Understanding Hierarchies in Computer Science Conferences

Aditya Garg
Aashutosh Trivedi
Ansa Mary Ephraim

Motivation

- Different conferences are rated differently
- We attempt to understand the structure of the community associated with conferences at different tiers of rating
- We find various trends of
    - Similarity between venues/conferences
    - Measure of research carried out
    - Measure of people getting into research
- These metrics are evaluated year on year as well as tier wise

The data

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

The data

<u>Initial hurdles</u> – 5 datasets but still couldn't get the data we were looking for

•SNAP database : very small dataset, for Physic papers not in interest set.

•Google Scholar : Used *Scrapy* to start building a web scraper. Involved sharp learning curve, limit on downloads. Required restarts, cleaning and manipulation.

•DBLP : 1.3 Gb of clean annotated metadata for papers published in Computer Science. Missing citation and keyword information.

•CiteseerX from the website : <xml> data, no citation information present.

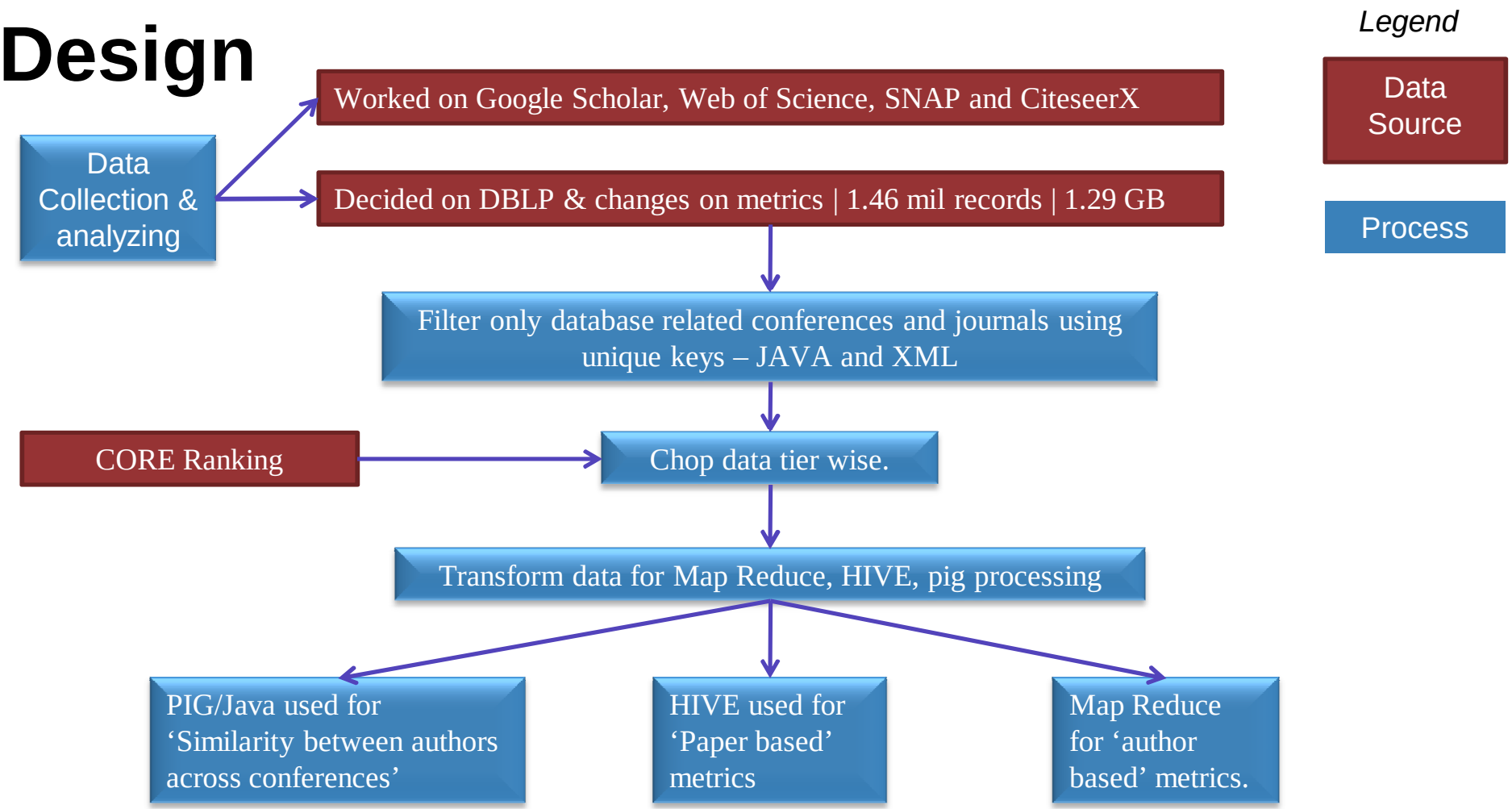
The data

Initial hurdles – 5 datasets but still couldn't get the data we were looking for

CiteseerX directly from Penn State :

- Over 40 Gb of data in XML and over 60 Gb MYSQL dump stored in Amazon S3.
- Stored on HPC.
- No mySQL on HPC.
- Tried parsing using Perl, Python and Bash scripts one line at a time.
- *Data not indexed.*
- Cleaning and clustering required due to data being automatically scraped by a crawler.

Design

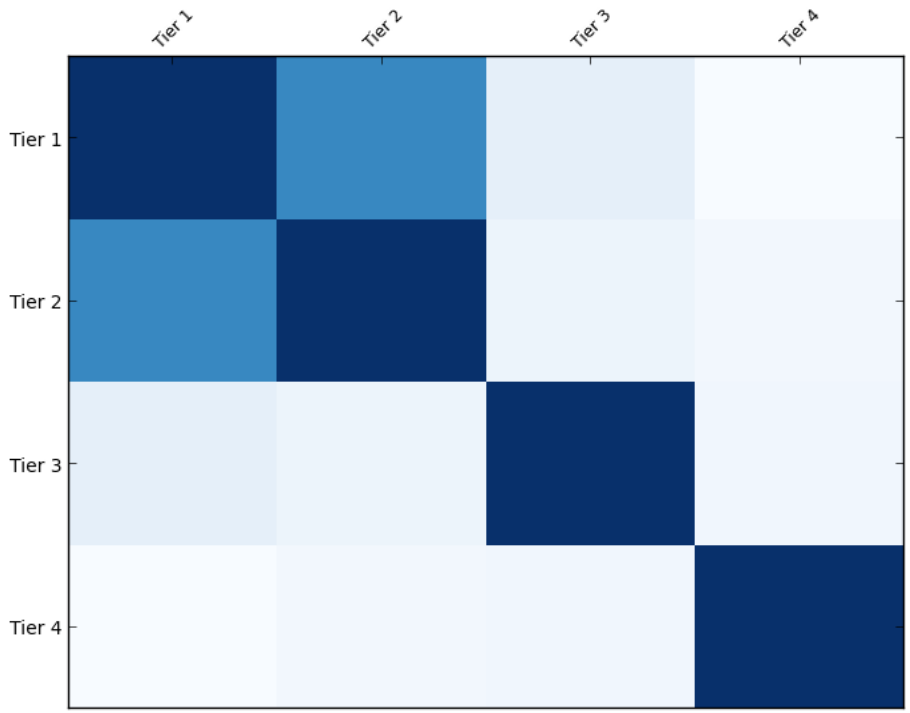

Legend
Data Source
Process
Data Collection & analyzing
Worked on Google Scholar, Web of Science, SNAP and CiteseerX
Decided on DBLP & changes on metrics | 1.46 mil records | 1.29 GB
Filter only database related conferences and journals using unique keys – JAVA and XML
CORE Ranking
Chop data tier wise.
Transform data for Map Reduce, HIVE, pig processing
PIG/Java used for 'Similarity between authors across conferences'
HIVE used for 'Paper based' metrics
Map Reduce for 'author based' metrics.

Design

Tiers

- *CORE* – Computing Research & Education – computer science rankings
- We decided to split the data into 4 different tiers as per ranking of the conference/journal.
- This gave us a better picture about the kind/measure of research at what level.
    - Tier 1 - flagship conferences
    - Tier 2 – excellent conferences
    - Tier 3 – good conferences
    - Tier 4 – other honorable conferences

Similarity between authors across conferences

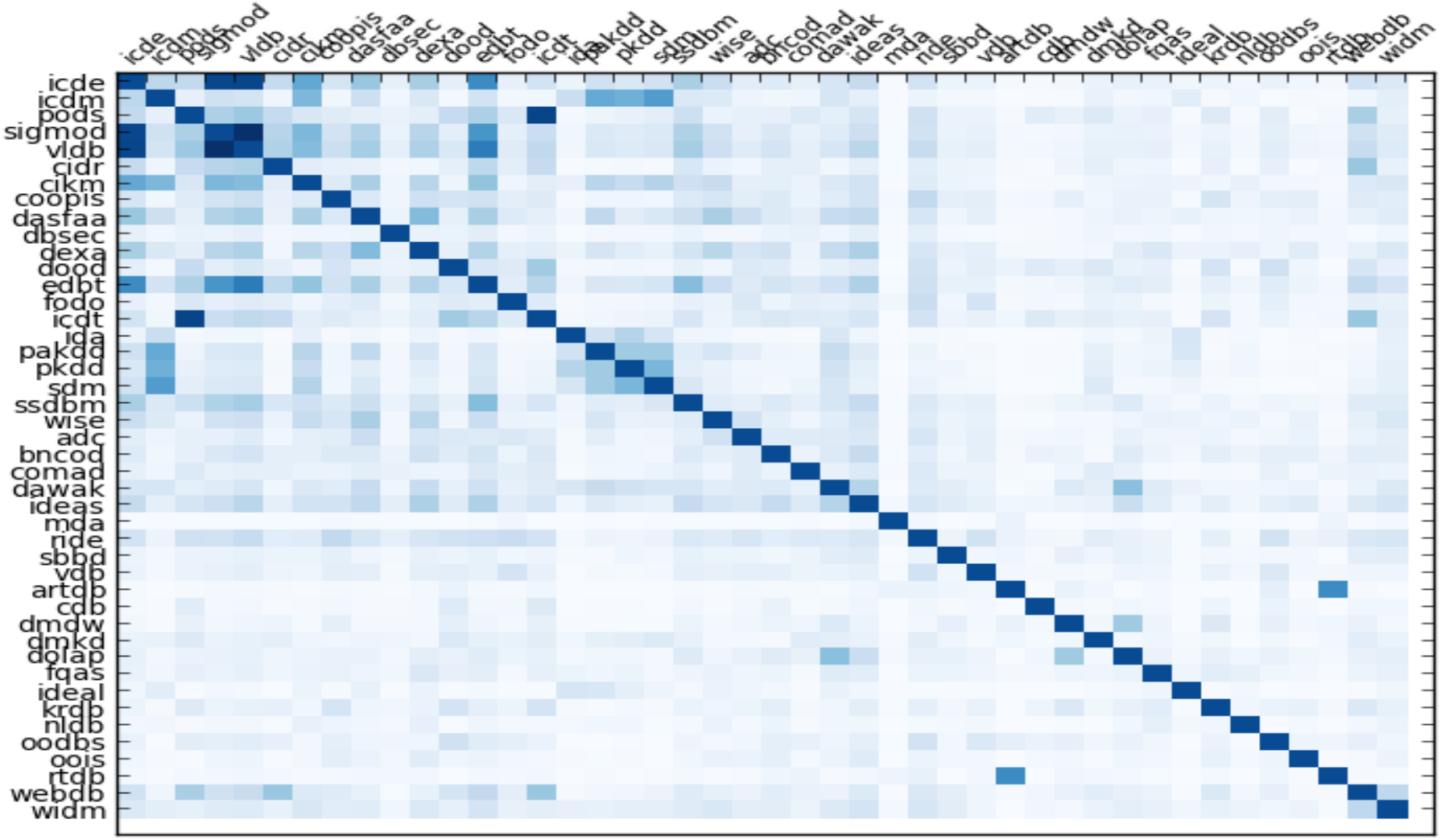


Focused on the Database community in DBLP

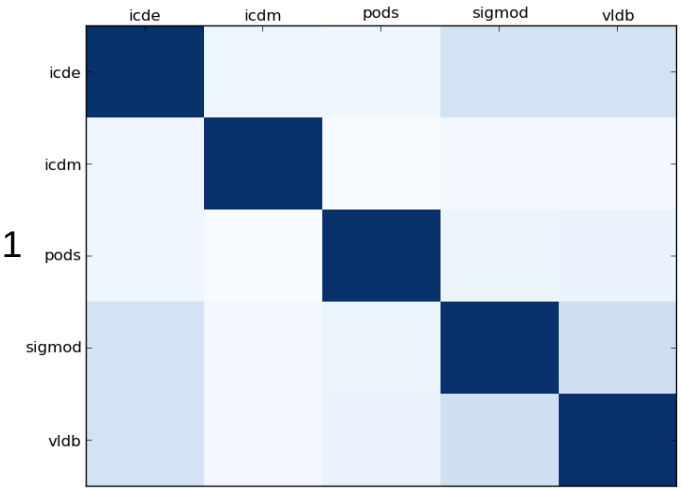
Similarity measured using Jaccard distance on the author sets

#1 Hive partitioning by conference.
2 Hive streaming with Python map reduce functions.
3 Pig followed by Java

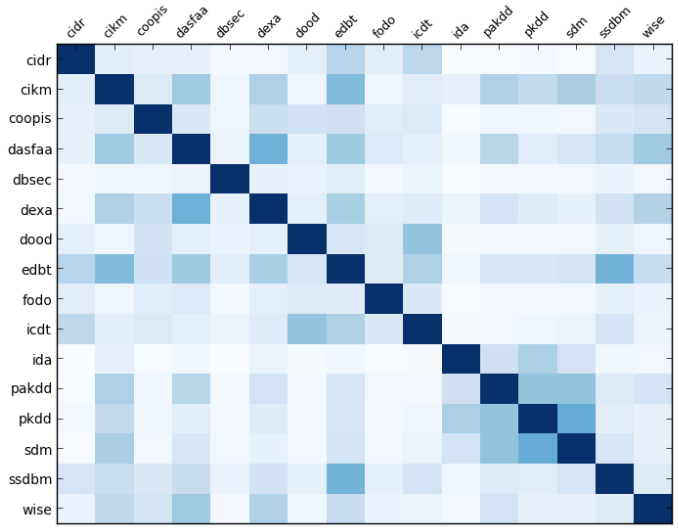
Jaccard distance = $A \cap B / A \cup B$


icde
icdm
pods
sigmod
vldb
cidr
cikm
coopis
dasfaa
dbsec
dexa
dood
edbt
fodo
icdt
ida
pakdd
pkdd
sdm
ssdbm
wise
adc
bncod
comad
dawak
ideas
mda
nde
sbbd
vdb
artdb
cdb
dmdw
dmkd
dolap
fqas
ideal
krdb
nldb
oodbs
oois
rtdb
webdb
widm

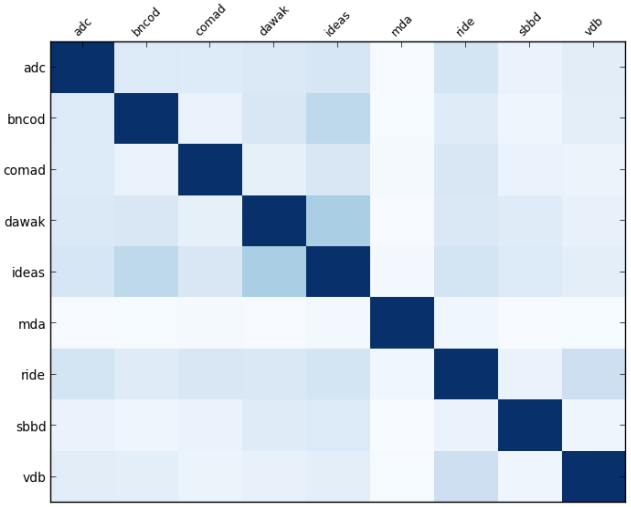
Tier 1



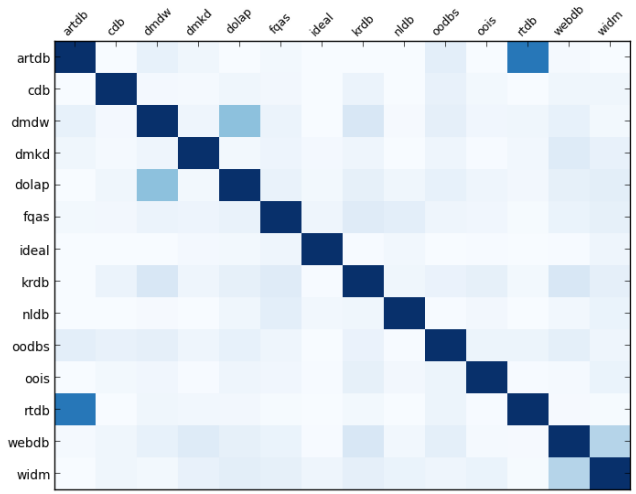
Tier 2



Tier 3



Tier 4



Author based statistics

Hadoop MR

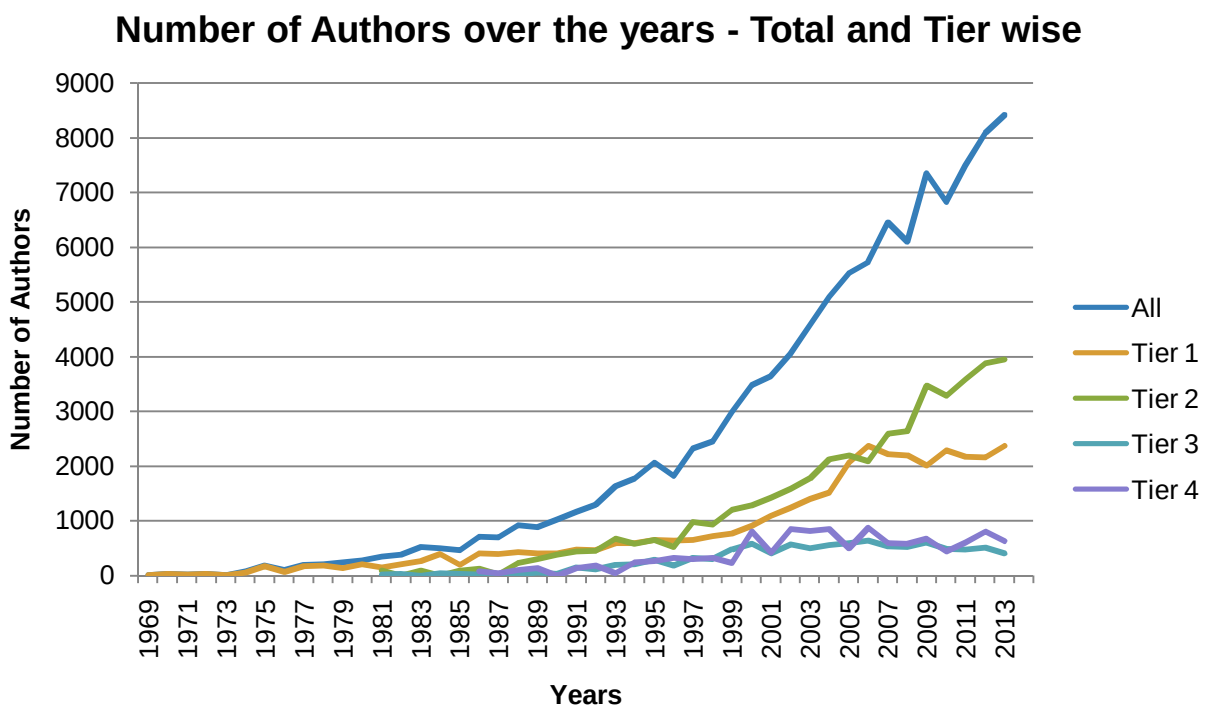
- Selected <xml> data was transformed and flattened to get each record into a single line for easier implementation of MR.
- E.g.

Inproceedings author:Roberto Brunelli|Ornella Mich title:Efficient Image Retrieval by Examples. year:2000 pages:145-162 crossref:conf/vdb/2000 booktitle:VDB url:db/conf/vdb/vdb2000.html#BrunelliM00

Author based statistics

•More people getting into research

•Healthy increase of top tier conferences instead of lower tier

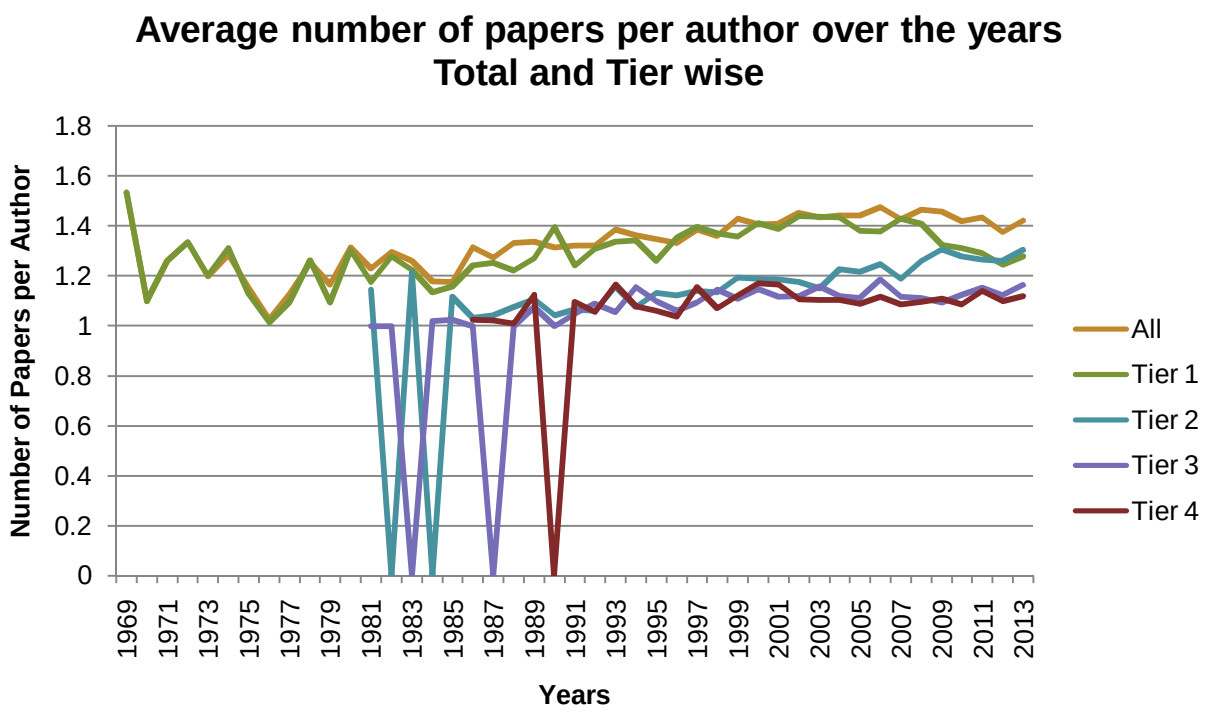


Author based statistics

•Steady amount of research per author, good as number of authors increasing

•Top tier conferences performing better than lower tier

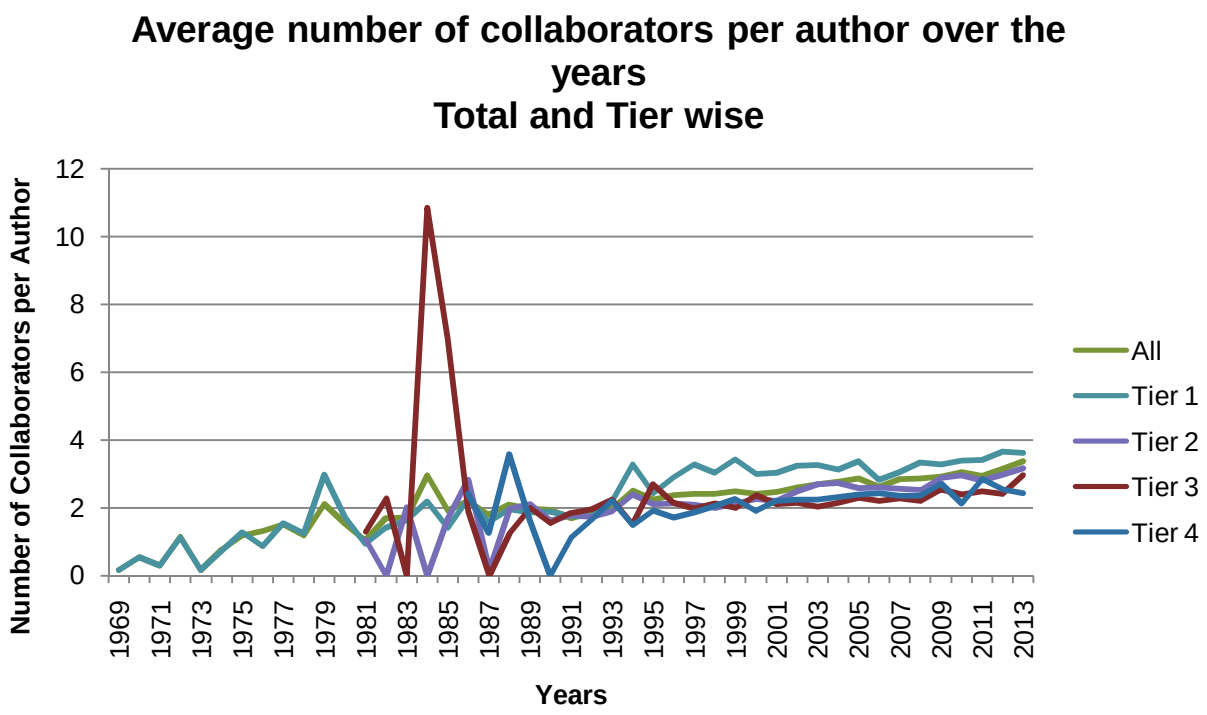
•Lotka Law of 60%



Author based statistics

•Steady number of collaborators

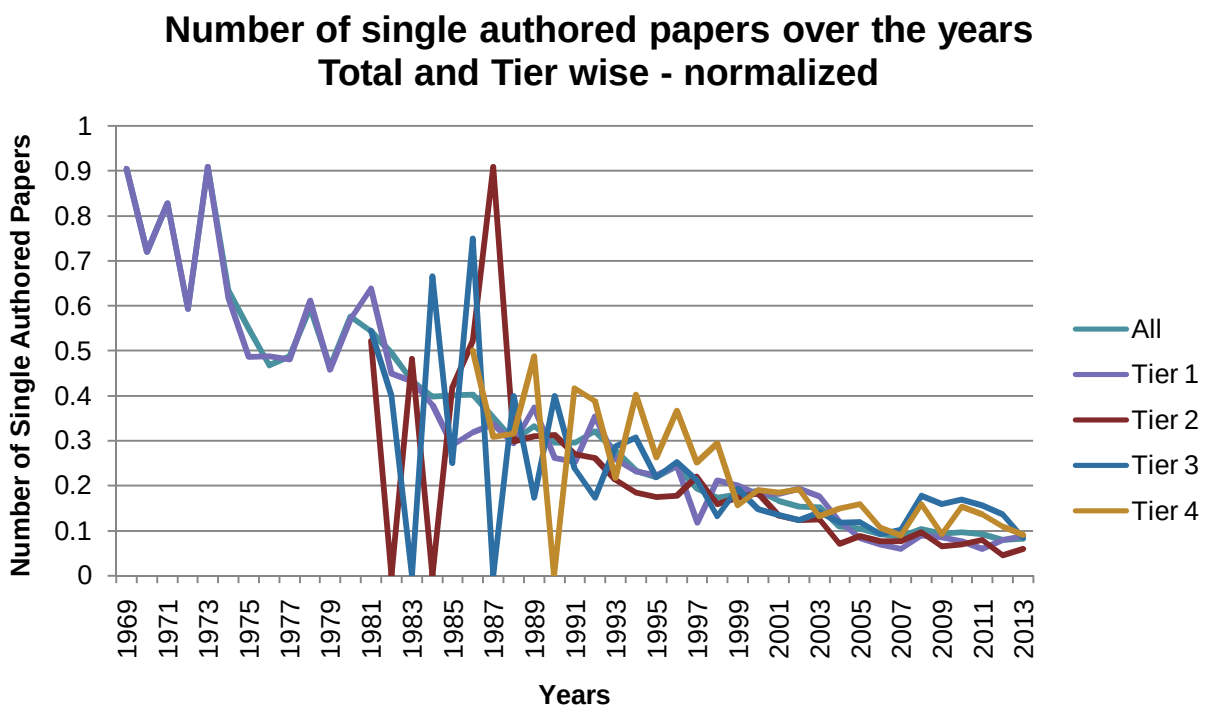
•Many conferences also have limit on maximum number of collaborators



Author based statistics

•Number of single authored papers reducing over time. Better networking & collaboration opportunity.

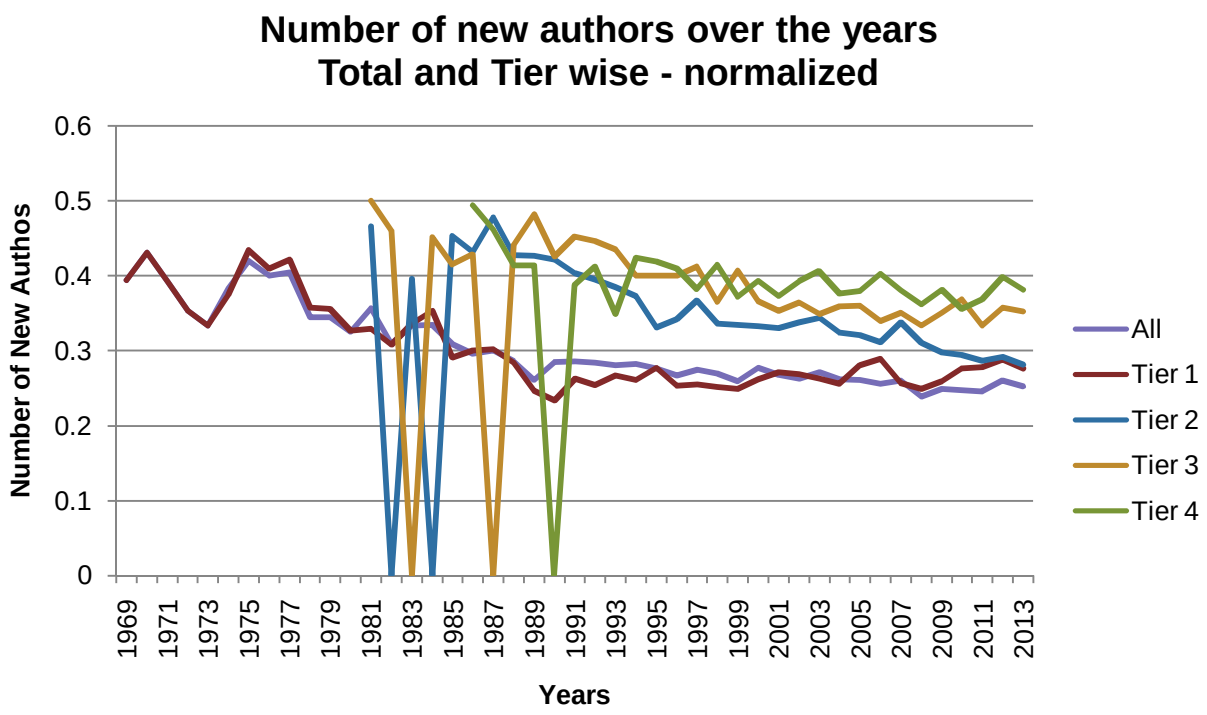
•Good as collaborators allow different perspective and usually allow deeper research



Author based statistics

•New people coming into research more or less stable

•Higher in lower ranked conferences. Makes sense since usually people will enter with lower ranked conferences



Paper based statistics

Hive:

- As only certain fields were required for Hive analytics, the files were transformed into a tab delimited file having the following format

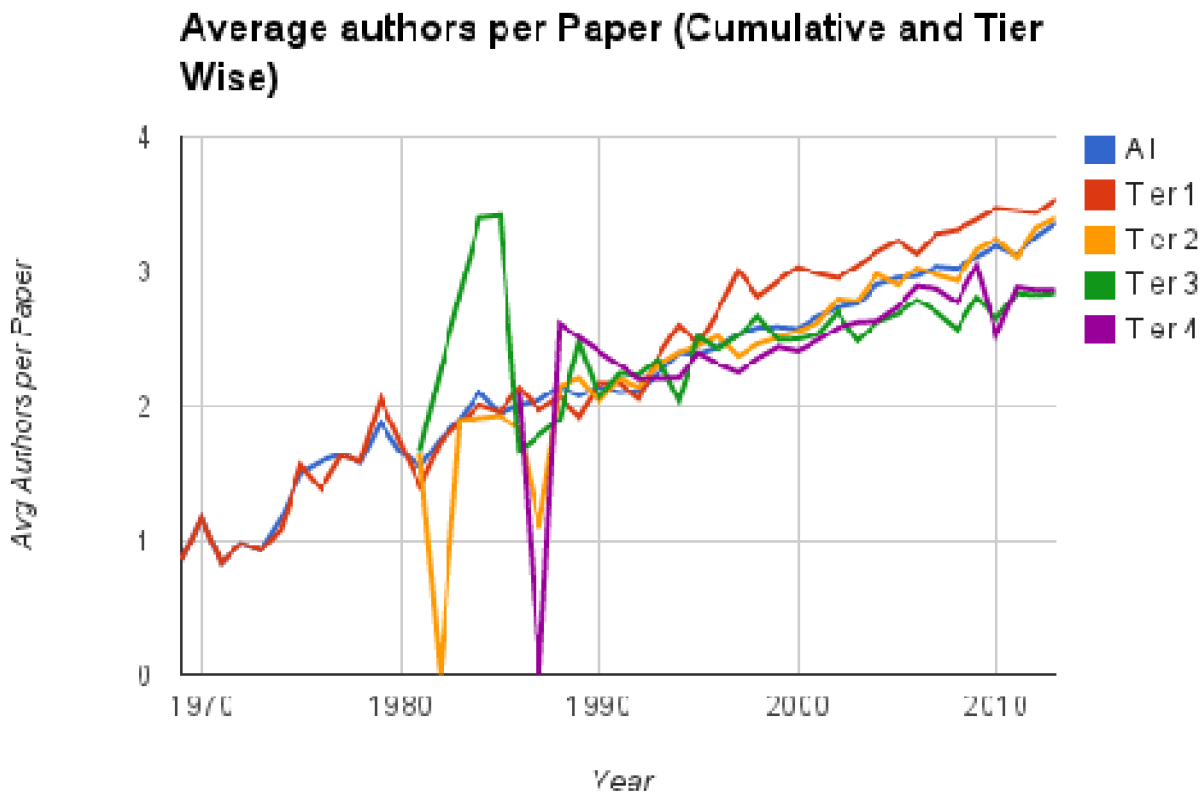
  key year conference authors-array

- Then these were put in the tables through Hive
- The configured file was also split according to the Tiers and the analytics were also found on the tier files

Paper based statistics

•Steady increase shows increasing trend of collaboration amongst authors.

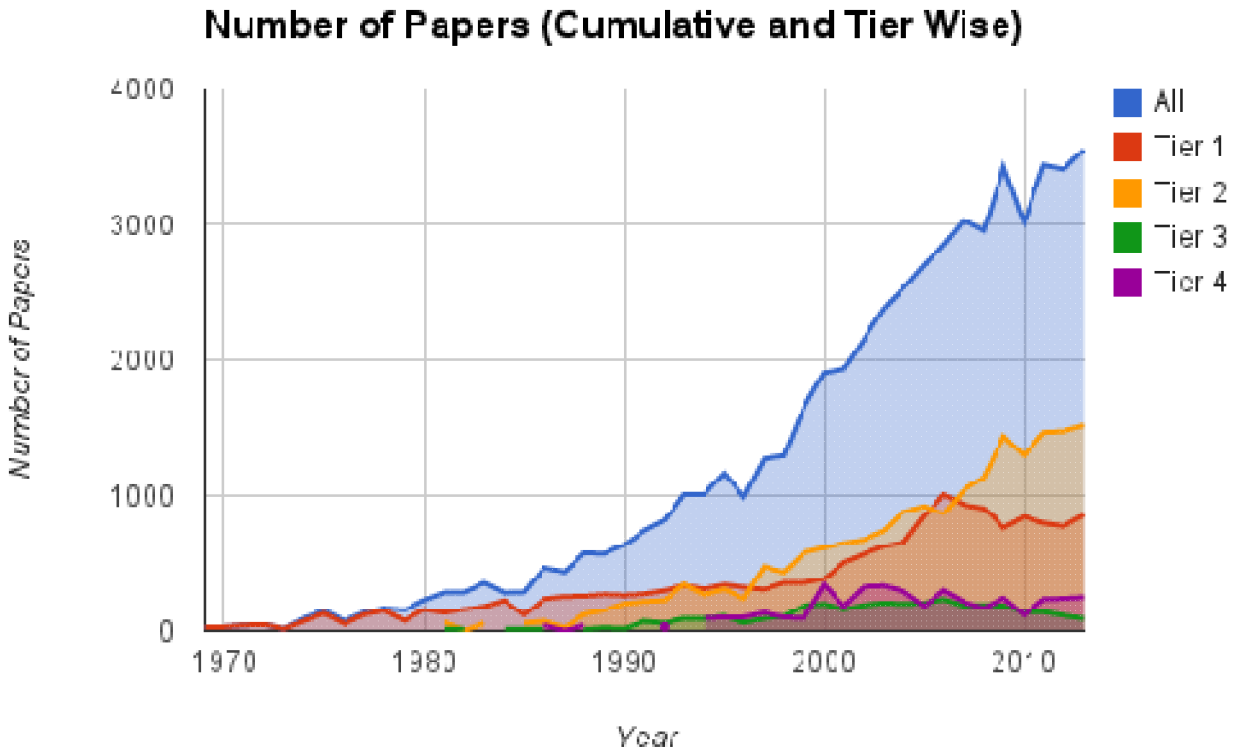
•Publish or Perish Theory



Paper based statistics

•Healthy and steady increase in number of papers published

•More papers published in top tier conferences.



Future Work

- Our data-set and associated techniques can also be used for further predictions like
  - What makes a paper acceptable
  - What gets a paper cited more often than others
  - Do people who get published, work in groups or alone?
  - What are the likely venues to publish given the authors one has worked with
  - Keyword analysis to identify what gets a paper cited more often
  - Structure of collaboration network/degrees of separation

References

- T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.
- J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In proceedings of 6th Symposium on Operating Systems Design and Implemenation, 2004.
- 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.
- Mario A. Nascimento, Jorg Sander and Jeffrey Pound. Analysis of SIGMOD's Co-Authorship Graph
- Vladimir Batagelj and Andrew Mrvar. Some Analyses of Erdos˝ Collaboration Graph.
- 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.
- Yi Han, Bin Zhou, Jian Pei, Yan Jia. Understanding Importance of Collaborations in Co-authorship Networks: A Supportiveness Analysis Approach.
- M.E.J. Newman. The structure of scientific collaboration networks.
- Alan F. Smeaton, Gary Keogh, Cathal Gurrin, Kieran McDonald and Tom Sodring. Analysis of Papers from Twenty-Five Years of SIGIR Conferences: What Have We Been Doing for the Last Quarter of a Century ?

Special Thanks

- Professor McIntosh for her continuous support in providing guidance and data options.

- Penn State University