Understanding Hierarchies in Computer Science Conferences

Aashutosh Trivedi
New York University
NY, USA
aashutosh.trivedi@nyu.edu

Aditya Garg
New York University
NY, USA
aditya.garg@nyu.edu

Ansa Mary Ephraim
New York University
NY, USA
ame378@nyu.edu

***Abstract*—**

A bibliometric study based on a number of publishing venues in the field of Databases in Computer Science has been carried out. We find the author communities of the top tier conferences to be more similar to each other than the lower tiers. In line with an increasing demand to publish more, as well as better connectivity across the board, scientists were found to be collaborating more. Publishing activity is highest in the topmost tier of flagship conferences.

Keywords*— *co-authorship, analytics, bibliometrics, similarity

I. Introduction

A considerable number of papers are published in the field of Computer Science at a variety of venues. These conferences and journals are not regarded to be of equal standing, some of them are known to be more prestigious than others. This paper explores and compares the structure of the community associated with different tiers of conferences in Computer Science. Analysis is carried out to detect similarities in the author community. Trends in paper publication and in author collaboration are studied over the years and across different tiers. Although the study was restricted to venues in Databases to make for coherent and meaningful results, it can easily be replicated for different fields.

II. Motivation

Conferences have been ranked over time and have *reputations* in the research community. Newcomers to the field are perplexed by the pecking order, and the decision of which venue to submit to is a daunting one. An understanding of the community associated with each tier can help an author choose a venue based on where the paper is more likely to be accepted. Further, based on his research interests, he can find other venues that would be similar to what would have been his first choice. A scientist could also find other researchers with common interests to collaborate with based on the co-authorship graph metrics. It would also help scientists to pace themselves against the average productivity of other researchers in the tier they want to belong to.

Finally, it is interesting to gain an understanding about how collaboration is increasing over the years.

III. Related Work

CSCW conferences Analysis [3] focused on finding the strongly connected core of the conferences. The authors looked for various clusters, specially the clustering of social science and computer science papers. They also realized that there were papers which were cited outside the conference more than they were cited in it. Studying 25 years of SIGIR [4] conferences allowed content analysis of the same. Researchers studied the trending topic and co-authorship in areas of databases, evaluation, probabilistic and language models.

[5] suggests that a large number of collaborators have a first degree connection amongst them. The empirical proof was given by analyzing 4 databases from different fields and comparing the results. It showed that, irrespective of size of database, 75% of the collaborators have a first degree connection. [6] supports this inference by providing a different view for the same problem. It clusters the collaborators by harmonic distance to find the degree of supportiveness, which is the metric of how much a co-author of a particular paper supports the research.

SIGMOD's co-authorship graph has been analyzed using data from DBLP [7]. This graph was analyzed in terms of number of authors per paper, number of papers per author and the longest path distance. It was shown that in no year was the co-authorship graph a single connected component. For the single largest connected component every year, the size and clustering coefficient was also studied. The paper also explored the existence of hubs in the graphs, and measured their centrality. In some Analyses of Erdos″ Collaboration graph, many metrics regarding co-authorship were analyzed [8]. It analyzed the top authors based on the number of their co-authors. It also described the concept of *cores* collaborativeness. Lords, vertices that have strong influence on their neighborhoods were studied. This paper analyzed the community from the point of view of a social network and characterized its graph properties.

All the above analyses were performed on a snapshot of co-authorship graphs. They focus on the nature of the connected graph as a whole, rather than properties of its nodes. Our work measures similar metrics, but observes them across venues and tiers, as well as over the years.

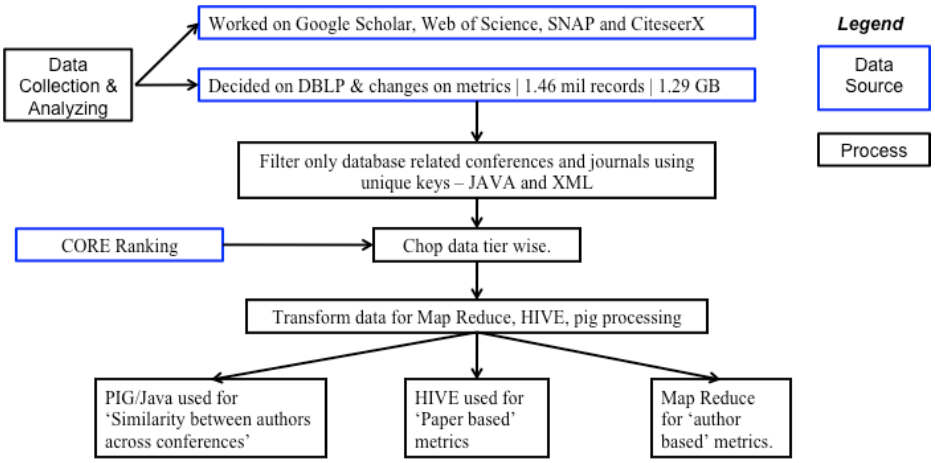

Figure 1. Design diagram for the research

IV. DESIGN

Figure 1 describes the work flow. Data was initially collected and analyzed from a number of different sources. They were evaluated in terms of their ease of collection, coherence and completeness (for collaboration and citation). The DBLP data was filtered to analyse papers published in database conferences. The data was split according to conferences tiers. It was then preprocessed to be used as input for Hive and Map Reduce. Pig, Hive, Java and Map Reduce were used to analyze similarity measures, and author and paper based metrics.

V. RESULTS

Data collection was the starting point, and the biggest hurdle of this project. Table 1 describes the data sources considered, but they all had drawbacks.

	ACM	DBLP	CiteSeer X	Google Scholar	Web of Science
Free	Partly	Yes	Yes	Yes	No
Download	No	Yes	Yes	No	Yes
Citation info	Yes	Few	Yes	Yes	Yes
# records	1.59 mil	1.46 mil	32.23 mil	NA	45.68 mil

Table 1. Databases considered for analysis [9]

- **SNAP co-authorship graphs**: This was restricted to Physics Arxiv papers [10], and hence, was outside our realm of interest. Also, there was not enough data for meaningful analysis.
- **Google Scholar**: Given that it was not available for direct download and did not have an API for streaming, we explored the utility of a Python based web crawler named Scrapy. This required a mapping of the html structure of the Google Scholar pages [11], and had other annoyances such as download limits requiring frequent restarts. It turned out to be a time consuming effort with meager results: the data streamed still required cleaning and indexing.
- **DBLP**: This database [12] offered 1.3 Gb of annotated XML data. This included information about year of publication, venue and authors, but was lacking citation and keyword data.
- **Citeseer from website**: Data downloaded from the website was found to be missing citation data. [13] It was discovered that Citeseer had migrated to a new version, CiteseerX, and this rendered all the older help pages and download tutorials ineffective.
- **CiteseerX**: We contacted Penn State University directly, and they were very kind to provide us with the metadata dump, 40 GB of XML and 60 GB of a mySQL dump. This is a d atabase built automatically using a web crawler and some basic indexing. The data was stored on HPC cluster and it took some time getting a handle on the structure of the data. Python, XSLT and Bash scripts were used to explore the data initially. HPC did not have MySQL installed, so a batch script was modified to work line by line to convert it to SQLite. However, at this point we realized that the data required cleaning, deduplication and indexing. There was no DOI mapping between the papers and those cited. This project did not afford us the time required for all this, but we believe that this is valuable data that can help answer many questions this research set out to answer.

We zeroed in on the DBLP database and found it satisfactory for the rest of the project. With entries having been made manually, they were clean, unique and well annotated. We decided to focus on one field so that our analysis was not diluted and the results could be cohesive and more meaningful.

81 conferences in the field of Databases were identified. XML/XSLT and Java programming were used to filter the data and retrieve only papers published at these venues. The goal was to study trends in the field across the years as well as compare them across differently ranked conferences.

Tier 1	ICDE, ICDM, PODS, SIGMOD, VLDB
Tier 2	CIDR, CIKM,COOPIS,DASFAA, DBSEC, DEXA, DOOD, EDBT, FODO, ICDT, IDA, PAKDD, PKDD, SDM, SIAM, SSDBM, WISE
Tier 3	ADBIS, ADC, BNCOD, COMAD, DAWAK, IDEAS, MDA, RIDE, SBBD, VDB
Tier 4	ARTDB, CDB, DBLP, DMDW, DMKD, DOLAP, EFIS/EFDBS, EFIS, EFDBS, FQAS, IDEAL, IW-MMDBMS, KR, KRDB, NLDB, OODBS, OOIS, RTDB, WEBDB, WIDM

Table 2. Conferences divided by tier

The data was divided into tiers based on the four tier grading awarded to conferences by CORE (Computing Research and Education) [14]. Tier 1 corresponds to Flagship Conferences, Tier 2 are considered excellent, Tier 3 are good, and Tier 4 receive honorable mention. Table 2 lists the conferences placed in each tier.

We experimented with Mahout library's XMLInputFormat for Map Reduce [15], but ended up chasing strange build errors. We could not find further online support for this, and hence switched to Java programming to flatten out the XML into a CSV. This could easily be read into MapReduce one record at a time, or loaded into Hive as a table. While we wrote plain Java code for this task, we also discovered JDOM Parsers that make dealing with XML files in Java extremely simple. JDOM parses the file as a tree and allows iteration across children at each level.

Our analytics span three topics: similarity between the author community, author based metrics and paper based metrics.

Conference Similarity:

Similarity for the author communities was measured using the Jaccard similarity measure [16]: if A and B represent the sets of authors at two venues, then the similarity between the two is defined as $|A \cap B| / |A \cup B|$.

Figure 2 (normalized to 0.3) shows that conferences in tier 1 and 2 are more similar to each other rather than those in 3 and 4.

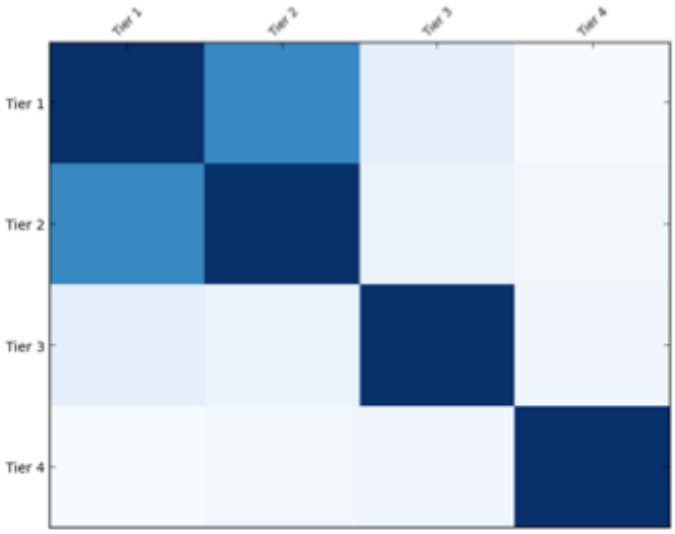

Figure 2. Similarities between the four tiers (normalized)

Figure 3 shows similarities across all the conferences ordered by ranking. The top left corner has the highest ranked conference. Other than a few exceptions, conferences in the top tiers are more similar, and get diluted down the tiers. SIGCOM and VLDB are similar to each other, as is ICDM. RTBD and ARTDB were found to be unusally similar, in tier 4.

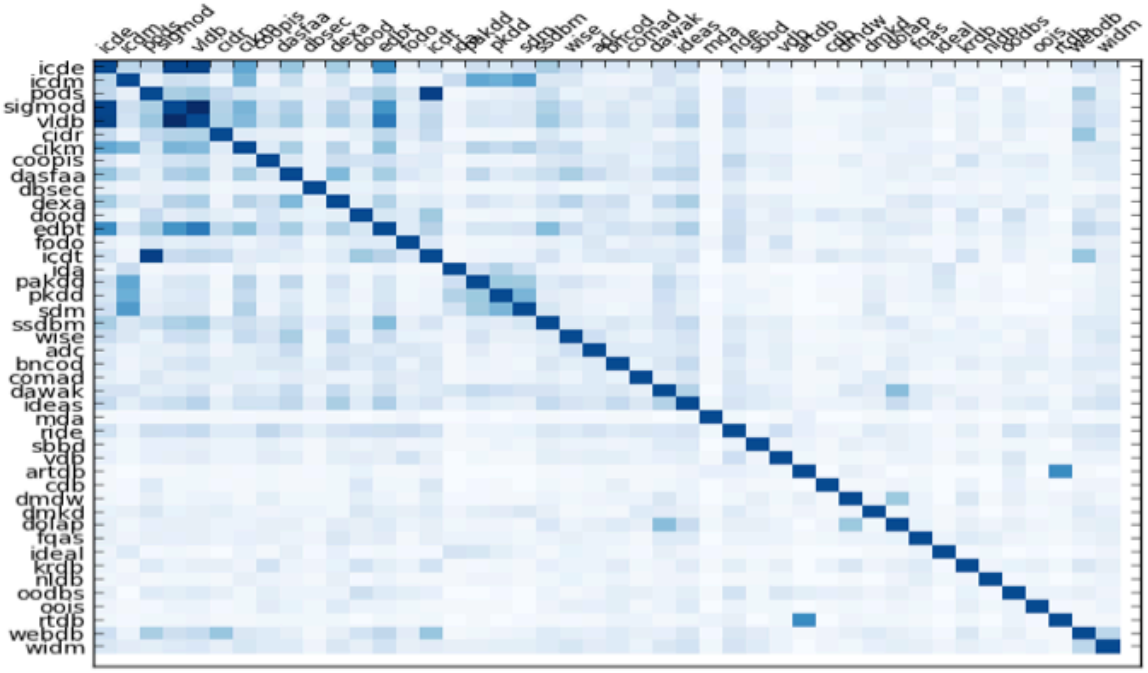

Figure 3. Similarities across conferences (normalized)

Author based Analyses:

Five statistics were analyzed using MapReduce, some of these required multiple MapReduce rounds.

We sought to find out the number of authors in the field over the years. This would give us some bearing of how much research was attracting people and how many people were putting in effort to find answers and create cutting edge technology. Figure 4 shows our findings for the same. The number of authors publishing every year has been increasing at a healthy rate, which is very encouraging as it goes on to show that more and more people are getting involved in research with every year. What is even more encouraging is that the rate of increase is higher for higher tiered conferences as compared to lower tier.

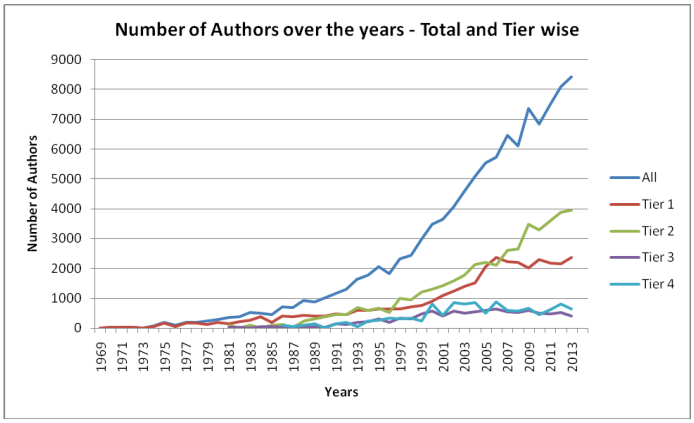

Figure 4. Number of Authors over the years

After realizing the increase in number of authors, we tried to measure how much research is being conducted by every author. One approach to this was, number of papers published for every author over the years. As see in Figure 5, while this has remain more of less constant over the years, (increase of 0.2 in 30 years), we still find this graph to somewhat expected and very healthy. As is evident, number of papers per author is higher in top tier conferences, a sign of higher and better research. This graph also to a very high extent depicts Lotka's law. As per the law, 60% of the authors will write/publish 1 paper and number of people publishing *'n'* papers is about *'1/n²'* of those making 1 publication.

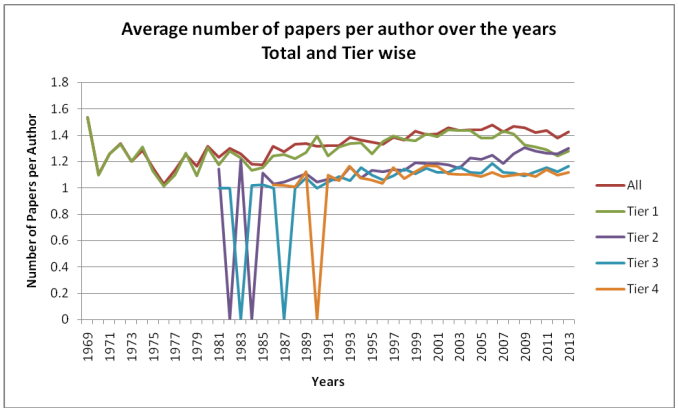

Figure 5. Number of papers per author over the years

We then turned our focus towards various factors that affect collaboration opportunities in the field. Figure 6 shows average number of collaborators for every author, and Figure 7 shows number of single authored papers over the years. As we see, number of collaborators are increasing, but at a very small rate. This can be attributed to the fact that most of the conferences have a limit on number of collaborators for every submission (2 to 4 usually). Figure 7 on the other hand shows a stark decrease in number of single authored papers. This decrease can be attributed to reasons like better networking and collaboration opportunities and technological advances to support the same. Higher numbers of collaborators also allow better quality research and a different perspective on the same problem. As is clear, these results are very strongly related. Tier 3 results show even more correlation in these 2 results. From 1983-85, the number of collaborators was very high in Tier 3 conferences, while when that reduced in 1986, there was a sharp increase in number of single authored papers.

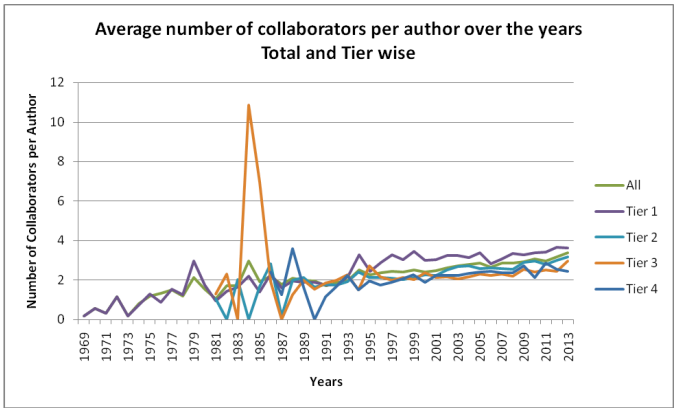

Figure 6. Average number of collaborators per author

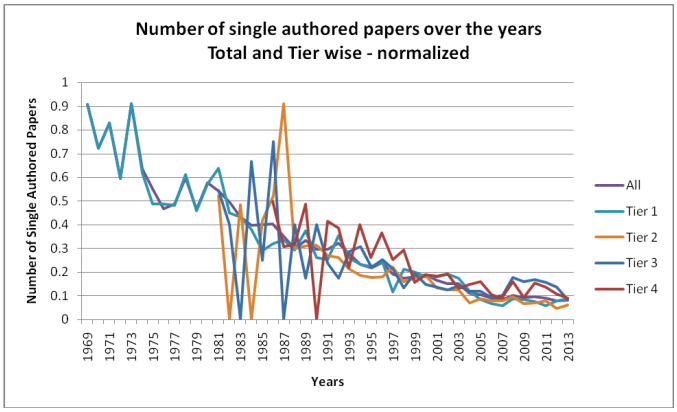

Figure 7. Number of single authored papers over the years

We realized that at any time, for any year, only about 20% of the DBLP database community is active. However, there was still an increase in number of papers every year. So we turned our focus to number of new authors that are coming in every year. Figure 8 shows our results for the same.

We find that this measure has been more or less stable for last 20 years or so. But when we analyze the same tier wise, we see more new authors in lower ranked conferences as compared to higher rank.

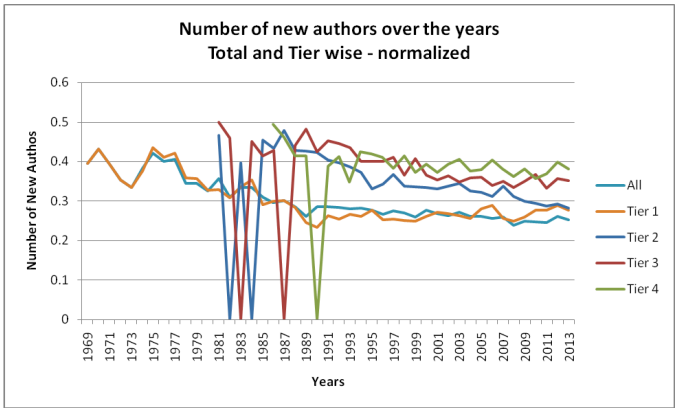

Figure 8. Average number of new authors over the years

We attribute this to the fact that any new author would usually try to first present their papers into lower ranked conferences, hoping the chances of publication to be higher.

We could not account for some of the strange blips in the graphs in some years. As it is unlikely that no paper at all was published in some tier, we believe that this must be due to limitations of the DBLP data.

Paper based Analyses:

For the paper based metrics, we used HIVE because of its easy support for data querying [17]. As only certain number of fields were required from the entire data, we used only the required fields (key, year, conference name, authors) from the flattened csv. The resulting record is as shown

conf/icod/AusielloBM80 1980 icod Giorgio Ausiello|Carlo Batini|Marina Moscarini

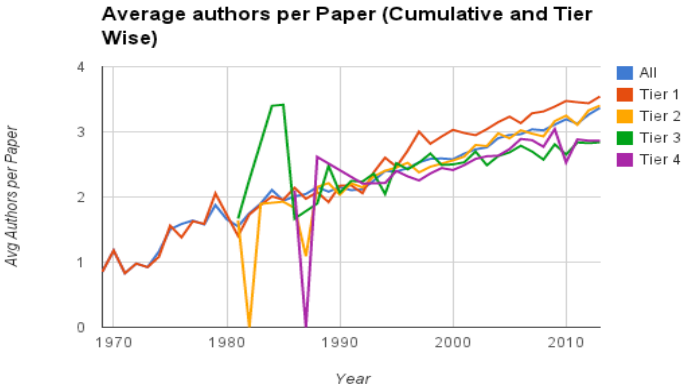

Figure 9. Average number of authors per paper

The various tier files were created by taking the names of conferences in each tier and putting them in a file. Then each conference name in a particular tier file was compared with the conference name in the original file. If they were a match, it meant that the paper was presented in a conference belonging to that tier, and was entered as such. This was done using a JAVA program and new files were made containing records for each tier. The analytics were run against the entire original file and each tier files as well. These records were loaded into various different HIVE tables.

As seen in Figure 9 the average number of authors per paper shows an increasing trend each year. This supports the theory that more authors are trying to collaborate and co-author research. This trend is seen irrespective of the Tier in which the paper is presented.

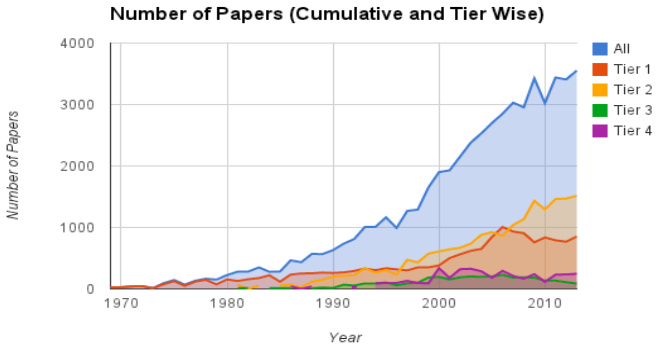

Figure 10. Total number of papers every year

The number of papers presented over the years have been on a high rise as well. Interesting to see in Figure 10 is that the top two tier conferences have a higher increase in the number of papers presented over the years. This is not because there are more conferences in the top two tiers. As a matter of fact, both the tiers combined have lesser conferences than the bottom two tiers. This shows that more papers are being presented at flagship and renowned conferences.

VI. Future Work

Combined with citation data, this work can be expanded to include citation graph metrics. Clustering of keywords in papers accepted can be used to identify popular topics at venues by year. An analysis of paper publication and citation metrics for panel members, together with information about the papers they accept could help identify any signs of panelists' bias towards collaborators, cited works or topics in the same (or different) domain of research. Put together, this could provide a strong venue recommendation system for new authors.

VII. Conclusion

A greater similarity between the author community of conferences in tiers 1 and 2 was established. An important caveat about our work is that all trends reflect on the accuracy and completeness of the DBLP data. A number of conferences were known to be about computer science but could not be ranked in one of the four tiers. There may be more conferences in each of the tiers that may have been missed. However, we still believe that the trends observed will hold in the most part and will not see much fluctuation because of these relatively minor omissions.

ACKNOWLEDGMENT

We are grateful to Lee Giles and Douglas Jordan of Penn State University for sharing the CiteseerX data with us. We also thank Shenglong Wang for support with the HPC cluster. We are especially grateful to Prof. Suzanne McIntosh for valuable discussion, guidance and facilitation.

REFERENCES

[1] T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.

[2] A. Gates. Programming Pig. O'Reilly Media Inc.,Sebastopol, CA, October 2011.

[3] Michal Jacovi, Vladimir Soroka, Gail Gilboa-Freedman, Sigalit Ur, Elad Shahar, Natalia Marmasse. The Chasms of CSCW : A Citation Graph Analysis of the CSCW Conference, In Proceedings of the 2006 Conference on Computer-Supported Cooperative Work

[4] Alan F. Smeaton , Gary Keogh , Cathal Gurrin , Kieran McDonald , Tom Sødring, Analysis of papers from twenty-five years of SIGIR conferences: what have we been doing for the last quarter of a century?, ACM SIGIR Forum, v.36 n.2, Fall 2002

[5] M.E.J. Newman. The structure of scientific collaboration networks, Proceedings of the National Academy of Sciences 98(2):404-409 (2001)

[6] Yi Han, Bin Zhou, Jian Pei, Yan Jia. Understanding Importance of Collaborations in Co-authorship Networks: A Supportiveness Analysis Approach.

[7] Mario A. Nascimento , Jörg Sander , Jeffrey Pound, Analysis of SIGMOD's co-authorship graph, ACM SIGMOD Record, v.32 n.3, p.8-10, September 2003

[8] Vladimir Batagelj and Andrew Mrvar. Some Analyses of Erdos˝ Collaboration Graph.

[9] Dalibor Fiala. Mining citation information from Citeseer data, University of West Bohemia, Czech Republic

[10] http://snap.stanford.edu/data/#citnets

[11] http://doc.scrapy.org/en/latest/intro/tutorial.html

[12] http://dblp.uni-trier.de/db/

[13] http://citeseerx.ist.psu.edu/index

[14] http://core.edu.au/index.php/categories/conference%20rankings/1

[15] http://xmlandhadoop.blogspot.com/

[16] Ergin Elmacioglu , Dongwon Lee, On six degrees of separation in DBLP-DB and more, ACM SIGMOD Record, v.34 n.2, June 2005

[17] E. Capriolo, D. Wampler and J. Rutherglen. Programming Hive. O'Reilly Media Inc.,Sebastopol, CA, October 2012

[18] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In proceedings of 6th Symposium on Operating Systems Design and Implemenation, 2004.