Modeling and Mining of Networked Information Spaces

http://www.mathstat.dal.ca/~mominis

The Opportunity

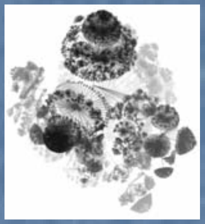
- *Show me your friends and I will tell you what you are.*
- Google: the more incoming links, the more worthy a web page is (Page Rank).
- Citation Graph (network)
- Email Graph (network)
- Phone Call Graph (network)
- Collaboration graph (network)
- **All analysis is based on the network links, not the content !!!**

Citation indexing of the scientific literature

- Used to be done manually, updated periodically
- Citeseer automated the process (in CSCI)
  - Focused crawler collects papers off the Web
  - Intelligent document processing extracts title, authors, abstract, references
  - Builds citation graph
    - nodes are papers
    - directed links are references/citations
  - Analyses citation graph

Citation graphs

http://www.informatik.tu-cottbus.de/~an/GD/ERLinLog/GDCitations.html

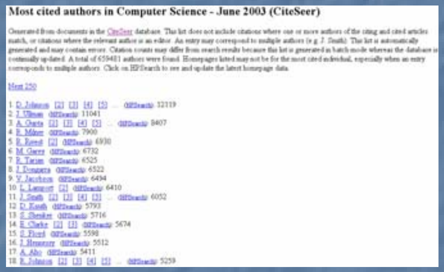


http://wwwcsif.cs.ucdavis.edu/~mcpherso/largegraph.html

How to find "hot spots" in dynamic graphs?

Open question

- Things to try:
  - Features of node neighbourhoods (as a time series)
  - Feature evolution
    - Over time
    - As neighbourhood size changes
- Features used to classify nodes

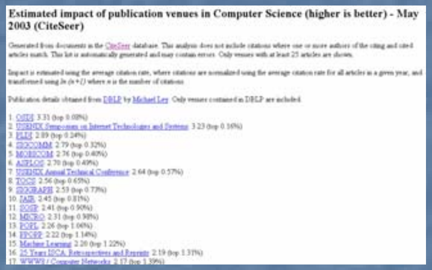


Most cited authors in Computer Science - June 2003 (CiteSeer)

Generated from documents in the CiteSeer database. This list does not include citations where one or more authors of the citing and cited articles match, or citations where the relevant author is an editor. An entry may correspond to multiple authors (e.g. *J. Smith*). This list is automatically generated and may contain errors. Citation counts may differ from search results because this list is generated in batch mode whereas the database is continually updated. A total of 659481 authors were found. Homepages listed may not be for the most cited individual, especially when an entry corresponds to multiple authors. Click on HPSearch to see and update the latest homepage data.

Next 250

1. D. Johnson [2] [3] [4] [5] ... (HPSearch) 32319
2. J. Ullman (HPSearch) 11041
3. A. Gupta [2] [3] [4] [5] ... (HPSearch) 8407
4. R. Milner (HPSearch) 7900
5. R. Reed [2] (HPSearch) 6930
6. M. Garey (HPSearch) 6732
7. R. Tarjan (HPSearch) 6525
8. J. Dongarra (HPSearch) 6522
9. V. Jacobson (HPSearch) 6494
10. L. Lamport [2] (HPSearch) 6410
11. J. Smith [2] [3] [4] [5] ... (HPSearch) 6052
12. D. Knuth (HPSearch) 5793
13. S. Shenker (HPSearch) 5716
14. E. Clarke [2] [3] (HPSearch) 5674
15. S. Floyd (HPSearch) 5598
16. J. Hennessy (HPSearch) 5512
17. A. Aho (HPSearch) 5411
18. R. Johnson [2] [3] [4] [5] ... (HPSearch) 5259



Estimated impact of publication venues in Computer Science (higher is better) - May 2003 (CiteSeer)

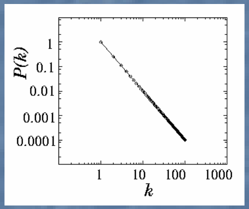
Generated from documents in the CiteSeer database. This analysis does not include citations where one or more authors of the citing and cited articles match. This list is automatically generated and may contain errors. Only venues with at least 25 articles are shown.

Impact is estimated using the average citation rate, where citations are normalized using the average citation rate for all articles in a given year, and transformed using *ln (n+1)* where *n* is the number of citations.

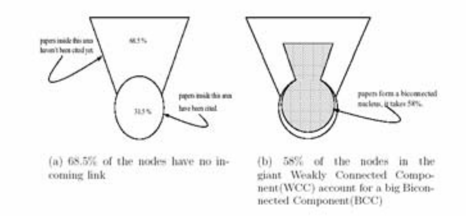
Publication details obtained from DBLP by Michael Ley. Only venues contained in DBLP are included.

1. OSDI 3.31 (top 0.08%)
2. USENIX Symposium on Internet Technologies and Systems 3.23 (top 0.16%)
3. PLDI 2.89 (top 0.24%)
4. SIGCOMM 2.79 (top 0.32%)
5. MOBICOM 2.76 (top 0.40%)
6. ASPLOS 2.70 (top 0.49%)
7. USENIX Annual Technical Conference 2.64 (top 0.57%)
8. TOCS 2.56 (top 0.65%)
9. SIGGRAPH 2.53 (top 0.73%)
10. JAIR 2.45 (top 0.81%)
11. SOSP 2.41 (top 0.90%)
12. MICRO 2.31 (top 0.98%)
13. POPL 2.26 (top 1.06%)
14. PPOPP 2.22 (top 1.14%)
15. Machine Learning 2.20 (top 1.22%)
16. 25 Years ISCA: Retrospectives and Reprints 2.19 (top 1.31%)
17. WWW8 / Computer Networks 2.17 (top 1.39%)

Power Law Distribution



Connectivity of the Citation graph

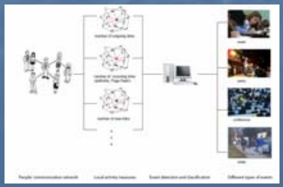

papers inside this area haven't been cited yet.
68.5 %
31.5 %
papers inside this area have been cited.
papers form a biconnected nucleus, it takes 58%.

(a) 68.5% of the nodes have no incoming link

(b) 58% of the nodes in the giant Weakly Connected Component(WCC) account for a big Biconnected Component(BCC)

Email graph of Dal Computer Science

- Classify nodes into
  - faculty,
  - students,
  - staff, or
  - mailing lists
- Detect events
  - exams,
  - Study breaks,
  - open house,
  - local conference


People communication network
Local activity measures
Event detection and classification
Different types of events

Identifying Interesting Regions

- Real world graphs are often too big to visualize.
- We need to extract features from these graphs to describe and compare them.
- Highly connected regions of a graph can viewed as forming a graph structure.
- We use this structure to characterize a large graph.

An induced subgraph of the collaboration graph (with Erdös number at most 2).
Made by Fan Chung Graham and Lincoln Lu in 2002.

A collaboration graph

Citation graph

- In-degree follows power law
  - Fraction of web pages having *k* incoming links is proportional to *1 / k*²
- Tightly connected **even** after removing high hub and authority articles.
- Bridges between subareas offer insight

Dynamic graphs

- Reflect evolution in social networks over time
  - Email graph
  - Phone call graph
  - Citation graph
  - Coauthorship graph
- Abnormal patterns may signify
  - Unusual event
  - Fraud
  - Terrorist activity

K-cores

- *k*-core is the subgraph generated by recursively removing all nodes of degree less than *k*.
- Here the 1-core is the full graph.
- The 2-core is composed of the red and green nodes.
- The 3 and 4-cores consist of only the red nodes.