Wandering Between Close and Apart: Are Scholarly Domains Crossable?

Lingkun Kong, Bo Wang, Jiaqi Liu, Luoyi Fu, Xinbing Wang
Shanghai Jiao Tong University, China
{klk316980786, wb1996, 13-liujiaqi, yiluofu, xwang8}@sjtu.edu.cn

***Abstract*—Interdisciplinary collaborations, i.e., scholarly cross-domain collaborations have generated huge impact to society, and has been previously proved to exhibit domain skewness [1]. To illustrate, scholarly cross-domain collaborations seldom emerge between irrelevant scholarly domains, e.g., mathematical and political science, astronomy and economics, etc. This viewpoint, however, is fundamentally limited by the paucity of experimental verification as well as the support of evaluation methods, whereby no previous work has proved the skewness with analysis on real-world database of large scale. Given this limitation, in this work we address a question to determine the possible existence of scholarly cross-domain collaborations, namely: "Are scholarly domains really crossable?"**

Using a real-world scholarly dataset, i.e., *Microsoft Academic Graph* (MAG) [2] with 126 million papers collected from 53,834 domains, we take the initiative to formalize a "crossability" quantification problem, where the "crossability" serves as an index that aims to evaluate the ability of two scientific domains to establish collaborations. In doing so, we propose two metrics, i.e., *co-paper ratio* and *hierarchical distance*, where the former one is the ratio of common papers in two domains to which in a single domain, and the later one is the difference of domains' levels according to their positions in the hierarchical structure. Combining two metrics together, we can calculate the value of "crossability". Interestingly, we observe a *peak* pattern, meaning that the influence of research work, i.e., number of citations, climbs to a peak when its domains' count goes to a certain number, after which the citation count decreases sharply. Our discovery indicates that a moderate amount of domain "crossability" helps to improve the impact of research work, which, however, could be weakened under excessive "crossability". With elaborately modeling, we reproduce this *peak* pattern and briefly discuss the reason of the existence of peak.

***Keywords*—Scholarly Data Analysis, Cross-domain Collaborations.**

I. Introduction

Interdisciplinary collaborations, i.e., scholarly cross-domain collaborations play vital role in scientific study, which have generated huge impact to society. For instance, collaborations between biology and computer science create the field of bioinformatics. Thanks to these cross-domain collaborations, originally extremely expensive tasks such as protein localization have turned out to be more scalable and affordable [3]–[5]. Another example belongs to the field of medical informatics, which, with the merge of data mining techniques nowadays, becomes a exponentially fast grown area that is expected to have huge impact on medicine [6] [7].

In literature, scholarly cross-domain collaboration is a theme that runs through large parts of scientific research [8] [9]. And scientists believe that this collaboration has great domain skewness [1], i.e., not all scientific domains are relevant for cross-domain collaborations. To illustrate, a widely accepted point of view is that scholarly cross-domain collaborations are seldom established in irrelevant scholarly domains, i.e., domains wandering apart. For examples, mathematical and political science, astronomy and economics, computer science and archaeology, etc., which have correspondingly far relationship, are unlikely to build collaborations. This viewpoint, however, is fundamentally limited by the paucity of experimental verification as well as the support of evaluation methods, since no previous work has proved the cross-domain collaborations' skewness with analysis on a real-world large-scale dataset. Faced with this limitation, in this work we address a question to determine the boundary between different scholarly domains and the possible existence of cross-domain collaborations, namely: "**Is there really an unbridgeable gap between some scientific domains?**", or in other words, "**Are scholarly domains crossable?**"

Based on the real-world scholarly dataset, i.e., *Microsoft Academic Graph* (MAG) [2] with 126 million papers collected from 53,834 domains, we formalize this problem to quantifying the "crossability" – an index we propose to evaluate the ability of two scientific domains establishing collaborations. Therefore, to calculate the value of "crossability", we propose two metrics, i.e., *co-paper ratio* and *hierarchical distance*, where the former one is the ratio of common papers in two domains to that in a single domain, and the later one is the difference of domains' levels according to their positions in the hierarchical structure.

Apparently, a larger co-paper ratio means better "crossability". However, merely using co-paper ratio cannot well quantify the "crossability" among different domains as it fails to incorporate domains' inherent hierarchical relationship. Therefore, based on scholarly domains' hierarchical positions in MAG, we further propose a method to calculate the hierarchical distance among different domains, which offsets the deficiency of simply evaluating domains' relationship by direct connection, i.e., co-paper ratio. Combining these two metrics together, we can thus properly quantify the "crossability" between different scholarly domains. While we defer to Section IV for more details of how to combine these two metrics, we would like to briefly disclose that the combined metrics more comprehensively evaluate the "crossibility", suggesting that the "crossability" between scholarly domains varies significantly, i.e., **scholarly domains are wandering between**

close and apart with each other.

As the impact of a research work is commonly assessed by its citation count, when exploring different works' citation numbers over their respective cross-domain performance, i.e., the number of their related domains, we surprisingly find a peak pattern, meaning that research work's influence climbs to a peak when its related domains' number goes to a certain value, while decreases when domains' number exceeds this certain value. Our discovery implies that moderate "crossability" between domains can potentiallly enhance a research work's impact, which, nevertheless, can be weakened under excessive "crossability", as opposed to a common conception that scholarly cross-domain collaborations can always generate work of broad impact [9]–[12]. Additionally, by the assist of several desirable properties of Gaussian function, we reproduce this *peak* pattern with elaborately modeling – presenting mathematical formula and learning algorithm. Our work concludes with the discussion on the underlying reason of why this *peak* pattern exists, followed by performing the prediction of research works' impact through our proposed Gaussian-like model.

To sum up, in this paper, we innovatively explore real-world datasets of scholarly cross-domain collaborations. Our contributions can be mainly divided into three parts:

- We make the first attempt to perform experiments on the large-scale real-world dataset to explore properties of interdisciplinary collaborations.
- We define the "crossability" between scholarly domains as index to evaluate the possibility of domains' collaborations, and quantify it by two proposed metrics, i.e., co-domain ratio and hierarchical distance.
- We discover the *peak* pattern in the relationship of research works' citation number and their related domains' count. Moreover, we reproduce this pattern with theoretical modeling and explain the cause of this pattern.

This paper is organized as follows. In Section II, we discuss related literatures. In Section III, we introduce the dataset we use and list several basic features of scholarly cross-domain collaborations. We give the quantification method of domains' "crossability" in Section IV and analyze the *peak* pattern in Section V. Finally, we draw conclusions in Section VI.

II. Related Work

To the best of our knowledge, there are no prior works, other than ours, that have explored whether two scholarly domains are crossable or not. However, there are indeed several related works regarding investigations of scholarly cross-domain collaborations.

In literature, scholarly cross-domain collaboration is a theme that runs through a large body of scientific research [8], [9], [13]. From the aspect of implementing valuable information from data in collaboration, Tang et al. [1] study the problem of cross-domain collaboration recommendation and propose a Cross-domain Topic Learning (CTL) model to learn and differentiate collaboration topics from other topics. Isenberg et al. [14] present the results of a comprehensive multi-pass analysis of visualization paper keywords supplied by IEEE Visualization conference series. By their visualization of relationship between papers' keywords, the collaborations of different scientific fields of study has been revealed. Aggrawal et al. [15] and McCarty et al. [16] explore collaborations among different authors through analysis of the co-authorship network by leveraging several statistical measures. The work provides useful guidelines for a better understanding of the collaborations between different communities. By analyzing the citation number of papers, Portenoy et al. [17] and Dawson et al. [18] evaluate the current impact of the scientific fields, which to some extent reveals the activity of collaborations between scholarly domains.

In the present work, we observe power-law distribution in scholarly cross-domain collaborations, which has essentially been intensively studied in classical social science theories [19]–[21]. Also, this paper discusses the *peak* pattern of how research works' cross-domain performance influences their academic impact, which has been previously investigated by Zhu et al. [22], [23] through some other different approaches. However, they have not taken advantage of the interdisciplinary collaborations' information, i.e., papers' cross-domain performance to study works' impact. Therefore, we explore works' influential performance with the concern of scholarly cross-domain collaborations, and propose a model to simulate the *peak* pattern, which is trained by methods of gradient descent optimization [24], [25].

III. Basic Discussion On Database

In this part, we introduce the database that we use by visualizing data structure. Also, by leveraging data-mining methodologies, we deliver basic discussions on some classical properties of the cross-domain collaboration, which have been well studied in social science while not been validated in real-world scholarly database.

A. Statistical Information of MAG

We perform 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. Approximately 126 million papers in 53,834 domains are included in this database, with the publication year of each paper varying from 1800 to 2016.

As we study the cross-domain collaborations of scholarly data, we mainly launch research on the scholarly data related with scholarly domains embodied in MAG. And the domains in MAG can be divided into four layers called as L0 domain, L1 domain, L2 domain and L3 domain in a hierarchical pattern where domain with smaller layer number represents larger scientific domain and can be further divided into many smaller sub-domains labeled with larger layer number. For instance, according to the MAG's hierarchy table, we can obtain a domain labeled with L0 (Layer 0) in MAG called "Computer Science", which contains several domains labeled

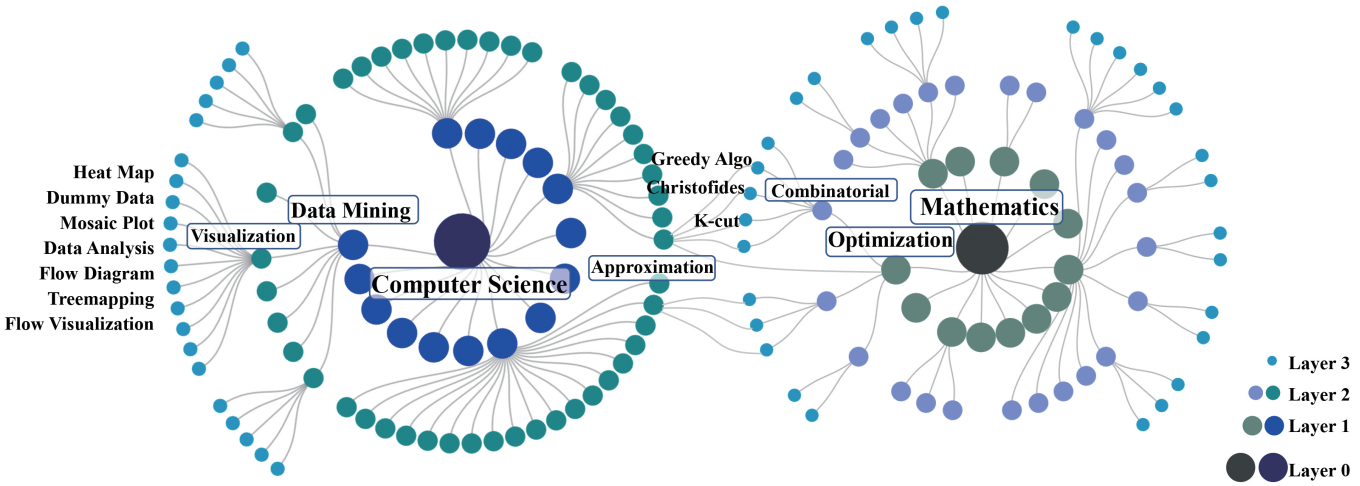


Fig. 1. This figure reveals hierarchy structure in MAG. Here we use domains in "Computer Science" and "Mathematics" as instances to explain the domain's hierarchy in MAG. In this figure, we use bigger nodes to present domains with higher level of hierarchy, which could contain domains with smaller size. For example, L0 layer domain – "Computer Science" contains "Data Mining" as its L1 sub-domains. And "Data Mining" also contains several other lower layer domains in MAG.

by L1, including "Artificial Intelligence, "Database", "Data Mining", "Computer Hardware", and etc. And "Data Mining" can contain domains with higher layer number such as "Conceptual Clustering", "Knowledge Extraction" in L2 Layer and "K-optimal pattern discovery", "XML Schema" in L3 layer.

Figure 1 is a hierarchy example of MAG from which we choose "Computer Science" and "Mathematics" as two particular domains for illustration. To present a clear overview of hierarchical structure of domains, from "Computer Science" and "Mathematics", we only pick, respectively, several specific sub-domains that are marked by their names. Although such selective choices lead to a far sparser domain distributions than which in the complete database, we note that it well preserves the overall hierarchy structure of real scholarly domains.

B. Rediscovery of the Power-law

Power-law distribution is a common feature in social science, which is also well studied by many existing literatures [19]–[21]. However, due to technical difficulties, there have been few studies that provide an experimental exploration and validation of power-law distribution in scholarly data of cross-domain collaborations, or even scholarly data at scale.

Motivated by this, we bridge this gap by studying real-world scholarly data of interdisciplinary research collaborations. And we validate the classical power-law distribution in our study, which pads the vacancy of the study about power-law distribution in scholarly collaborations. In our study, we formalize the abstract concept of scholarly cross-domain collaborations to the network which combines two fundamental elements in scholarly collaborations, i.e., research papers and scientific domains. In fact, in academic research, a paper serves as the bridge of cross-domain collaborations, as papers that simultaneously cross multiple different scientific domains, represent the blending of multi-domains' knowledge and the achievements of scholarly cross-domain collaborations.

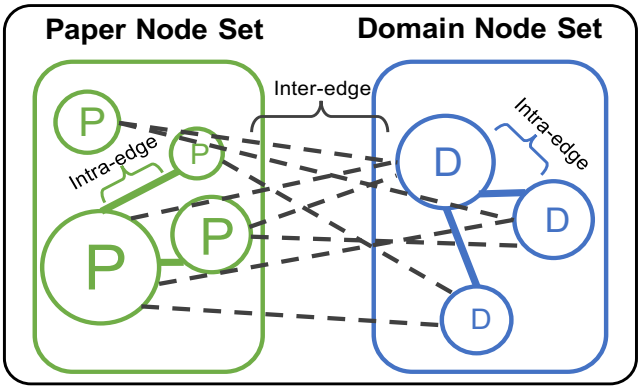


Fig. 2. Network structure of scholarly cross-domain collaborations. This network contains paper node set and domain node set, where the inter-edges in the network present the correlation between papers and domains – paper's content being related with several domains. And the intra-edges in paper node set reflect the citation among papers, while the intra-edges in domain set represent the hierarchy among domains, i.e., one domain might be connected to several its sub-domains in database.

Here, we regard these two elements, i.e., research papers and scientific domains as two node sets in the network of scholarly collaborations. While the inter-edges between these two node sets denote the collaborating relationship between papers and domains. Also, inside each node sets, there exist intra-edges, which in paper node set reflect the citation among papers, in domain node set present the domains' hierarchical relationship. Figure 2 helps to illustrate the construction of this network of scholarly collaborations.

Based on our network construction and analysis result on real-world scholarly data, we prove the existence of power-law distributions of both inter and intra nodal degrees in scholarly cross-domain collaborations. Figure 3 represents four different types of power-law distributed degree by our empirical results. The first one is a power-law distributed domains' number with papers' count in these domains, as shown in the Figure

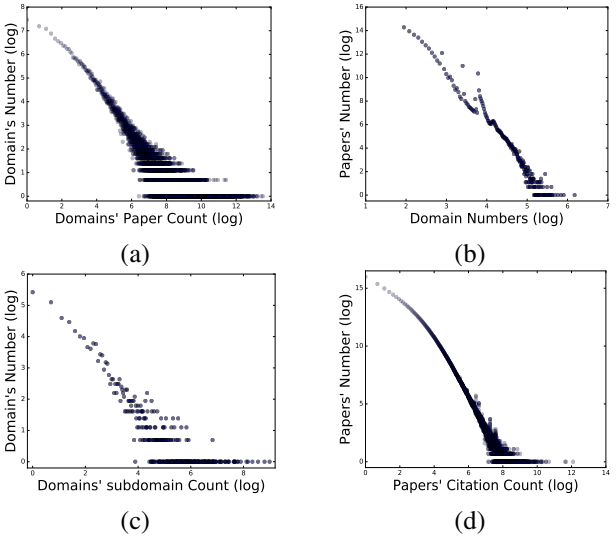


Fig. 3. Illustration of power-law distributions in scholarly cross-domain collaboration from four aspects. Figure 3(a) and 3(b) show the two kinds of power-law distributed inter-degree, i.e., nodes' degree in two node sets, contributed by edges cross two domains. Particularly, Figure 3(a) reflects the power-law distribution in domain set, while Figure 3(b) shows the power-law distribution in paper set. Figure 3(c) and 3(d), however, present the other two kinds of power-law distributed intra-degree, i.e., nodes' degree in two node sets, contributed by intra-edges. And Figure 3(c) portrays power-law distribution inside domains' hierarchy structure while Figure 3(d) illustrates the power-law in paper's citation relationship.

3(a). The power-law distributed papers' number vs. papers' domain count is plotted in Figure 3(b). And Figure 3(c) represents power-law distributed domains number with their sub-domains' count, while Figure 3(d) reflects the power-law distribed papers' number with its citation count.

All those observations confirm, from different perspectives, the existence of classical power-law distribution in scholarly cross-domain collaborations.

IV. The Quantification of "Crossability"

In this section, faced with the question that "Are scholarly domains crossable?", we try to quantify the "crossability" between different scholarly domains by measurement of jointly using *co-paper ratio* and domains' *hierarchical distance*.

A. The Co-paper Ratio between Domains

First of all, as we illustrated earlier, papers in real-world database are always related with several domains. This discovery indicates that between two domains, there might exist many co-papers, which are the bridges of scholarly cross-domain collaborations, as illustrated in Figure 4. And the co-paper ratio can be computed as one important criteria of domains' "crossability". For a clearer explanation, let us consider the "crossability" from "Data Analysis" to "Data Modeling" as an example. In this scenario, we can obtain the co-paper ratio through dividing the co-paper number between "Data Analysis" and "Data Modeling" by the total paper number from "Data Analysis". Therefore, we can obtain the co-paper ratio as a fundamental evaluating metric for quantifying the "crossability" between different scholarly domains.

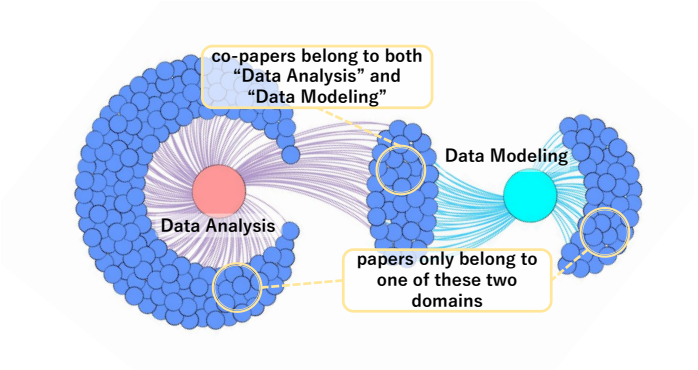


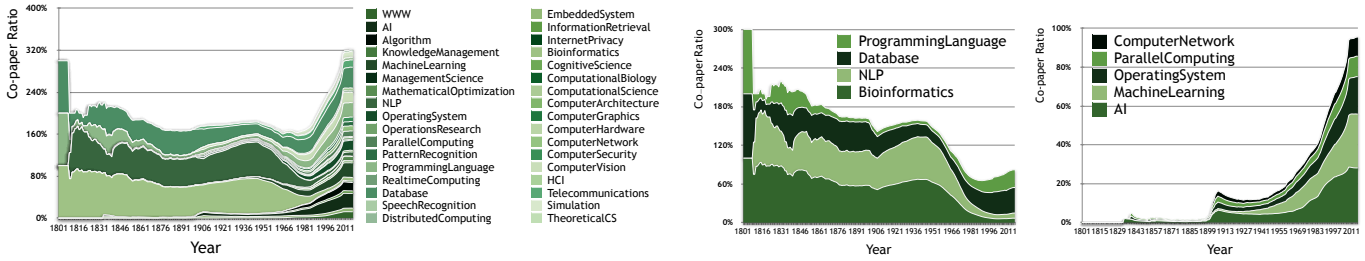
Fig. 4. Co-papers between two scholarly domains. Here these two are both L3 layer domains, called "Data Analysis" and "Data Modeling".

Further, when observing the time information from the data of paper publication date in MAG, we find that the co-paper ratio evolves, i.e., fluctuates with the time. Here, we generate a case study on co-paper ratio between "Data Mining" and other L1 sub-domains in "Computer Science". To illustrate, we label every paper from "Data Mining" and other sub-domains in "Computer Science" by their publication date. Thus, all co-papers between "Data Mining" and other domains are labeled by time information. Then, we calculate the co-paper ratio for every time slot and get the visualizing results in Figure 5.

In Figure 5, we can view the evolving co-paper ratio between "Data Mining" and other different L1 domains, such as "Artificial Intelligence", "Algorithm", "Operating System", and etc. By this figure, we can see the evolving collaboration strength between "Data mining" and other L1 domains in "Computer Science". Interpreted more clearly, in the early stage, "Data Mining" has high co-paper ratio with "Bio-information" – almost 90% which indicates high collaboration strength among these domains. While in recent years, the co-paper ratio with "Bio-information" has been decreased to a very low level, which indicates the weaker intimacy between "Data Mining" and "Bio-information". Moreover, it can be viewed that recently "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, which, to some extent, reveals the rapid development of these two subjects.

B. Hierarchical Distance Between Domains

Co-papers between scientific domains can directly present collaborations in different scholarly domains. However, these collaborations cannot fully reflect the "crossability" as domains' inherent hierarchical relationship is neglected. Take Figure 1 as example, "Flow Diagram" and "Heat Map" are both L3 sub-domains which belong to the L2 domain of "Visualization", meaning that they have really close hierarchical relationship. And according to database, the co-paper ratio between these two domains is fairly small, since basically they are two different visualizing tools. However, when academic researchers are trying to implement visualizing tools, they are likely to further dive into the sub-domains of "Visualization" domain. And these two domains might be employed as presentation methods by researchers. In fact, in this paper, we



(a) Overall Evolving Co-paper Ratio (b) Major Decreasing Ratio (c) Major Increasing Ratio

Fig. 5. This figure reflects the evolving pattern of co-paper ratio between the domain of "Data Mining" and other L1 sub-domains in the L0 domain – "Computer Science". Here, we use stack flow graph to present the ratio's evolving value, i.e., the wide flow means the large co-paper ratio between two domains. And flows' values are not summed up to 100% since the overlapping of papers' domains, i.e., one paper could have multiple different domains. Figure 5(a) shows all co-paper ratios between "Data Mining" and other subdomains in "Computer Science". And Figure 5(b) presents major decreasing ratios between "Data Mining" and four L1 domains – "Programming Language", "Database", "Natural Language Processing" and "Bio-informatics". While Figure 5(c) shows major increasing ratios from five domains – "Computer Network", "Parallel Computing", "Operating System", "Machine Learning" and "Artificial Intelligence".

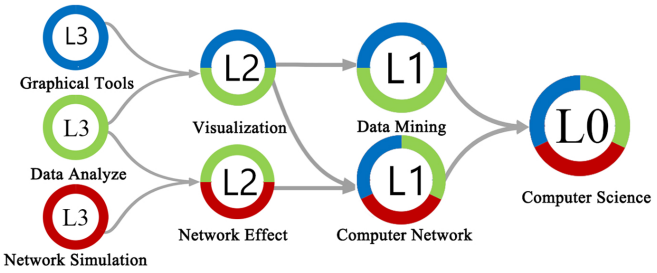


Fig. 6. This figure presents an instance of how the proposed Algorithm 1 of calculating hierarchical distance between scholarly domains works. Here it selects several scholarly domains from *MAG* which are marked by circles. And the color of the circle indicates the hierarchical relationship between domains – circles possessing same color are hierarchically correlated in real-world database.

also use both two of these graphs. Hence, the collaboration between these two domains is founded, meaning that the two domains are actually crossable. And if we just depend on co-paper ratio to quantify the "crossibility" between "Flow Diagram" and "Heat Map", the result will nearly be equal to 0, which betrays the fact that the collaboration is likely to happen between these two domains. In this case, only using co-paper ratio to determine the "crossability" is not comprehensive nor accurate enough. Therefore, we further introduce the conception of hierarchical distance between scholarly domains as an additional important feature for quantifying the "crossability" between scholarly domains. And we propose an algorithm to calculate this hierarchical distance between different scholarly domains, as presented in Algorithm 1. The major notations that are used in Algorithm 1 are listed in Table I.

TABLE I
NOTATIONS ADOPTED IN ALGORITHM1

Notations	Definations
D_a, D_b	The scholarly domain 'a' and 'b'
$\text{Que}_a, \text{Que}_b$	Queue structures for data from D_a, D_b
$\text{Ancs}(D_a)$	All ancestors of D_a
$\text{Layer}(D_a)$	The layer of D_a in database $\in \{0, 1, 2, 3\}$
$\text{Dist}(D_a, D_b)$	The hierarchical distance between D_a and D_b

To gain a better understanding of Algorithm 1, Figure 6 provides an illustration of hierarchical distance. We pick up three L3 scholarly domains – "Graphical Tools", "Data Analyze", "Network Simulation" as well part of their ancestors, i.e., "Visualization", "Network Effect", "Data Mining" and etc.

Algorithm 1 Calculating Hierarchical Distance

Input: Two scholarly domains, namely D_a and D_b.
Output: The hierarchical distance between domain 'a' and domain 'b', i.e., the final $\text{Dist}(D_a, D_b)$.

Add D_a and D_b into Que_a and Que_b respectly.
while Que_a not empty **do**
  Add the accestors of Que_a's head into Que_a.
  Pop out the head of Que_a and add it into $\text{Ancs}(D_a)$.
end while
while Que_b not empty **do**
  Add the accestors of Que_b's head into Que_b.
  Pop out the head of Que_b and add it into $\text{Ancs}(D_b)$.
end while
$\text{Dist}(D_a, D_b) \leftarrow 3$ // Initialisation
for d_a in $\text{Ancs}(D_a)$ **do**
  for d_b in $\text{Ancs}(D_b)$ **do**
    if $d_a = d_b$ and $3 - \text{Layer}(d_a) < \text{Dist}(D_a, D_b)$ **then**
      $\text{Dist}(D_a, D_b) \leftarrow 3 - \text{Layer}(d_a)$
    end if
  end for
end for
return $\text{Dist}(D_a, D_b)$

In Figure 6, circles, possessing the same color indicates that they are hierarchically correlated domains in real-world database, e.g. the L1 domain "Computer Network" and L3 domain "Data Analyze" exhibit hierarchical relationship as they both share the red color in their borders.

As can be revealed by Figure 6, "Graphical Tools" and "Network Simulation" share two common hierarchical ancestors, i.e., ancestors in Algorithm 1 – "Computer Network" and "Computer Science". In our algorithm, we choose the ancestor domain at the lowest layer to calculate the final hierarchical distance. And in this case, it is "Computer Network" at L1 layer. According to our algorithm, we use 3 to minus the layer number of lowest ancestor – here "$3 - 1$", where 1 denotes the layer number of "Computer Network". Therefore, the hierarchical distance between "Graphical Tools" and "Network Simulation" is calculated as 2. The same approaches can also be applied to quantify the hierarchical distance between different scholarly domains.

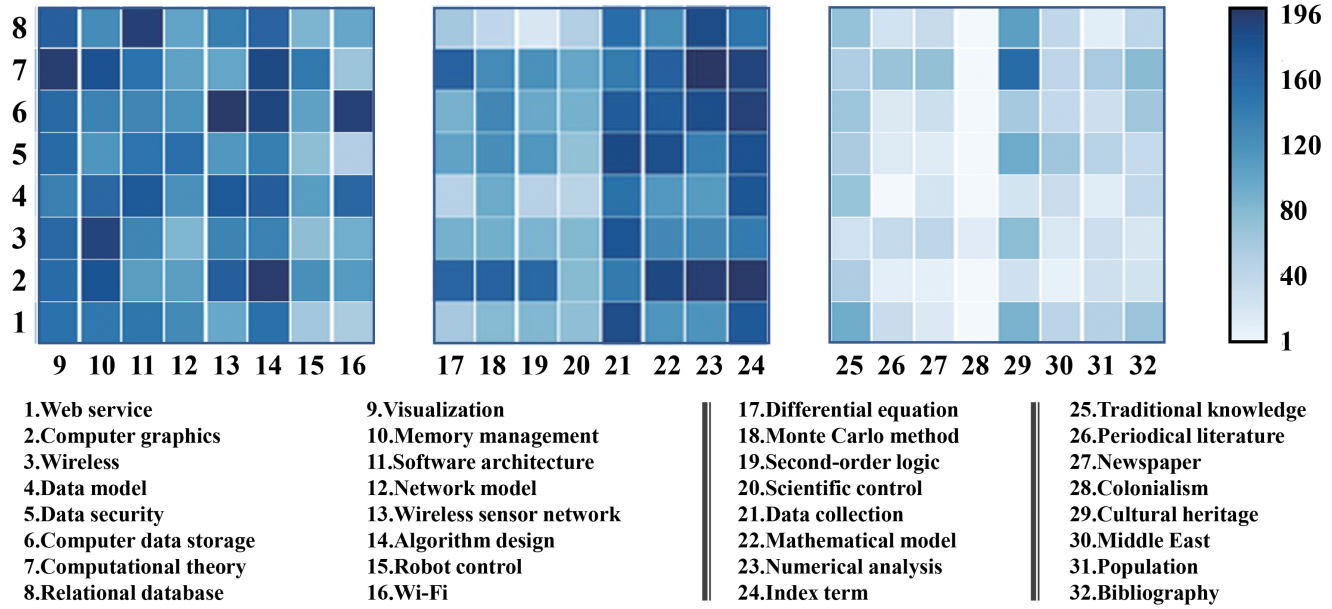


Fig. 7. This figure shows the quantified "crossability" between different scholarly domains, where domains from No.1 to No.16 are sub-domains of L0 domain "Computer Science", domains from No.17 to No.24 are sub-domains of L0 domain "Mathematics", while domains from N0.25 to No.32 are sub-domains from L0 domain "Art". Due to the huge gap in magnitude among values of "crossabilities" we calculate, we range the heat bar from 1 to 196, which represents the mapped ranking of values about "crossabilities" between domains, where 1 means the smallest value while 196 means the largest.

C. *Quantification Results*

As we have already obtained co-paper ratio between scholarly domains and domains' inherent hierarchical distance according to previous discussion, combining them together, we can quantify the "crossability" between different domains. However, since the question – "Are scholarly domains crossable?" receives no prior exploration and the "crossability" is an index property first being proposed, there is no ground truth of the "crossability" between different scholarly domains. And this makes evaluating the accuracy of our quantifying result even impossible.

Therefore, in this paper, we regard our quantification method as a reasonable possible solution to the open question. And here we use the equation below to quantify the "crossability" between different scholarly domains:

$$\text{Score} = \frac{R}{D+1},$$

where Score denotes the final quantified value of "crossability", R denotes the paper ratio between scholarly domains, D means the domains' hierarchical distance.

By this equation, we visualize the "crossability" between different L1 domains from "Computer Science" in Figure 7. As can be seen from the figure, the "crossability" intra sub-domains of "Computer Science" is generally larger than "crossability" between sub-domains from "Computer Science" and "Mathematics", which means the collaborations can be more easily established between sub-domains from "Computer Science". In contrast, the "crossability" between sub-domains of "Computer Science" and those of "Art" are significantly much smaller.

V. The Discovery And Analysis of *Peak* Pattern

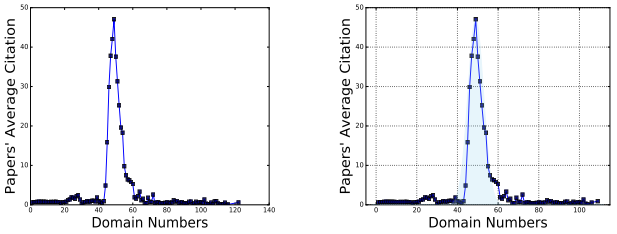
As we have argued in previous section, research works serve as the bridge in scholarly cross-domain collaborations. And collaboration's "crossability", from these bridges' perspective, can serve as research works' cross-domain performance, since if domains share better "crossability", the papers in these domains will be more likely to be related with multiple domains. In this section, we explore research works' citation number, i.e., the impact of collaborations' products over its cross-domain performance, i.e., related domain number. And surprisingly, we discover a *peak* pattern, meaning that research work's influence climbs to a peak when its domains' count being accumulated to a certain number, and then declines sharply hereafter. This indicates that the better performance of domains' "crossability" not always leads to the improvement of research works' quality. Further, taking advantage of properties of Gaussian function, which desirably meet some regularities of *peak* pattern, we generate modeling on the *peak* pattern based Gaussian function. And we validate our model by experiments. Also, we estimate one research work's reputation based on its cross-domain performance.

A. *Observation of* Peak *Pattern*

Intuitively, a research work which combines multiple disciplines' knowledge is supposed to own high citations. This is because the combination of different scientific domains' theories and methodologies can well produce desirable novelty. Also, the impact of works with multiple domains can be well broadcasted among researchers since every certain domain captures attention of researchers who launch studies in this certain field. Therefore, it seems to be reasonable that

the research work's impact and quality will increase with the growing number of domains it belongs to.

However, through our investigation into real-world scholarly data in MAG, surprisingly we get a *peak* pattern which contradicts our assumption, i.e., the work's influence climbs to a peak and drops from the peak with the increase of domains' number as opposed to our intuition that it will grow ceaselessly. For details, we explore the papers', i.e., research works' influential performance with the concern of scholarly cross-domain collaborations. Based on paper's membership relation with the domains it belongs to, we study paper's citation number which we use to evaluate the paper's influence. And surprisingly we find a peak in data, i.e., the papers' average citation number is likely to reach a maximum value when papers' domain number comes to a certain amount and drop sharply when domain number exceeds this amount.



(a) Observation (b) Simulation

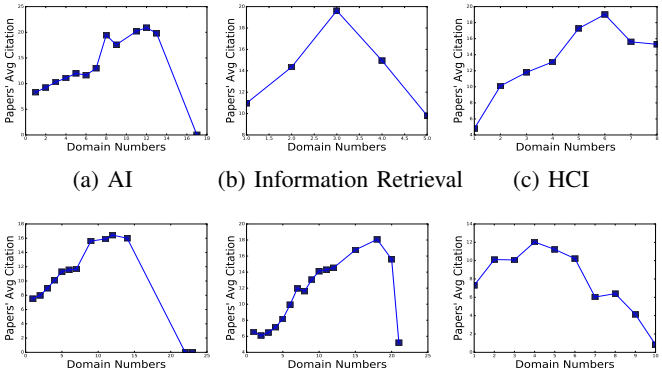
Fig. 8. This figure shows papers' average citation count with certain domain numbers of all papers in *MAG*. Figure 8(a) portrays real-world scholarly data, and Figure 8(b) contains simulation result of our model.

Figures 8 and 9 present this pattern clearly. To illustrate, Figure 8 counts the overall papers' average citation number over their related domains' number. And we can clearly find a peak when paper's domain count goes to around 50. As a matter of fact, here we set a threshold of these papers' number when calculating average citation in case it is too small to represent a pattern. And we find when paper's domain count is less than 120, the paper's number is large enough to support such a pattern. In other words, this *peak* pattern does exist. Moreover, when exploring smaller domains, we find that the *peak* pattern is uniformly valid. Figure 9 selects several L1 sub-domains from the L0 domain of "Computer Science" as instances, and shows that the peak still exists, though the range of paper's domain count is confined.

The intuitive explanation of *peak* pattern is that the small range of research works' related domains constrains their impact in small area, which brings indistinctive influence, i.e., low citations, while too many domains might distract or diffuse author's attention and thus harm the paper's profundity, which decreases the citation number of research works' with more related domains.

B. The Modeling of Peak *Pattern*

In order to predict the papers' citation, i.e., research works' impact, we use the Gaussian function to build our model due to several desirable properties of Gaussian function.



(a) AI (b) Information Retrieval (c) HCI

(d) Operating System (e) Tele-communications (f) WWW

Fig. 9. This figure shows papers' average citation count with their related domains' numbers in selected L1 sub-domains of "Computer Science" in MAG, including "Artificial Intelligence", "Information Retrieval", "Human-Computer Interaction", "Operating System", "Tele-communications" and "World Wide Web". In this figure, the points as well the blue line represent real-world data in MAG. And these domains are divided into test set for model learning, whose simulating, i.e., predicting results will be discussed later.

In mathematics, a Gaussian function, often simply refers to as a Gaussian, is a function of:

$$g(x) = c \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{1}$$

where c denotes a constant.

And the normalized Gaussian function is constructed as:

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

which is plotted in the Figure 10.

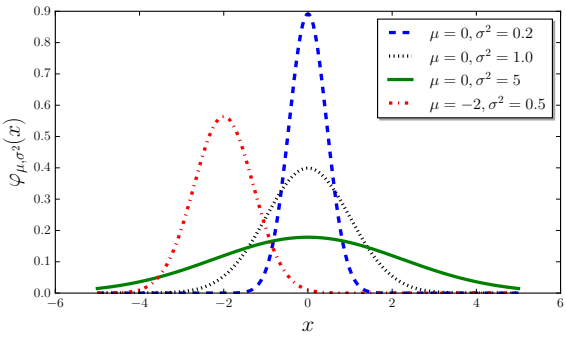


Fig. 10. Normalized Gaussian curves with expected value μ and variance σ^2.

In Figure 10, a peak is presented by Gaussian function, where the Gaussian curve climbs to the maximum when $x = \mu$. And, the curve becomes smoother with the growth of σ^2, while in contrast, the peak of Gaussian curve is steeper when σ^2 being smaller. Similarly, in the study of *peak* pattern in real-world scholarly database, we find that when the x axis has wider range, i.e., the papers' domain number varies in a wider range, the peak becomes much steeper, while in contrast, when x axis has narrower range, i.e., the domain number is limited in a short scope, the peak becomes much gentler. And by comparing the peak's steepness with domain number's range in Figures 8 and 9, this property is clearly viewed.

Therefore, we can leverage this similarity between *peak* pattern of cross-domain collaborations and the peak-like distribution in Gaussian function to fit the real-world data. To illustrate, we are supposed to replace μ and σ in classical Gaussian model by parameters with practical meanings. Hence, we model the *peak* pattern by mirroring the Gaussian function and get series of equations to reproduce the peak. In fact, our model uses the data of peak point, i.e., the most influential research work (with highest citation) as input to simulate the other research works citation distribution in a certain scientific domain. The equation is shown as below:

$$p_\theta(x) = M \cdot e^{-\frac{\left(x - \arg\limits_{x\in\mathbb{R}} M\right)^2}{f_\theta(\mathbb{R})}}, \tag{2}$$

where x denotes the paper's domain number, $p_\theta(x)$ represents the expected papers' average citation count with input x, M is a constant which equals to the maximum citation number for papers in a certain domain, and $f_\theta(\mathbb{R})$ is a linear function with the range $\mathbb{R}$ of paper's domain number as input and θ as weights. And the equation for $g(\mathbb{R})$ is:

$$\begin{aligned} f_\theta(\mathbb{R}) &= 2 \cdot \left(\frac{1}{h_\theta(|\mathbb{R}|)}\right)^2 \\ &= 2 \cdot \left(\frac{1}{h_\theta\left(\max\limits_{x\in\mathbb{R}} x - \min\limits_{x\in\mathbb{R}} x\right)}\right)^2, \end{aligned} \tag{3}$$

where the hypothesis $h_\theta(t)$ is given by the linear model:

$$h_\theta(t) = \theta^T t = \theta_0 + \theta_1 t.$$

And here t equals to $\max\limits_{x\in\mathbb{R}} x - \min\limits_{x\in\mathbb{R}} x$.

Combining equations (2) and (3) together, our model is presented by:

$$p_\theta(x) = M \cdot e^{-\frac{1}{2}\cdot\left[\left(x - \arg\limits_{x\in\mathbb{R}} M\right) h_\theta\left(\max\limits_{x\in\mathbb{R}} x - \min\limits_{x\in\mathbb{R}} x\right)\right]^2}. \tag{4}$$

Comparing our modeling Eqn. (4) for *peak* pattern with the classical Gaussian function which is presented by Eqn. (1), our model utilizes $\arg\limits_{x\in\mathbb{R}} M$ to replace the classical μ in Gaussian function and uses $\frac{1}{h_\theta\left(\max\limits_{x\in\mathbb{R}} x - \min\limits_{x\in\mathbb{R}} x\right)}$ to replace the classical σ. And by this construction, in our model, we can find that the average papers' citation number approaches to peak exactly like the real-world data. And the peak's steepness in model is controlled by the range of papers' domain number.

By using Eqn. (4), we plot our manually simulating graph in Figure 8, in which the simulating peak is able to well fit the peak in real-world scholarly data.

C. Experiments And Simulating Results

Here, we train our model by gradient descent optimization. First of all, we give our cost function of our model as below:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(p_\theta(x^{(i)}) - y^{(i)}\right)^2 + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2,$$

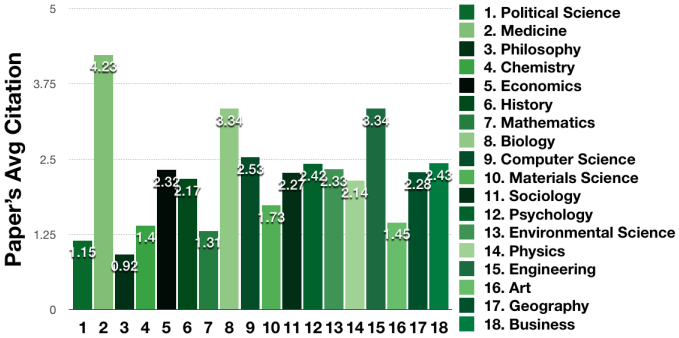


Fig. 11. This figure calculates research works' average citation number in every L0 domains from MAG. And for different scholarly L0 domains, the average citation number of research works varies significantly.

where m denotes the number of training samples, function p_θ is given by Eqn. (4), $x^{(i)}$ is the value of input in the i^{th} training sample while $y^{(i)}$ is the real-world labelled output value for the i^{th} training example, λ represents the regularization parameter which prevents overfitting problem, n denotes the number of features we try to learn and θ_j means the weight of feature j.

Based on our cost function, we compute the gradient descent function for our model:

- If $j = 0$, we have:

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{A}{m}\sum_{i=1}^{m} p_\theta(x^{(i)}) h_\theta(A) \cdot (B - x^{(i)}),$$

  where $A = \max\limits_{x\in\mathbb{R}} x - \min\limits_{x\in\mathbb{R}} x$, $B = \arg\limits_{x\in\mathbb{R}} M$.
- If $j > 0$, we have:

$$\frac{\partial J(\theta)}{\partial \theta_j} = \left(\frac{A}{m}\sum_{i=1}^{m} p_\theta(x^{(i)}) h_\theta(A) \cdot (B - x^{(i)})\right) + \frac{\lambda}{m}\theta_j^2,$$

  where A and B stay the same when $j = 0$.

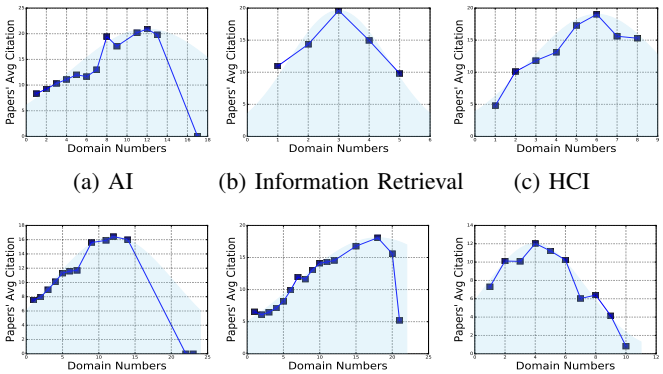
Therefore, we can fix parameters in weight vector θ by gradient descent optimization. And during the learning process, the learning rate is set to be α, i.e.,

$$\theta := \theta - \alpha \cdot \frac{\partial J(\theta)}{\partial \theta},$$

where θ denotes the weight vector including all values of θ_j, in which $j \in \{1, 2, ..., n\}$.

However, in different academic communities, the research works' citation count can be far cry from each other, due to the diverse activity in different scholarly domains, different popularity of digital library which facilitates easy citation, as well the disparate citing habits. And in Figure 11, it can be easily found that in our database, there exists significant difference between research works' average citation numbers in different scholarly domains. For instance, papers in the domain of "Medicine" have 4.23 citations on average while papers in "Philosophy" only have an average of 0.92 citations.

In this situation, it is inappropriate to model all scholarly data in different domains by same parameters, i.e., θ in equation 4. Therefore, in experiment, we separate scholarly data into several parts, i.e., for every L0 domain in MAG,



(a) AI (b) Information Retrieval (c) HCI

(d) Operating System (e) Tele-communications (f) WWW

Fig. 12. This figure shows our model's predicting results of papers' average citation count with their related domains' number in selected L1 sub-domains of "Computer Science" in MAG, including "Artificial Intelligence", "Information Retrieval", "Human-Computer Interaction", "Operating System", "Telecommunications" and "World Wide Web". In this figure, points represent real-world data, while the Gaussian function shaped curves, with their pale blue background, show the peak which is predicted by our model.

we train parameters for *peak* pattern. More specifically, for eight selected L0 domains in database, we use their 60 % sub-domains as training set, 20% as validation set and 20% as test set. And when tuning parameters in validation set, we typically choose learning rate α from 0.0001, 0.001, 0.01, 0.1, 1 and choose the regularization parameter λ from 0, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10 to see which pair of them can get the best learning performance. The final testing result is listed in table II, where domain name represents the selected L0 domain's name, α and λ reflect the best learning rate and regularization parameter for every domain, final cost is the final score of our cost function in validation set, and the $[\theta_1, \theta_0]$ is the final fixed parameter vector in our model for every different scholarly domain.

TABLE II
RESULTS OF GRADIENT DESCENT OPTIMIZATION

Domain Name	α	λ	Final Cost	$[\theta_1, \theta_0]$
Computer Science	0.001	0.3	7.345	[4.34, 13.23]
Chemistry	0.001	0.1	9.978	[5.17, 16.19]
Biology	0.01	0.03	7.672	[6.34, -1.58]
Physics	0.001	0.03	10.198	[7.34, 4.72]
Mathematics	0.001	0.1	9.910	[4.34, 2.20]
Materials Science	0.001	0.3	12.345	[17.34, -1.92]
Art	1	3	17.253	[16.22, -12.91]
Business	0.001	0.03	8.021	[7.17, 5.03]

Therefore, the parameters in our model for different kinds of scholarly domains are properly trained. And then, we employ these parameters to generate simulation results, which will be compared with the result from real-world data in test set. In other words, in next step, research works' citation number will be predicted by trained model. And we will compare these predicting results with real-world data in testing set to evidence the accuracy as well usefulness of our *peak* pattern modeling.

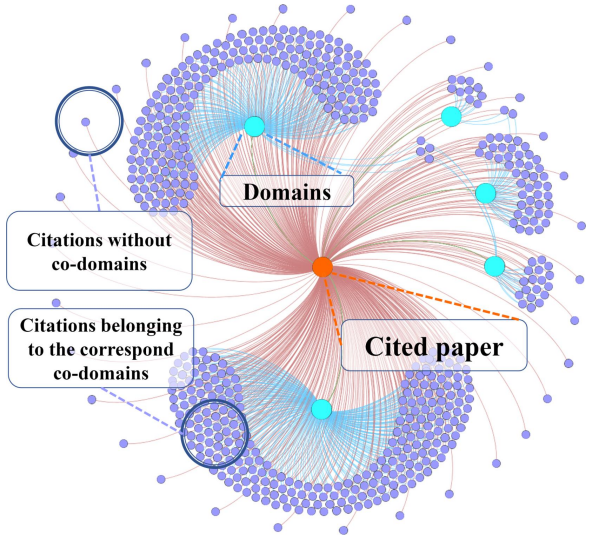


Fig. 13. This figure presents research works' cross-domain performance. The red node in the center represents a specific paper we study. The blue nodes connected to the central node denote the domains related to paper we study. The purple nodes mean papers which the paper in the middle. And they also might link to domains in blue color, indicating those papers make citation by co-domains with the cited paper. In fact, there are citations having more than one co-domains with the cited paper, and there also exist citations which do not connect any domain related to the cited paper, i.e., the purple nodes scatter around the corner.

D. Prediction of Paper Citation

Based on the model discussed in this section, the research works' influence, i.e., citation number can, to some extent, be forecasted by our model. To illustrate, after knowing the most influential paper, i.e., paper with most citations in a certain scholarly domain, as well its domain count, other papers' citation number can be predicted once we know how many domains it related to in database.

And the Gaussian-like curve in Figure 12 presents our predicting results of several L1 domains in "Computer Science" from testing set, which have been mentioned in Figure 9. It can be viewed that the curves are able to well fit the real-world scholarly data, i.e., the points connected by blue lines in Figure 12.

E. The Insight of Peak Pattern

According to the above subsections, we have observed the *peak* pattern in the relationship between research work's impact and its cross-domain performance. And we also give a naïve explanation for the *peak* pattern that the small number of related domains constrains a paper's impact in a small area, which brings few citations, while too many domains might distract authors from focusing on specific problems and harm the paper's quality as well impact.

Here, however, we dig into the scholarly data about research works' cross-domain performance and try to explore the insight of *peak* pattern. We visualize these data as Figure 13 presents. By this figure, we learn more detailed knowledge of how research works' cross-domain performance impacts their citations. In Figure 13, some related domains of the cited paper bring high citations, while others bring low citations. When studying the pattern of which related domains bring

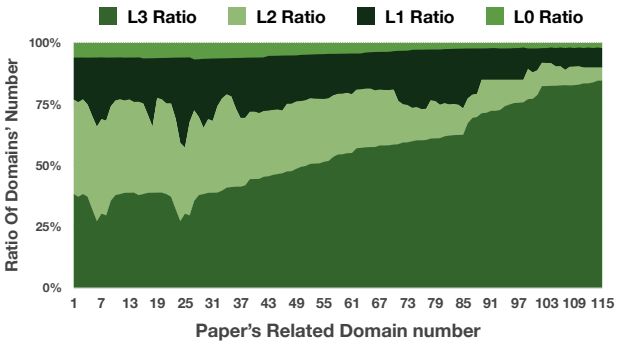


Fig. 14. This figure shows the number ratio of domains from different layer in database. For those papers related with more than 100 domains, more than 75 % of their domains are from L3 layer, which in fact affords small citation number.

higher citations, we find that the co-domain from layer with smaller number in database brings better impact, i.e., the L0 co-domains are more likely to contribute the higher citations than L1, L2 and L3 co-domains. However, in *peak* pattern, with the growth of paper's related domains number, the ratio of domains with higher layer number also increases rapidly, as shown in Figure 14. This indicates though the research work seems to be related with many domains, there only few of these domains could bring citations for the research work. And that is reason why research work's citation cannot increase ceaseless with the growth of its related domain number.

VI. Conclusion and Future Works

Scholarly cross-domain collaborations have generated huge impact to society. In general, based on observations and experiments on real-world scholarly database with big data, our work helps to judge and explain whether it is possible to build collaborations between different scholarly domains. Also, by exploring the *peak* pattern in the correlation of research work's influence, i.e., citations, and its cross-domain performance, i.e., related domains' number, we discover the interesting truth that only research works focusing on a certain number of domains can produce significant impact. And the citation count of a paper is fitted with our proposed Gaussian-like model, which can also predict the paper's future citation, thereby helping afford researchers further knowledge when conducting literature review and offers scholarly recommendation system more features in paper recommendation.

Besides, there are a lot of interesting future directions which are worthwhile exploration. Since no efforts have been made to study the "crossability" between different scholarly domains, it is important to find proper ways to rigorously evaluate the accuracy of the quantified "crossability". And when studying *peak* pattern, we only consider literature's domain number as cross-domain performance while not digging into these domains. Therefore, in future study, we will try to find the ground truth for evaluating the quantified "crossability" between domains. Last but not least, we will explore more detailed features of paper's domain distribution to present research work's cross-domain performance.

References

[1] J. Tang, S. Wu, J. Sun, and H. Su, "Cross-domain collaboration recommendation," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1285–1293.

[2] (2016) Microsoft academic graph. [Online]. Available: https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

[3] N. Kory, R. V. Farese, and T. C. Walther, "Targeting fat: mechanisms of protein localization to lipid droplets," *Trends in cell biology*, vol. 26, no. 7, pp. 535–546, 2016.

[4] T. Mikuni, J. Nishiyama, Y. Sun, N. Kamasawa, and R. Yasuda, "High-throughput, high-resolution mapping of protein localization in mammalian brain by in vivo genome editing," *Cell*, vol. 165, no. 7, pp. 1803–1817, 2016.

[5] R. Tachibana, T. Terai, G. Boncompain, S. Sugiyama, N. Saito, F. Perez, and Y. Urano, "Improving the solubility of artificial ligands of streptavidin to enable more practical reversible switching of protein localization in cells," *ChemBioChem*, 2017.

[6] M. F. Collen and M. J. Ball, *A history of medical informatics in the United States*. Springer, 2015.

[7] A. Burnett-Hartman, P. A. Newcomb, C. X. Zeng, Y. Zheng, J. M. Inadomi, C. Fong, M. P. Upton, and W. M. Grady, "Abstract pr05: Using medical informatics to evaluate the risk of colorectal cancer in patients with clinically diagnosed sessile serrated polyps," 2017.

[8] D. Hristovski, A. Kastrin, and T. C. Rindflesch, "Implementing semantics-based cross-domain collaboration recommendation in biomedicine with a graph database," *DBKDA 2016*, p. 104, 2016.

[9] S. J. Derry, C. D. Schunn, and M. A. Gernsbacher, *Interdisciplinary collaboration: An emerging cognitive science*. Psychology Press, 2014.

[10] B. Taebi, A. Correlje, E. Cuppen, M. Dignum, and U. Pesch, "Responsible innovation as an endorsement of public values: The need for interdisciplinary research," *Journal of Responsible Innovation*, vol. 1, no. 1, pp. 118–124, 2014.

[11] B. K. Sovacool, "Energy studies need social science," *Nature*, vol. 511, no. 7511, p. 529, 2014.

[12] H. Rabbani, "Interdisciplinary researches in iran v: Toward interdisciplinary technologies," *Journal of medical signals and sensors*, vol. 6, no. 3, p. 129, 2016.

[13] D. He and W. Jeng, "Scholarly collaboration on the academic social web," *Synthesis Lectures on Information Concepts, Retrieval, and Services*, vol. 8, no. 1, pp. 1–106, 2016.

[14] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller, "Visualization as seen through its research paper keywords," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 771–780, 2017.

[15] N. Aggrawal and A. Arora, "Visualization, analysis and structural pattern infusion of dblp co-authorship network using gephi," in *Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on*. IEEE, 2016, pp. 494–500.

[16] C. McCarty, J. W. Jawitz, A. Hopkins, and A. Goldman, "Predicting author h-index using characteristics of the co-author network," *Scientometrics*, vol. 96, no. 2, pp. 467–483, 2013.

[17] J. Portenoy and J. D. West, "Dynamic visualization of citation networks showing the influence of scholarly fields over time," in *International Workshop on Semantic, Analytics, Visualization*. Springer, 2016, pp. 147–151.

[18] S. Dawson, D. Gašević, G. Siemens, and S. Joksimovic, "Current state and future trends: A citation network analysis of the learning analytics field," in *Proceedings of the fourth international conference on learning analytics and knowledge*. ACM, 2014, pp. 231–240.

[19] M. E. Newman, "Power laws, pareto distributions and zipf's law," *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.

[20] C. Xie, L. Yan, W.-J. Li, and Z. Zhang, "Distributed power-law graph computing: Theoretical and empirical analysis," in *Advances in Neural Information Processing Systems*, 2014, pp. 1673–1681.

[21] L. Muchnik, S. Pei, L. C. Parra, S. D. Reis, J. S. Andrade Jr, S. Havlin, and H. A. Makse, "Origins of power-law degree distribution in the heterogeneity of human activity in social networks," *arXiv preprint arXiv:1304.4523*, 2013.

[22] T. Honicke and J. Broadbent, "The influence of academic self-efficacy on academic performance: A systematic review," *Educational Research Review*, vol. 17, pp. 63–84, 2016.

[23] X. Zhu, P. Turney, D. Lemire, and A. Vellino, "Measuring academic influence: Not all citations are equal," *Journal of the Association for Information Science and Technology*, vol. 66, no. 2, pp. 408–427, 2015.
[24] D. Needell, R. Ward, and N. Srebro, "Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm," in *Advances in Neural Information Processing Systems*, 2014, pp. 1017–1025.
[25] S. Sra, S. Nowozin, and S. J. Wright, *Optimization for machine learning*. Mit Press, 2012.