Building and Analyzing a Global Co-Authorship Network Using Google Scholar Data

Yang Chen[1,2], Cong Ding[3], Jiyao Hu[1,2], Ruichuan Chen[4], Pan Hui[5] and Xiaoming Fu[6]

[1]School of Computer Science, Fudan University, China
[2]SKLCS, Institute of Software, Chinese Academy of Sciences, China
[3]Department of Computer Science, Cornell University, USA
[4]Nokia Bell Labs, Germany
[5]Department of Computer Science & Engineering, The Hong Kong University of Science and Technology, Hong Kong
[6]Institute of Computer Science, University of Goettingen, Germany
{chenyang,hujy13}@fudan.edu.cn, cong@cs.cornell.edu,
ruichuan.chen@nokia-bell-labs.com, panhui@cse.ust.hk, fu@cs.uni-goettingen.de

ABSTRACT

By publishing papers together, academic authors can form a **co-authorship network**, modeling the collaboration among them. This paper presents a data-driven study by crawling and analyzing the vast majority of author profiles of Google Scholar. We make the following major contributions: (1) We present a demographic analysis and get an informative overview of the authors from different aspects, such as the distribution of countries, scientific labels, and academic titles. (2) Based on the publication lists of crawled authors, we build a global co-authorship network with 402.39K authors to study the collaboration among authors. With the aid of social network analysis (SNA), we observe several unique features of this network. (3) We explore the relationship between the co-authorship network and citation metrics. We find a strong correlation between PageRank and h-index.

Keywords

Co-Authorship Network, Social Network Analysis, Citation Metrics, Google Scholar

1. INTRODUCTION

Examining the scientific impact of an individual scholar is useful in various scenarios. For instance, a funding agency's review panel may want to know which applicant performs better in a particular domain, a faculty recruitment committee might want to rank the candidates accurately, and a prospective graduate student might wish to find out the most promising advisor. On one hand, an author can be evaluated by simply looking at citations of her published papers. There are a number of *citation metrics* to evaluate an individual's scientific impact, such as total number of citations [6], h-index [6], and g-index [3]. On the other hand,

how well an individual is connected with other scholars is also important. Co-authorship network [4, 9, 15] has been proposed to model collaborations among scholars. By undertaking a social network analysis (SNA), several network-based metrics can be calculated, such as the number of co-authors an individual has. For both citation metrics and the co-authorship network, conducting a massive data-driven study based on authors from all over the world is a challenging but rewarding task, as we can gain numerous insights for evaluating the scientific impact of authors from a global and interdisciplinary perspective.

Based on the powerful Google search engine, Google Scholar [1] is a representative online search engine that collects and indexes the information of massive scholarly literatures. It covers major scientific disciplines and academic publishers, and it has been a useful platform for scholars worldwide. Since November 2011, a feature known as the **Google Scholar author profile**, has become a public service. With a Google account, any author is able to manage her publicly visible author profile by importing and organizing her existing publications. Moreover, Google helps every author calculate her citation metrics including total number of citations, h-index, and i10-index. We choose Google Scholar for our data-driven analysis because 1) it automatically collects the latest publications from databases across various disciplines to achieve a wide coverage, and it timely updates the citation status among its collected publications; 2) the self-management nature of author profiles allows each author to explicitly select the papers published by herself. This feature is very useful for solving the well-known author name disambiguation problem [14] in creating online author profile pages. Instead, if we just refer to an author's name to find all her publications, we might get some papers published by someone else, as different authors might have the same name.

In this paper, we carry out our study by crawling and analyzing author profiles of Google Scholar. By performing extensive crawling using a distributed method, we have collected 812.98K author profiles from Google Scholar, which covers most if not all existing public author profiles. By scanning these profiles and filtering out some inaccurate ones, we create a dataset with more than 402.39K authors. This dataset gives us an unprecedented opportunity to study

---

[1]https://scholar.google.com/

several interesting problems. Our main contributions include:

1. We perform a demographic analysis of crawled authors, and obtain the distribution of scientific labels, affiliation countries, and academic titles. We can see that a large part of authors are related to computer science or biology. Also, authors from the United States cover more than 30% of the entire author base.

2. We build a global co-authorship network of more than 402.39K authors. We use a set of representative metrics in SNA, such as degree, clustering coefficient, PageRank, and connected components to evaluate the collaboration among authors. We have found a number of unique features of the co-authorship network.

3. We explore the correlation between the co-authorship network and citation metrics. We conclude that getting an "important" position in the co-authorship network is a good indicator of a higher h-index.

2. BACKGROUND AND RELATED WORK

2.1 Online Author Profile Pages

Several major online bibliographic databases such as Google Scholar, Microsoft Academic Service [12], and AMiner [13] allow individual scholars to create profile pages for themselves. In our study, we focus on Google Scholar due to its wide coverage. We first give a brief overview of an individual's author profile on Google Scholar. There are 6 main parts of each page as follows.

1. **Basic Information:** The basic information is manually entered by the author, covering several fields such as name, title/affiliation, scientific labels, email domain, and personal homepage. Particularly, Google provides email address verification. If an author can provide a professional email address at her institution, and verify the ownership of it, her email address can be authenticated as "verified".

2. **Photo:** Every author can optionally upload a photo.

3. **Citation Metrics:** Google Scholar calculates an author's three citation metrics including total citation, h-index, and i10-index. Moreover, the values of these metrics based only on the citations received in the last five years are also shown. These values are being updated from time to time.

4. **Yearly Citations:** The number of citations received in each year is displayed.

5. **Article List:** This list contains the information of the author's all publications, where each entry represents one paper. For each entry, the title, author list, journal/conference name, number of citations, and the publication year are shown. An author can easily import her publications by a convenient interface. In addition, an author can request Google Scholar to automatically discover and add new publications of her.

6. **Co-Author List:** Once two authors are listed in the author list of a paper, they are regarded as co-authors.

In Google Scholar, the co-author list is manually entered by each author. According to our observation, many authors have added only a small subset of her co-authors to this list, or even choose not to add any co-author.

Furthermore, Google Scholar profile also provides a "search authors" module. By inputting a keyword, Google Scholar is able to return hyperlinks to author profiles which contain this keyword.

2.2 Citation Metrics

To evaluate the scientific impact of an author, there are a number of existing citation metrics to evaluate an author's scientific achievements, according to the citations her papers have received. Representative citation metrics include:

1. Total citation: The number of citations to all publications [6].

2. H-index: H-index is defined as the largest number h, which satisfies that each of the top h articles has received at least h citations [6].

3. G-index: G-index is defined as the largest number g, which satisfies that the top g articles received at least g^2 citations in total [3].

Most of these metrics have been widely used for evaluating a scholar from different aspects. In particular, *total citation* and *h-index* have been chosen for Google Scholar's "author profile" pages. To gain a statistical view of these metrics, the real data of a large number of scholars is desired. Unfortunately, many of existing studies are based on small datasets with only tens or hundreds of authors [6, 10]

2.3 Co-Authorship Networks

A *co-authorship network* can be used to understand the collaboration status among authors. The network $G = (V, E)$ has a node set V and an edge set E. V contains all authors in the selected dataset. An edge (x, y) in E indicates author x and author y have published at least one paper together.

Uddin et al. [15] studied a co-author network with 5251 authors, by referring to the "steel structure" articles published in the Scopus bibliographic database. Morel et al. [9] built a co-authorship network with 174 Brazilian authors focusing on the dengue fever based on the data from the "Web of Knowledge" database. These networks, however, cover only a small number of authors from a specified research area.

3. DATA COLLECTION AND PREPROCESSING

There are three key requirements in data collection and preprocessing: (1) We need to crawl as many author profiles as possible on Google Scholar. (2) To ensure the trustworthiness of the data set, we need to discover and filter out the authors who have added other people's publications to their profiles. (3) We need to extract information from crawled author profiles. Besides the information explicitly displayed in author profiles, we need to do further calculation to get additional information.

In this section, we describe how we collect the author profiles from Google Scholar (§ 3.1). Then we demonstrate how

we manage to remove some inaccurate data (§ 3.2). Finally, we apply some pre-processing for our further comprehensive analysis (§ 3.3).

3.1 Data Collection

As discussed in [10], crawling a large number of author profiles from Google Scholar is not easy. There are several challenges for collecting a snapshot of all public author profiles on Google Scholar. First, different from Twitter and Facebook, there is no public API for obtaining author profiles from Google scholar. Still, we are able to obtain an author's profile through parsing HTML source code, using the user ID as a key. Second, traditional graph-based crawling/sampling methods, such as Breadth-First Search (BFS) and Random Walk (RW) [5] are all relied on the explicit hyperlinks between users. In an author profile page of Google Scholar, only the "Co-Author List" part contains links between authors. Unfortunately, these author profiles are loosely connected, e.g., up to 63.07% of authors have not listed any co-author. Therefore, the traditional web hyperlink-based crawler has a high chance of being trapped.

Luckily, we can leverage the "search authors" module to discover user IDs of all authors. We choose letters from `a` to `z` as keywords. Based on the search result of these 26 keywords, we are able to get the user IDs of 812,984 author profiles (as of May 29th, 2015). We believe that we have collected the vast majority of the public author profiles on Google Scholar as of that day. Then we deploy a cluster of 10 virtual instances from the Microsoft Azure platform. Each virtual instance has a unique IP address. We further crawl the profiles of all these authors using a distributed method [2]. Each crawler works at a moderate crawling rate to avoid generating too much traffic to Google Scholar.

After collecting the HTML source code of each author's profile, we implement an HTML parser to parse its useful information including user ID (the `user` parameter in the URL), name, title and affiliation, list of scientific labels, email domain, homepage URL, with photo or not, citation metrics (total citations, h-index, i10-index, and these values in the last five years), number of citations of each year, each article's information with its number of citations, and self-claimed co-author list.

3.2 Data Cleaning

In Google Scholar, some authors have added papers not published by themselves to their profiles. There might be several reasons. One possible reason is an author might intentionally add some papers to make her citation metrics look better. Another reason is due to the "automatic import" function of Google Scholar. Ideally, this function can help an author discover papers published by herself, and add them automatically to her profile. Unfortunately, some irrelevant papers might be added as well. This is due to the name representation format of Google Scholar, i.e., using the first name initial and last name. In other words, an author with a name "James Chen" will be represented as "J Chen" in Google Scholar. As a result, the "automatic import" function might add papers published by "John Chen" to the author profile page of "James Chen". According to our observation, many articles have been added in this way. Such articles will mislead us when discovering co-authorship collaborations.

By scanning the publication entries of each author, we group authors into three sets, i.e., positive, negative, and neutral. "Positive" means we tend to trust the article lists of authors within this set, while "negative" indicates we feel the article lists of authors within this set are not reliable. There is also a "neutral" set, covering the authors that we are not able to judge the trustworthiness of their article lists. For each author using English names, we examine all of her published articles one by one. Note that for each article, only the first name initial and last name of each author are shown on the crawled profile page. However, for each article, there is a linked web page, containing the detailed information of this article. By further crawling this page, we can get a list of full names of all authors. If an author's full name is shown among more than 95% of her publications' author lists, this author will be put into the "positive" set. In contrast, if her full name is not shown in more than 5% of her publications' author lists, and these papers are written in English, the author will be put into the "negative" set. For the rest of authors, some of them might use non-English names, and some might have added too many non-English publications. Since we are not able to judge the trustworthiness of their publication lists, we put them into the "neutral" set. There are 464.57K (57.14%), 142.36K (17.51%) and 206.05K (25.34%) authors in positive, negative and neutral sets, respectively. In our study, we focus on the authors in the "positive" set, and we require each selected author to be "verified". Finally, we get 402.39K authors to construct the co-authorship network. We believe these authors are more reliable.

3.3 Data Pre-Processing

After obtaining a "raw" data set, We need to pre-process the dataset before performing detailed analysis. For each author, we compute four additional metrics, i.e., g-index, country, scientific domain and co-authorship information.

G-index: We calculate g-index [3] as it is another widely-used citation metric, and it has not been offered by Google Scholar. Given the number of citations of each article a scholar has published, we can calculate the g-index of this author according to the definition in [3].

Country: To understand the country distribution of Google Scholar authors, we need to know which country each author is working in. Since verified authors all use the professional email addresses from their institution, we utilize the email address information for country detection. We first extract the email domains from author profiles, and then query the domain name servers (DNSes) to get the associated IP addresses. Finally, we translate the IP addresses to country codes using IPInfoDB API[2].

Scientific Domain: There are a large number of scientific domains. In our study, we pick *computer science* (CS) and *biology* (Bio) as two example domains to study. Let us take CS as an example for scientific domain detection. We scan the profile of every author, if any CS-related substring (e.g., `computer` and `cs dep`) appears in the affiliation field, or any CS-related keyword (e.g., `computer`) shows up in the scientific label field, we regard this scholar as a CS author. We also examine the email address. If there is a substring like (@cs, @informatik, @computer), this scholar will be regarded as a CS author, too. Among all authors, 20.61% of them are CS authors, and 20.47% of them are Bio authors.

Co-Authorship: Building the global co-authorship network of Google Scholar is not trivial. As we mentioned, an

---

[2]http://www.ipinfodb.com/

author might have not listed all the co-authors on her profile page. Therefore, we cannot simply use the "Co-Author List" part of the profile pages. Instead, to discover the undeclared co-authorship information, we scan the publication list of every author. If a publication has been listed by both author A and author B, then we regard A and B are co-authors, i.e., an edge (A, B) will be added to E. In the constructed network G, there are 402,392 nodes and 1,234,019 edges.

4. DATA ANALYSIS

In this section, we study the constructed global co-authorship network from different aspects. First, we conduct a demographic analysis to see the composition of the authors (§ 4.1). Second, we use a number of classic network metrics to provide a clear understanding of the network (§ 4.2), and further compare between authors from different groups (§ 4.3). Last but not least, we study the relationship between the co-authorship network and citation metrics (§ 4.4).

4.1 Demographic Analysis

Table 1: Top 10 Scientific Labels

Labels	#Authors
machine learning	9833
artificial intelligence	6725
computer vision	5522
bioinformatics	4608
data mining	3925
neuroscience	3727
robotics	3164
image processing	3067
software engineering	2782
ecology	2482

In this subsection, we present a series of demographic analysis for authors in our data set. We are interested in the distribution of the authors' country, scientific labels and academic titles.

The top five countries are United States, United Kingdom, Italy, Germany and India. Each of them covers 32.32%, 6.10%, 4.79%, 4.00% and 3.72% of the entire population, respectively. Clearly, United States has covered more than 30% of author profile owners, which represents the most influential country in the scientific community.

Table 1 lists the top 10 scientific labels. We find that many of the popular labels are related to computer science. The most popular scientific label is "machine learning". Also, we find that biology-related authors (including three labels: `bioinformatics`, `neuroscience`, and `ecology`) contribute a lot to the whole author set. Therefore, we make a comparative study between computer science scholars and biology scholars.

We also study the academic title distribution, and we focus on the titles in the academia, i.e., professors, postdocs (research associates), and students (research assistants). Among all authors, we are able to place 31.03% of them into one of these three categories. We can see that 19.27% of all authors can be identified as professors, 3.87% of all authors can be determined as postdocs, and 7.89% of all authors can be concluded as students. The rest 68.97% of authors have not provided enough information to identify their academic titles, or, some of them might work in the industry.

4.2 Co-Authorship Network Analysis

To analyze the network G, we examine the following representative network metrics, i.e., degree, clustering coefficient (CC), PageRank, and connected components. The definitions of these metrics are:

Degree: The degree of a node in a network denotes the number of edges connected to the node. For an author in G, the degree represents the number of her co-authors.

Clustering coefficient: The clustering coefficient (CC) of a node is defined as the fraction of pairs of the node's neighbors that are directly connected with each other.

PageRank: PageRank [11] is an algorithm to estimate the importance of the nodes in a network. It has been used by the Google Search Engine to rank webpages in its search results.

Connected components: A connected component is a subgraph in which any two nodes are connected to each other by paths. In addition, any node in this subgraph is not connected to any additional node in the supergraph.

The average degree of network G is 6.13. Differently, many of the mainstream online social networks (OSNs) have a much larger average degree. For example, Renren has an average degree of 20.95 [17], and Cyworld has an average degree of 31.64 [1]. Fig. 1(a) shows the cumulative distribution function (CDF) of the degrees in G. The median value of the degrees in G is 2. This loosely connected property is due to the fact that the network G is constructed in a professional way. Two authors cannot be connected if they have not published any paper together. Differently, in OSNs like Facebook or Renren, two users can be friends whenever both of them agree. Therefore, gaining a connection in the co-authorship network is much harder than in OSNs.

An author's clustering coefficient (CC) is the ratio of the number of links over all possible connections between her co-authors, which describes the regional connecting tightness of a graph. Fig. 1(b) shows the CDF of the clustering coefficients of the nodes in G. According to our study, the co-authorship network's average clustering coefficient is 0.20, which is much larger than that in the OSNs (0.14 in Renren [17], 0.16 in Cyworld [1]). This indicates that co-authors are often tightly connected. For instance, people from the same research group, or the same research project, have a higher chance to publish papers together. Similarly, we can see the CDF of PageRank in Fig. 1(c).

Fig. 1(d) shows the sizes of the largest 10 connected components. There are 133,159 connected components. The size of the largest connected component is 258,949, which covers 64.35% of all authors. The 2nd to 5th largest connected components have 14, 14, 13, and 13 nodes, respectively. Among all 133,159 connected components, 125,318 (94.11%) of them are singletons, i.e., each with one disconnected node only. In short, we can see the co-authorship network has one giant connected component, and a number of small connected components.

Fig. 1(e) shows the distribution of the shortest path lengths of all node pairs in the largest connected component (LCC). The average shortest path length is 5.96. Also, the average clustering coefficient is 0.30. Therefore, G has a small average shortest path length, and a high average clustering coefficient, both of which are key properties of small-world networks [16].

As in [8], we further study whether the "core" of the co-authorship network is densely connected. We undertake the
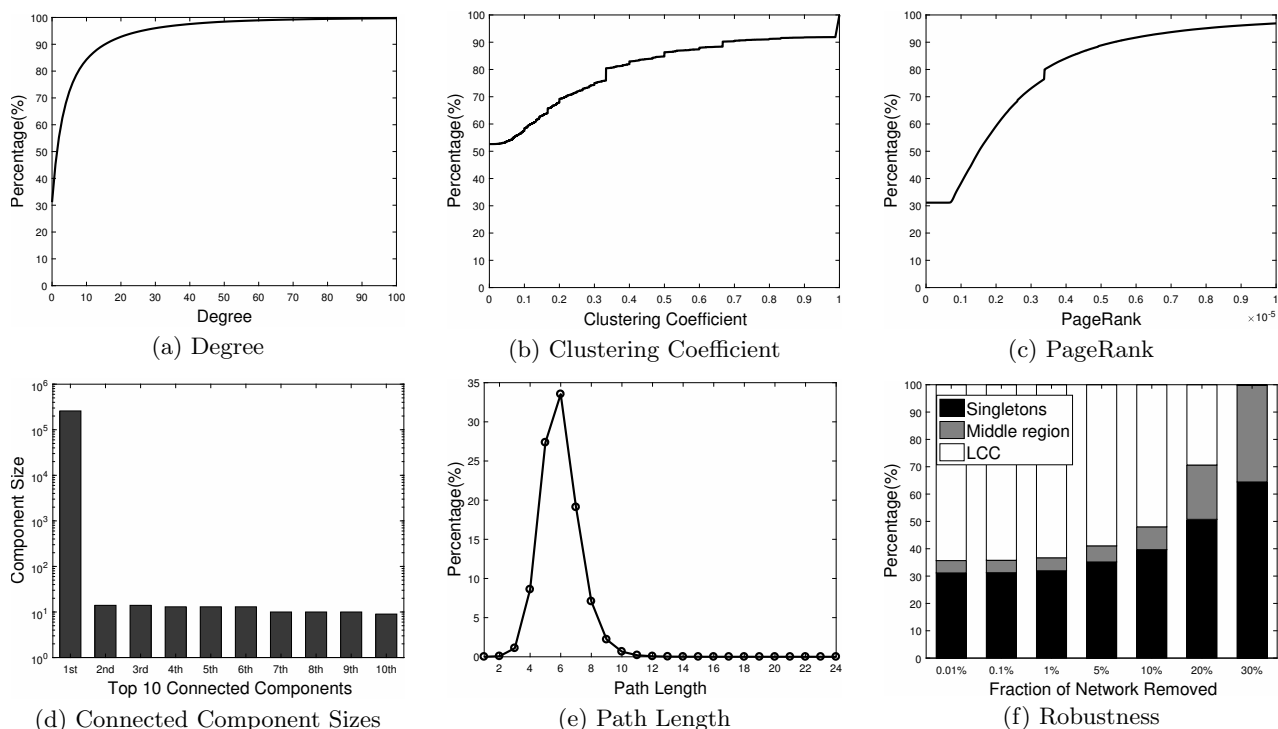
(a) Degree   (b) Clustering Coefficient   (c) PageRank

(d) Connected Component Sizes   (e) Path Length   (f) Robustness

Figure 1: Analysis of the Co-Authorship Network

investigation by removing some highest degree nodes from the network, and analyze the remaining nodes and edges. We increase the percentage of removed nodes from 0.01% to 30%, and we show the distribution of connected components of the remaining network in Fig. 1(f). To make a clear comparison, we group all connected components into three categories, i.e., the LCC, the singletons, and the middle region. We can see that even when we have removed 20% of highest degree nodes, we still have a huge LCC, covering about 30% of the remaining nodes. Therefore, the network is still well connected.

4.3 Comparison Among Different Author Groups

In addition to study the whole author base, we also compare among authors from different groups and show our results in Table 2. Considering the academic title, we can see the average degrees of professors, postdocs and students are 10.32, 4.75 and 1.85, respectively. The reason is simple as senior authors typically have more co-authors.

The average clustering coefficients of professors, postdocs, and students are 0.16, 0.27, and 0.22, respectively. We can see that both the postdocs and students have a higher clustering coefficient than professors. This is due to the fact the a professor has a higher chance to collaborate with researchers here and there. Differently, a postdoc or student would have more collaborations within her advisor's team, leading to a higher clustering coefficient.

Regarding PageRank, professors have the largest average value, while the students have the smallest average value. This indicates that compared to junior authors, senior authors are more "important" in the global co-authorship network.

We also classify users according to their research domains. In particular, we focus on authors related to computer science and biology. On average, computer science authors have more co-authors, and have a higher average clustering coefficient. This reflects the difference among the collaboration styles of different disciplines. Also, the average PageRank values of both computer science authors and biology authors are larger than that of the entire G. Therefore, these two groups of authors play a significant role in G.

4.4 Co-Authorship Network and Citation Metrics

As shown in Table 3, we evaluate the correlation between network metrics and citation metrics. We use Pearson correlation coefficient for the evaluation. A correlation coefficient is between -1 and 1. A positive value means a positive linear correlation. A value close to 1 denotes a strong linear correlation. Similarly, a negative value indicates a negative linear correlation. A value of 0 represents no linear correlation. We can see that the clustering coefficient metric has almost no linear correlation to each of the three citation metrics. Differently, either the degree metric or the PageRank metric has a moderate or strong positive correlation with each of the three citation metrics. In particular, the correlation coefficient between PageRank and h-index is as large as 0.73, which is quite strong. Therefore, the "importance" of an author in the global co-authorship network is a good indicator for her h-index.

Table 2 also shows the average values of the h-index and g-index of all authors and different author groups. The number of covered authors is much larger than existing work such as [6, 10]. We can see that senior authors perform better in

Table 2: Comparison Among Different User Groups

Criteria	Group	Avg. Degree	Avg. CC	Avg. PageRank	Avg. H-index	Avg. G-index
-	All	6.13	0.20	$2.49*10^{-6}$	8.34	16.67
Academic title	Professors	10.32	0.16	$3.60*10^{-6}$	13.86	28.00
	Postdocs	4.75	0.27	$2.13*10^{-6}$	6.17	12.52
	Students	1.85	0.22	$1.31*10^{-6}$	2.14	4.12
Domain	CS	8.54	0.24	$2.84*10^{-6}$	8.32	16.83
	Bio	7.80	0.22	$2.88*10^{-6}$	10.42	21.29

Table 3: Graph Metrics v.s. Citation Metrics (Pearson Correlation Coefficient)

	Degree	Clustering Coefficient	PageRank
Total Citations	0.53	-0.03	0.53
H-index	0.68	0.02	0.73
G-index	0.65	0.02	0.68

citation metrics. Also, the average values of citation metrics for computer science authors are smaller than that of biology authors. This confirms the finding that citation metrics are discipline dependent [7].

5. DISCUSSION AND FUTURE WORK

Google Scholar has a wide coverage of academic articles from different fields, but its relatively small number of author profiles limits the generality of our study. Nevertheless, it is expected that more and more authors from different disciplines will join this service due to Google's great user coverage.

As the next step, we plan to further explore the co-authorship network from the following aspects.

- The global co-authorship network covers authors from different scientific areas. We plan to explore this large network from an interdisciplinary perspective. We wish to see the difference and connections between different research disciplines.

- Besides studying an aggregate snapshot, we will further investigate the yearly evolution of the co-authorship network. More insights may be explored through the analysis of the dynamic features.

- We will further apply our findings from the co-authorship network in practical cases, such as evaluating the scientific impact of an individual or an institution.

Acknowledgement

6. REFERENCES

[1] Y.-Y. Ahn, S. Han, H. Kwak, Y.-H. Eom, S. Moon, and H. Jeong. Analysis of Topological Characteristics of Huge Online Social Networking Services. In *Proc. of WWW*, 2007.

[2] C. Ding, Y. Chen, and X. Fu. Crowd Crawling: Towards Collaborative Data Collection for Large-scale Online Social Networks. In *Proc. of ACM COSN*, 2013.

[3] L. Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[4] B. Fonseca, R. B. Sampaio, M. V. d. A. Fonseca, and F. Zicker. Co-authorship network analysis in health research: method and potential use. *Health Research Policy and Systems*, 14(1):34, 2016.

[5] M. Gjoka, M. Kurant, et al. Walking in Facebook: A case study of unbiased sampling of OSNs. In *INFOCOM*, 2010.

[6] J. Hirsch. An index to quantify an individual's scientific research output. *PNAS*, 102(46):16569–16572, 2005.

[7] C. Mccarty, J. W. Jawitz, A. Hopkins, and A. Goldman. Predicting Author H-index Using Characteristics of the Co-author Network. *Scientometrics*, 96(2):467–483, Aug. 2013.

[8] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proc. of ACM IMC*, 2007.

[9] C. M. Morel, S. J. Serruya, G. O. Penna, and R. Guimarães. Co-authorship Network Analysis: A Powerful Tool for Strategic Planning of Research, Development and Capacity Building Programs on Neglected Diseases. *PLOS Neglected Tropical Diseases*, 3(8):1–7, 08 2009.

[10] M. Olensky, T.-H. Tsai, and K.-T. Chen. H-index Sequences Across Fields: A Comparative Analysis. In *Proc. of WWW Companion*, 2016.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Stanford InfoLab Technical Report 1999-66, November 1999.

[12] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang. An Overview of Microsoft Academic Service (MAS) and Applications. In *Proc. of WWW Companion*, 2015.

[13] J. Tang. AMiner: Mining Deep Knowledge from Big Scholar Data. In *Proc. of WWW Companion*, 2016.

[14] J. Tang, A. C. M. Fong, B. Wang, and J. Zhang. A unified probabilistic framework for name disambiguation in digital library. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):975–987, 2012.

[15] S. Uddin, L. Hossain, A. Abbasi, and K. Rasmussen. Trend and efficiency analysis of co-authorship network. *Scientometrics*, 90(2):687–699, 2012.

[16] D. J. Watts. Networks, Dynamics, and the Small-World Phenomenon. *American Journal of Sociology*, 105(2):493–527, 1999.

[17] X. Zhao, A. Sala, C. Wilson, X. Wang, S. Gaito, H. Zheng, and B. Y. Zhao. Multi-scale Dynamics in a Massive Online Social Network. In *Proc. of ACM IMC*, 2012.