems to be still power-law distribution. Right graph is computer science's. This graph more smoothly simulates the power-law distribution since there are much



(a) (b)

Fig. 5. Paper-oriented power-law distribution

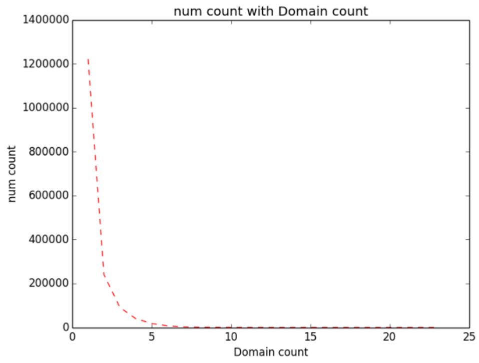


Fig. 6. Power-law in "AI"

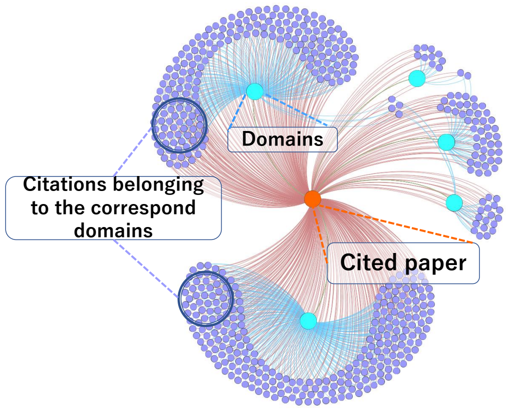


Fig. 7. Paper's citation domain structure

more papers in computer science and much more cross-domain collaboration.

Figure 6 is the power-law of AI (L1 layer). We find that the paper has fewer domains as we only take AI's fields into count – the AI contains several smaller L2 and L3 domains.

Paper citation distributions

We believe cross-domain paper citation performance is an essential indicator for paper's quality, i.e. whether the paper is good or not, since a good paper might cross several different domains as nowadays research emphasizes on the study's width. More specifically, a scientific study combining several fields of studies knowledge together, might be more likely to catch others' attention and generate stimulating results. For instance, paper in computer science domain with solid mathematical foundation or theory, i.e. also in mathematical domains are always more likely to be a good paper. Moreover, recently, scientific study launched in biology domain make many refreshing breakthroughs by combining computer science, especially data mining and machine learning's knowledge.

We extract papers' citation information and use their citations domain information to draw paper citation distribution map, which intuitively study the structure of paper citation distribution with domain information.

In figure 7, the green node in the center of the graph indicates the paper we want to study, and the pink node represents paper which cite the paper we study. The light green node refers to L1 layer domain, red node is L2 layer domain while blue is the L3 layer domain. And the line between green node to pink node means the paper's citation, while other lines among papers to domains describes membership relation between papers and domains. More specific information need to be updated.

What we can see is that the paper's citation can be divided into several cliques according to the paper's domain info. And there exists clique overlapping as citation papers might have more than single domains which are similar with the paper being cited. And we consider these papers to be successors of the original paper, which we will discuss later as these papers are more likely to be papers which inherit studying paper's idea or methods. successor phenomenon

Paper's average citation distribution over domain number.

We believe that paper's citation is related to its domain count, as the paucity of domains might constrain a paper's impact in a small domain, while too many domains might decrease paper's quality as too many domains might distract or diffuse author's attention and harms the paper's depth. Therefore, we plot the paper's average citation number over paper's domain count in below figures.

Figure 8 is the overall paper's citation count over their domain count. And we can clearly find a peak when paper's domain count goes to about 50. And we check these paper's number in case paper's number is too small to represent a pattern when paper's domain count exceeds a certain number, e.g. 50. And we find when paper's domain count is less

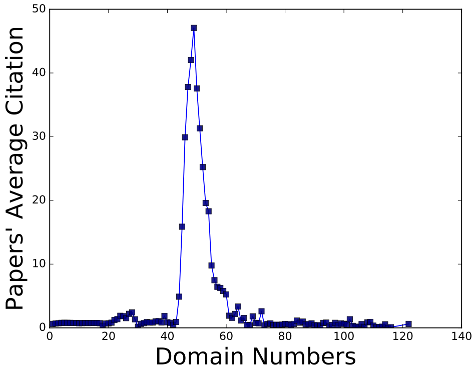


Fig. 8. Papers' average citation count with certain domain numbers of overall paper in the database

than 100, the paper's number is large enough to calculate the average citation count. In other words, this ”peak” pattern does exist.

Moreover, when we look into smaller domains, computer science's sub-domains, for instance, we find that this rule is almost generally valid. And in fact though in some subdomains, theoretical cs for instance, though paper's domain count is small, the peak still exist.

In figure 9, we can easily view these properties.

Paper's citation performance over more specific domains Further, we look into the paper's citation distribution. And we divide one paper's citation into several different types according to the cross-domain step. Intuitively speaking, the cross-domain step is a parameter we use to evaluate the paper's cross-domain distance. For instance, a paper in ” Astrology” domain cites paper in ”Data-mining” has longer cross-domain distance than paper in ”Data Base” cite paper in ”Data-mining”. But how to quantify this difference? We take advantage of our dataset's level partition. And we set cross-domain step into four class, i.e. merge at L3 layer, L2 layer, L1 layer and L0 layer. Noticed that a paper may belong to sevral domains, the domains of two papers can merge at different layers. We choose the lowest merge layer as the cross-domain distance since the domains at lower layer can represent the paper more specifically. figure 10 presents our dividing rules:

In figure 10, the domains at layer 3 such as ”Graphical Tools” and ”Data Analyze” merge at layer 2, while ”Graphical Tools” and ”Network Simulation” merge at layer 1. If the cited paper and its citation belongs to ”Graphical Tools” and ”Network Simulation” respectly, their domains merge at layer 1 and layer 0, and we chose the lowest merge layer, i.e. layer 1, as the merge layer of this citation. And according to our rules, it is easily can be found that a citation type is supposed to only belong to one class. And the possibility of a citation belongs to these 4 class should be summed into exactly 1. Drawbacks: we currently cannot figure the distance between math to cs and art to cs which is longer.

As we can see in figure 11, we pick up some papers with high citations to draw this distribution and the citation in cross-L3 type's possibility is much higher than cross L1 and cross L2 type's. And in fact cross L2 type's possibility is slightly higher than cross L1 types. And the cross L0's possibility is very low, almost 0.

In figure 12 we calculate the average cross domains ratio in several domains. The result shows that the citations of a paper are tends to be in a relatively close distance of the cited paper. In some domains such as ”Literature”, the ratios of cross L1 level and L2 level are much lower than other domains, which indicates that in those domains, the interdisciplinary trend is not so strong.

successors In figure 6, we find some citation papers which might have more than one domains that overlap with the fields of cited paper. This phenonmenon exposes that the cited paper introduces more papers to cross the domains as the cited paper does. It illustrates the real ability of the cited paper — the ability to lead more papers to be interdisciplinary. So we call this phenonmenon as successor phenonmenon.

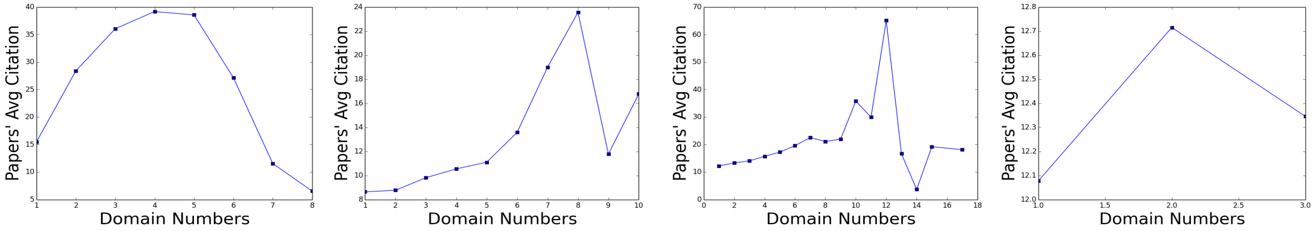
In figure 13, we count the number of citation papers averagely for one cited paper and the number of the corresponding overlaping domains with the cited paper. We find that successors which have more overlaping domains with the cited paper are relatively less. It means that even the cited paper crosses many domains, only a few citation papers cross the same domains as the cited paper does. Waiting for clearer explaination and more graphs

IV. Modeling and Analysis on Observation

In this section, based on the previous observations in real world dataset, we use three models to help to simulate or evaluate the scholarly data's cross-domain performance.

A. *Power-law Distribution Modeling*

In our observation, it can be viewd clearly that there exists two kinds of power-law distributions in our study. The first is the power-law distributed domain's number with paper's number in this domain. And the second is the power-law distribution between the paper's number and the this paper's relative domains' number. And here we construct a evovlving model which can properly reproduce these power-law distribution in



(a) Cognitive Science

(b) Data-mining

(c) Machine Learning

(d) Theoretical CS

Fig. 9. Papers' average citation numbers under certain domain number in Computer Science's subdomains

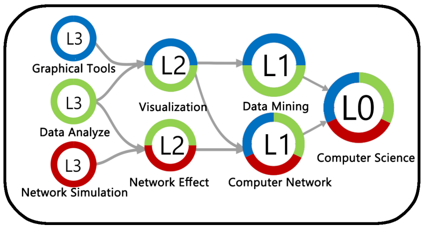


Fig. 10. The example of domains merging at different layers

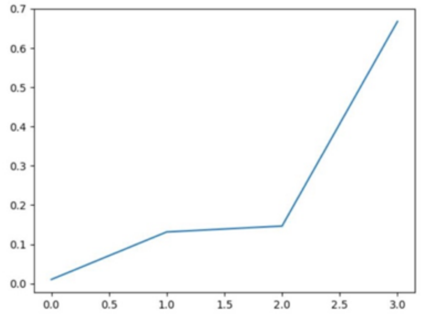


Fig. 11. 1000 papers cross L0, L1, L2, L3 domains ratio distribution

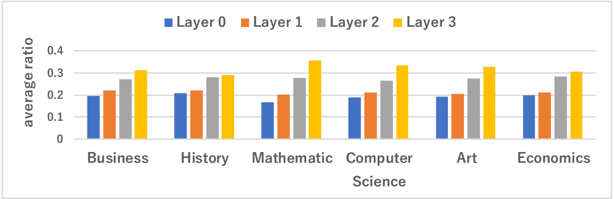


Fig. 12. The average cross domains ratio distribution over L0 domains

the cross-domain scholarly data. And we call this model as *model of cross-domain power-law* – MCP.

MCP Construction:

We use network structure to present the scholarly data. And in MCP, the graph is denoted as $G(P, D)$. Then, we use bipartite 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 IV-A. It contains:

(1) Two node sets: Paper node set N_p, and domain node set N_d. The node in each node set is marked as n_p and n_d.

(2) Inter-edge sets: We denote the edges between every two node sets as inter-edge sets. And, we refer all edges between paper node and domain node as $E_{pd}(E_{dp})$, then an edge $e_{n_pn_d}$ which belongs to E_{pd} means paper n_p belongs to the domain n_d or equivalently domain n_d has paper n_p.

(3) Two intra-edge sets: Intra-edge is the edge in the same node set, and our graph has two intra-edge sets, which we refer as E_{pp} and E_{dd}. If an edge $e_{n_p^i n_p^j} \in E_{pp}$, then we know that paper n_p^i and n_p^j have reference or citation relationship. While if an edge $e_{n_d^i n_d^j} \in E_{dd}$, then we know that domain n_d^i and n_d^j are directly connected in the domain hierarchy dataset.

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

TABLE I
NOTATIONS AND DEFINATIONS

Notations	Definations
N_p, N_d	Node set of Paper and Domain
E_{pd}	Inter-edge set between nodes in N_p and N_d
E_{pp}, E_{dd}	Intra-edge set of N_p and N_d
α_p, α_d	Probability that a new node arrives in N_p and N_d
β_p, β_d	Probability that an edge added in set E_{pp} and E_{dd}
c_{pd},c_{dp}	Number of edges added to set E_{pd} at one time slot
$G(P, D)$	Graph of our cross-domain scholarly model
$B(N_p, N_d)$	Bipartite graph with sets N_p, N_d, and E_{pd}

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 $i \neq j \in \{p, d\}$, 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 or a domain, is added to the graph. For clarity, here we only take the arrival of a new paper as example for explanation of the subsequent steps. And the symmetry also holds for the domain. **2)** With probability proportional to degree in $B(N_p, N_d)$, paper node n_p^d is chosen as prototype for the new node n_p. **3)** c_{pd} neighbors (n_d^1 ,..., $n_d^{c_{pd}}$) of n_p^d in N_d are randomly chosen to have connections with node n_p. **4)** c_{pd} edges are added between n_d^1 ,..., $n_d^{c_{pd}}$. **5)** Edges between every two paper nodes are added with probability β_{pd} if they have a common domain.

For a better intuitive understanding of this evolving process, let us, for instance, consider the arrival of a new paper. This paper is likely to learn from an influential paper, which is thus selected as a prototype and influences the new paper on choosing research domains. To illustrate, this new paper will have high possibility to generate studies in the same domains like the old, influential paper. Besides, the papers belongs to the same domain are often relevant, indicating that these papers belong to these domains with a higher possibility to be connected, i.e. cite each other than those belong to different domains.

Similarly, when a new topic emerges in the literature, it is usually inspired by some existing topics (prototypes) and these topics are more likely to be related by the same papers they have.

```
Algorithm 1 Evolving Process
Parameters: Simulated time steps: $T$, Fixed probability $\alpha_i$
    that a new node arrives in $N_i$, fixed $\beta_{ij} \in (0, 1)$ and
    integers $c_{ij} > 0$ where $i \neq j \in \{p, d\}$.
Initialisation: In initial graph, the node in paper set has
    a certain number of neighbors with domain set. For
    example, a paper node $n_p$ connects to at least $c_{pd}$ domain
    nodes. So the inter-edge set $E_{pd}$ has at least $c_{pd}$ edges in
    the beginning.
for $1 \leq t \leq T$ do
    1) Node arrival: According to $\alpha_p$, $\alpha_d$, we decide the
    type of node to join the graph. In later discussion, we
    take the arrival of a new paper node $n_p$ as example, and
    the symmetry also holds for domain.
    2) Preferentially chosen ProtoType: A node $n_p^d \in N_p$ is
    chosen as prototype for the new node, with probability
    proportional to its degree in $B(N_p, N_d)$.
    3) Edge copying: $c_{pd}$ edges are copied from $n_p^d$, that is,
    $c_{pd}$ neighbors of $n_p^d$, denoted by $n_d^1$ ,..., $n_d^{c_{pd}}$ in $N_d$ are
    chosen uniformly at random, and the edges ($n_p, n_d^1$), ...,
    ($n_p, n_d^{c_{pd}}$) are added to the graph.
    4) Evolution inside: For every two nodes $n_p^x$ and $n_p^y$
    ($x \neq y$), if they have a common domain, then with
    probability $\beta_{pd}$, an edge ($n_p^x$, $n_p^y$) is added in $E_{pp}$.
end for
```

B. Peak

C. Evaluation of Paper's Influence's Broadness

V. Theoretical Analysis

In this section, we mathematically analyze our MCP model and the peak model to confirm that our models can well reproduce properties in the real-world database, i.e. the scholarly network in this paper.

A. MCP Analysis

Here we prove that our MCP can well reproduce two kinds of the power-law distributions in our observation.

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_{pd}, we call it d^{ir} – ir here means inter. And the second is the *intra-degree*, i.e. the node degree inside node set, related with the growth of E_{ii}, i.e., E_{pp} or E_{dd}, we call it d^{ia} – ia here means intra.

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, and $i \in \{p, d\}$.

In fact, an implications can be deduced by this result, that the inter-degree $d_i^{ir}(t)$ grows with polynomial rate in time t, following the power $\lambda_i \in (0,1)$.

This implication gives the growth rate of node's inter-degree. And the detailed proof is given in Theorem 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

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

where λ_j represents the constant λ in N_j. For instance when i is p, i.e., the paper, the j represents the d, i.e., the domain.

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 2.

Also, we analyze nodes' power-law distribution in two cases – inter and intra-degree respectively.

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

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

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 3 and verified by experimental measurements.

Distribution of intra-degree: For the node n in $G(P, D)$ with $t \to \infty$, the intra-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' intra-degree. And results are proved in Theorem 4.

Combining above four results together, it can be easily viewed in our model – MCP that the degree of the node in graph $G(P, D)$ grows with polynomial rate in time t, and the growth rate differs from inter-degree to intra-degree of the node. Moreover, the inter-degree as well the intra-degree are proven to be powerlaw-distributed in MCP. Therefore, our MCP model can well simulate and reproduce our observation in real-world database.

Theorem 1: For graph $G(P, D)$ 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_d, 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 $n \in N_d$ with symmetrical expressions.

Proof: At each time slot t, the inter-degree of node $n \in N_p$ in $B(N_p, N_d)$, i.e. $d_p^{ir}(t)$, can only increase in follow case: a new node arrives at N_d and is connected to n, which results in $d_p^{ir}(t) = d_p^{ir}(t-1) + 1$.

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 $B(N_p, N_d)$ points to a certain node n is $\frac{d_p^{ir}(t-1)}{s_p(t-1)}$, where $s_p(t-1)$ denotes the sum number of edges in $B(N_p, N_d)$ at time $t-1$, and we have

$$s_p(t-1) = (\alpha_p c_{pd} + \alpha_d c_{dp})(t-1).$$

Then, we get

$$d_p^{ir}(t) - d_p^{ir}(t-1) = \alpha_d c_{dp} \frac{d_p^{ir}(t-1)}{s_p(t-1)}$$

With the initial condition that

$$d_p^{ir}(t) = \left(\frac{t}{t_0}\right)^{\lambda_p} d_p^{ir}(t_0), \tag{1}$$

where $\lambda_p = \frac{\alpha_d c_{dp}}{\alpha_p c_{pd} + \alpha_d c_{dp}}$.

By same approach we can obtain the expression result of $d_d^{ir}(t)$ for nodes in N_d. That is

$$d_d^{ir}(t) = \left(\frac{t}{t_0}\right)^{\lambda_d} d_d^{ir}(t_0),$$

where $\lambda_d = \frac{\alpha_p c_{pd}}{\alpha_p c_{pd} + \alpha_d c_{dp}}$. Thus we complete the proof. ■

Theorem 2: For graph $G(P, D)$ generated after t time slots ($t \geq t_0$), with the condition that inter-degree in node set N_d growing with the power λ_d, the intra-degree of $n \in N_p$ at time t satisfies

$$d_p^{ia}(t) = \Theta\left(t^{\frac{1}{\lambda_d}+1}\right).$$

This result also holds for $n \in N_d$ with symmetrical expressions.

Proof: The intra-degree in N_p is generated by common neighbors in N_d.

When a certain node $d \in N_d$ has node degree x from N_d to N_p, it has exactly x neighbors in N_p. Thus, the expected intra-degree in N_p added by this node is $2\beta_p\binom{x}{2}$, where β_p is the linking probability when two nodes inside node set N_p have a common neighbor node in N_d. And the number of nodes in N_d who have x neighbors in N_p is expected as $|N_d|\mathbb{P}\left\{d_{ap}^{ir}(t) = x\right\}$ 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

$$\text{Contribution}(x) = 2\beta_p\binom{x}{2}|N_d|\mathbb{P}\left\{d_d^{ir}(t) = x\right\}. \tag{2}$$

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 3, we get the intra-degree $d_p^{ia}(t)$ in node set N_p contributed is

$$\begin{aligned} d_p^{ia}(t) &= \sum_{x=1}^{\max} \text{Contribution}(x) \\ &= \sum_{x=1}^{\max} 2\beta_p \binom{x}{2} |N_d| \mathbb{P}\left\{d_d^{ir}(t) = x\right\} \\ &= \Theta\left(\sum_{x=1}^{\max} x^2 x^{-\frac{1}{\lambda_d}-1} t\right) \\ &= \Theta\left(\sum_{x=1}^{t} x^{-\frac{1}{\lambda_d}+1} t\right), \end{aligned}$$

where max presents the maximum inter-degree in E_{pd} which satisfies $\max = \Theta(t)$. By using the sum of *p-series*, we get

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

Therefore, we have $d_p^{ia}(t) = \Theta\left(t^{\frac{1}{\lambda_d}+1}\right)$.

By same approaches, we can also obtain the expression result of d_d^{ia} for nodes in N_d, thus we complete the proof. ■

Theorem 3: For graph $G(P, D)$ generated after t time slots, when $t \to \infty$, the inter-degree sequences of $n \in N_p$ in $B(N_p, N_d)$ follows 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 $n \in N_d$ sharing symmetrical expressions.

Proof: First of all, we consider the distribution of $d_p^{ir}(t)$ 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

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

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

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

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

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

By same approaches, we can also calculate the distribution of $d_d^{ir}(t)$, and thus the proof is complete. ■

Theorem 4: For graph $G(P, D)$ generated after t time slots, when $t \to \infty$, the nodes' intra-degree sequences of $n \in N_p$ follow power-law distribution that

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

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 $n \in N_d$ as they share symmetrical expressions.

Proof: The proof uses the result of *Silvio Lattanzi and D. Sivakumar*'s research work.
citelattanzi2009affiliation. In their work, the model's bipartite network's structure is similar to our model's bipartite networks' which are disconstructed from $G(P, D)$.

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 $n \in N_d$ and thus complete the proof. ■

VI. Conclusion

The conclusion goes here.

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

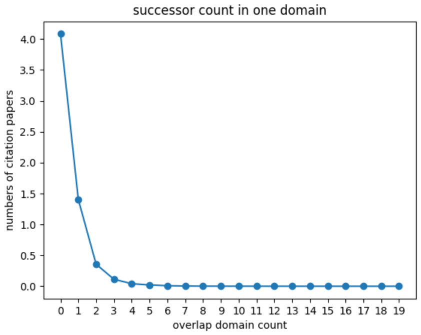


Fig. 13. overlap domain count and the number of cited papers in one domain

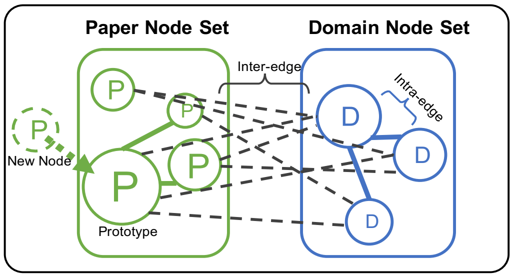


Fig. 14. Structure of evolving scholarly model