Paper Title

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

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

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

***Abstract*—**

Many Computer Science conferences are held every year with a large number of papers published. Many conferences are considered prestigious, while some are not. We attempt to understand the motivation behind the hierarchy of these conferences in terms of the papers they publish. We propose to study this based on the papers they cite and the authors they collaborate with. It also intends to derive meaningful results about popular topics at a given conference, and the nature of the citation and collaboration graphs that make up its community.

Keywords*— *citation, co-authorship, analytics

I. Introduction

We access citation database of conferences over a period of time. Some analytics we hope to derive are

1) Potential Hot Topic : Paper with the least citation distance from all the published papers.

2) Find citation distance and collaboration distance between authors

3) Help suggest should two authors work together in future.

4) Suggest/predict conferences to submit entries to for authors

II. Motivation

(Write a couple of paragraphs describing why you think this analytic is important. Why should people care about this analytic?)

III. Related Work

Analysis of SIGMOD's Co-Authorship Graph :
The first input gathered from this paper was a valuable data source : a DBLP XML data file that contains authorship info and panel members. A few points that were analyzed were : number of authors per paper, number of SIGMOD papers per author, the longest path distance of a collaboration graph. They found that in no year was the co-authorship graph a single connected component. For the single largest connected component every year, they also analyzed the size, the clustering coefficient, and tried to see if it exhibited the small world phenomenon. The paper also explored the existence of hubs in the graphs, and measured their centrality.

Some Analyses of Erdos˝ Collaboration Graph :
This was not exactly a well published paper, but it was hard to find popular publications that relate to our project as closely. It sets out some features of analysis that can be performed on large network graphs, and specifically carries them out on Erdos graphs.
It analyzed features regarding the co-authorship , ie the mean,median, average degree, maximum and maximizer of of vertex degrees in Erdos graphs. It analyzed the top authors based on number of co-authors. It also described the concept of *cores* and went on to analyze authors and number of co-authors in the main core, total number of co-authors, average core and and average degree of all their co-authors, and their collaborativeness. Another interesting idea they proposed was that of 'Lords' - vertices that have strong influence on their neighbourhoods. A recursive process, much like page ranking, finally assigns the 'power' to each vertex. Another analytic used was block modelling.

The Chasms of CSCW: A Citation Graph Analysis of the CSCW Conference :
The authors look at papers from the CSCW conference and tries to prove the following hypothesis.

- H1: There is a strongly connected core of the CSCW conference : Proven to some extent
- H2a: The CSCW conference is divided into several thematic clusters : Clearly evident from results of clustering
- H2b: Social science and computer science papers will reside in different clusters : Again evident from clustering
- H3: There are chasm-papers in the CSCW conference that are cited outside the conference significantly more than within it : Proven to some extent.

The paper divides the corpus into clusters using the betweenness centrality algorithm implemented in JUNG framework, which iteratively removes edges from citation graph. It also defines a success function and chasm potential to find chasm papers.

Analysis of Papers from Twenty-Five Years of SIGIR Conferences: What Have We Been Doing for the Last Quarter of a Century?:

The authors try to look at all the papers published in last 25 years in SIGIR conferences and perform a content analysis on the same. They try to determine how the trending topics have changed over time, which topics have come and gone. They have categories like databases, evaluation, Probabilistic and language models, Conceptual IR, users and search and general among others. They also created co-authorship graph (including cleaning up author names) and perform some analysis on that. These include authors with maximum papers, authors with greatest number of collaborators, and analyze the path between the authors, Erdos-type analysis. Authors then try to predict the hottest topic for next year and the co-authorship combination.

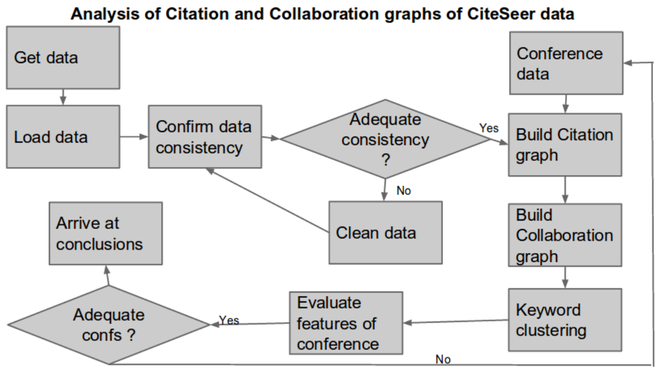
The structure of scientific collaboration networks: M. E. J. Newman:

The author considers 2 scientists to be connected to each other if they have co-authored a paper together. The paper argues that most people who have written a paper together will know one another quite well. Thus it is a moderately stringent definition. The author has constructed collaboration graphs for scientists in a variety of fields. The data come from four databases: MEDLINE (which covers published papers on biomedical research), the Los Alamos e-Print Archive (preprints primarily in theoretical physics), SPIRES (published papers and preprints in high-energy physics), and NCSTRL (preprints in computer science). In each case, he has examined papers that appeared in a 5-year window, from 1995 to 1999 inclusive. The sizes of the databases range from 2 million papers for MEDLINE to 13,000 for NCSTRL. The distance is calculated based on the co-authoring of the paper. 2 scientists are said to be at distance 1 if they have co-authored a paper together. Any authors who have worked whit these co-authors come at a distance 2 from the original author and so on. The paper takes a number of different scenarios such as the number of authors (and the errors such as same author giving different initials or different authors having same initials), Mean Papers per Author and Authors per Paper and number of collaborators. The paper also suggests that with a very large database and huge number of connections, the first degree connection takes up around 3/4 part of the database. Also the second connection group is much smaller than the first degree connection group. The author concludes that the maximum distance between 2 scientists is an average around 6 irrespective of the size of the database.

Understanding Importance of Collaborations in Co-authorship Networks: A Supportiveness Analysis Approach

This paper supports the theory that 2 authors co-authoring a paper means that one author supports the research of the other author. Thus it provides various such supportiveness measures. The definition of supportiveness here is given as "For an author a, the supportiveness from author b to a is used to measure how close the collaborations from b to a." The authors develop efficient methods to extract top n most supportive authors in co-authorship networks. They model the supportiveness ranking problem as a reverse k nearest neighbor (k-RNN for short) searching problem on graphs. To better model the co-authorship relation, we use hyper graphs in this paper. The authors use harmonic distance measure to calculate the support. It shows that support can differ in different directions based on the number of papers written in total and the number of papers co-authored. Thus the closeness and distance is calculated to be the harmonic mean of thecontribution from one author to the other. The k nearest neighbor algorithm is used to find the author with maximum support. The algorithm finds the neighbor having the lease harmonic distance and thus concludes that it can be a neighbor. Here one author can have multiple or one nearest neighbor, depending on different factors. The authors also have tried to expand the algorithm to work for not just a single vertex but also a group of vertices. They have discussed the efficiency of both problems on large co-authoring networks and come up with very interesting solutions.

Iς. DESIGN



A number of data sources such as Google Scholar, ACM, DBLP, CiteSeer and Web of Science were considered. Citeseer was chosen because of its availability, completeness, and categorization. The data was available in XML and MySQL files. The volume of 40 Gb necessitated the use of NYU's HPC cluster. The SQL format was chosen for ease and speed of querying. The data was not structured well and had to be cleaned over multiple rounds of querying till it could give accurate results.

<Future> Once the data has been cleaned, the citation graph will be built using a Breadth First Search starting with the list of papers published at a select conference, as well as papers published by panel members who are also authors. An Author collaboration graph will also be built similarly. Different analytics such as longest distance, clustering coefficient, central node, keyword clustering will be carried out. This will be repeated for a number of popular conferences – each conference will have its own graph and the same metrics calculated. Based on this, different inferences about each conferences such as whether they exhibit Small World phenomenon, popularity of papers linked to this conference and the most popular keyword will be deduced.

<Note : If the CiteSeer data proves futile after a number of iterations of cleaning and querying, we will switch to the DBLP database. It is already downloaded, and clean. However, then we will have to forgo the analysis on the citation graph, the rest of the analysis will proceed as in the original design>

RESULTS

(Future… In this section, you can describe: Your experimental setup/issues with data/performance/etc. Describe your experiments, describe what you learned. Did you prove or disprove your hypothesis? Were some results unexpected? Why?)

ς. FUTURE WORK

(Future… Given time, how would you expand your analytic? Could it be applied to other areas? Etc…)

CONCLUSION

(Future… One or two paragraphs about the value/accuracy/goodness of your analytic.)

ACKNOWLEDGMENT

(This section is optional. It can be used to thank the people/companies/organizations who have made data available to you, for example. You can list any HPC people who were particularly helpful, if you used the NYU HPC.)

REFERENCES

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

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

[3] S. Ghemawat, H. Gobioff, S. T. Leung. The Google File System. In Proceedings of the nineteenth ACM Symposium on Operating Systems Principles – SOSP '03, 2003.

[4] Mario A. Nascimento, Jorg Sander and Jeffrey Pound. Analysis of SIGMOD's Co-Authorship Graph

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

[6] 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.

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

[8] M.E.J. Newman. The structure of scientific collaboration networks.

[9] Alan F. Smeaton, Gary Keogh, Cathal Gurrin, Kieran McDonald and Tom Sødring. Analysis of Papers from Twenty-Five Years of SIGIR Conferences: What Have We Been Doing for the Last Quarter of a Century ?