Agenda

1. About The Alpha Lab
2. What is “Big Scholarly Data”? 
3. Scholarly Data Analysis
4. Quantifying Scientific Impact
5. Identifying Academic Rising Stars
6. Academic Recommendation
**Why?**

Alpha has the meaning of first in Greek. We borrow this word to express the idea that we pursue being extraordinary not only in academic research, but also in fully exploiting the potential of ourselves. We value hard work and talents. We embrace the change and the differences.

**What?**

Our goal is to create innovation through conducting interdisciplinary, application-driven academic research. We are interested in a broad spectrum of cutting-edge research topics including social computing, computational social science, big data, and mobile social networks.

**Where?**

Full name of the Lab: The Alpha Lab @ Dalian University of Technology, China.
Address: School of Software, Dalian University of Technology, Development Zone, Dalian 116620 China.
URL (website): http://TheAlphaLab.org
Supervisors

Feng Xia
Professor

Jeremy Kong
Associate Professor

Ivan Lee
Distinguished Guest Professor

Alex Ning
Assistant Professor

Jenny Xu
Assistant Professor

Isabelle Liu
Assistant Professor
From 7 different countries

PhD, master, undergraduate students: 30+

We are family!
Research Interests

Computational Social Science
- social informatics, human dynamics, network science, social computing and networking

Big Data
- data analytics, scholarly big data, digital libraries, urban big data, data science, open data

Mobile Social Networks
- vehicular social networks, ad hoc social networks, socially-aware networking, intelligent systems
Computational Social Science

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, Marshall Van Alstyne

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven “computational social science” has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or reviewed by the long latency, verifying science—base—offering of individuals.

Harvard University, Cambridge, MA, USA. 2Massachusetts Institute of Technology, Cambridge, MA, USA. 3University of Michigan, Ann Arbor, MI, USA. 4New York University, New York, NY, USA. 5Northeastern University, Boston, MA, USA. 6Interdisciplinary Scientific Research, Seattle, WA, USA. 7Northwestern University, Evanston, IL, USA. 8University of California–San Diego, La Jolla, CA, USA. 9Columbia University, New York, NY, USA. 10Cornell University, Ithaca, NY, USA. 11Boston University, Boston, MA, USA. E-mail: david_lazer@harvard.edu. Complete affiliations are listed in the supporting online material.
Research Areas/Topics

Science of Success in Science

Scholarly Recommendation

Big Trajectory Data
Research Areas/Topics

Socially-Aware Networking

Science of Scientific Team Science

Scientific Collaboration Dynamics
What is “Big Scholarly Data”? 

Feng Xia, Wei Wang, Teshome Megersa Bekele, Huan Liu. 
Big Scholarly Data: A Survey 
IEEE Transactions on Big Data, accepted for publication 2016
Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time.
What is “Big Scholarly Data”? 

Big Scholarly Data is coined for the rapidly growing scholarly data, which contains information including millions of authors, papers, citations, figures, tables, as well as scholarly networks and digital libraries.
Big Scholarly Data is Heterogeneous
## Major Digital Libraries and Search Engines

<table>
<thead>
<tr>
<th>Name</th>
<th>Discipline</th>
<th>Description</th>
<th>Access</th>
<th>Reference Management</th>
<th>Provider</th>
<th>Search Engine/Digital Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACM Digital Library</td>
<td>Computing and Information Technology</td>
<td>Comprehensive collection of full-text articles and bibliographic records</td>
<td>Subscription</td>
<td>No</td>
<td>Association for Computing Machinery</td>
<td>Digital library</td>
</tr>
<tr>
<td>Arnetminer</td>
<td>Computer Science</td>
<td>Comprehensive search and mining services for researcher social networks</td>
<td>Free</td>
<td>No</td>
<td>Tsinghua University</td>
<td>Both</td>
</tr>
<tr>
<td>arXiv</td>
<td>Multidisciplinary</td>
<td>Highly-automated electronic archive and distribution server for research articles</td>
<td>Free</td>
<td>No</td>
<td>Cornell University</td>
<td>Both</td>
</tr>
<tr>
<td>CiteSeerX</td>
<td>Computer and Information Science</td>
<td>Evolving scientific literature digital library and search engine</td>
<td>Free</td>
<td>No</td>
<td>Pennsylvania State University</td>
<td>Both</td>
</tr>
<tr>
<td>DBLP</td>
<td>Computer Science</td>
<td>Open bibliographic information on computer science journals and proceedings</td>
<td>Free</td>
<td>No</td>
<td>University of Trier</td>
<td>Digital library</td>
</tr>
<tr>
<td>Google Scholar</td>
<td>Multidisciplinary</td>
<td>Indexing the full text or metadata of scholarly literature across disciplines</td>
<td>Free</td>
<td>No</td>
<td>Google</td>
<td>Search engine</td>
</tr>
</tbody>
</table>
## Major Digital Libraries and Search Engines

<table>
<thead>
<tr>
<th>Service</th>
<th>Domain</th>
<th>Description</th>
<th>Access Model</th>
<th>Subscription</th>
<th>Publisher</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEE Xplore</td>
<td>Computer Science, Engineering, Electronics</td>
<td>Online service used to index and search social networks</td>
<td>Subscription</td>
<td>No</td>
<td>IEEE Computer Society</td>
<td>Digital library</td>
</tr>
<tr>
<td>Mendeley</td>
<td>Multidisciplinary</td>
<td>Crowdsourced database of research documents</td>
<td>Free</td>
<td>Yes</td>
<td>Mendley</td>
<td>Search engine</td>
</tr>
<tr>
<td>Microsoft Academic Search</td>
<td>Multidisciplinary</td>
<td>Provides many innovative ways to explore scientific papers, conferences, journals, and authors.</td>
<td>Free</td>
<td>No</td>
<td>Microsoft Search Engines</td>
<td>Search engine</td>
</tr>
<tr>
<td>PubMed National</td>
<td>Medicine</td>
<td>Accessing primarily the MEDLINE database of references and abstracts on biomedical topics</td>
<td>Free</td>
<td>No</td>
<td>U.S. National Library of Medicine</td>
<td>Both</td>
</tr>
<tr>
<td>ScienceDirect</td>
<td>Multidisciplinary</td>
<td>A leading full-text scientific database offering journal articles and book chapters</td>
<td>Subscription</td>
<td>No</td>
<td>Elsevier</td>
<td>Digital library</td>
</tr>
<tr>
<td>Scopus</td>
<td>Multidisciplinary</td>
<td>A bibliographic database containing abstracts and citations for academic journal articles</td>
<td>Subscription</td>
<td>No</td>
<td>Elsevier</td>
<td>Digital library</td>
</tr>
<tr>
<td>Web of Knowledge</td>
<td>Multidisciplinary</td>
<td>An academic citation indexing and search service</td>
<td>Subscription</td>
<td>No</td>
<td>Thompson Reuters</td>
<td>Digital library</td>
</tr>
</tbody>
</table>
Open Access Scholarly Data

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Discipline</th>
<th>Size</th>
<th>Updated time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aminer</td>
<td>Computer Science</td>
<td>710MB</td>
<td>2013-02-26</td>
</tr>
<tr>
<td>APS</td>
<td>Physics</td>
<td>1.21GB</td>
<td>2014-07-21</td>
</tr>
<tr>
<td>DBLP</td>
<td>Computer Science</td>
<td>297MB</td>
<td>2015-09-05</td>
</tr>
<tr>
<td>MAG</td>
<td>Multidisciplinary</td>
<td>29.8GB</td>
<td>2015-08-31</td>
</tr>
</tbody>
</table>

Affiliations.txt
1. Affiliation ID
2. Affiliation name

Authors.txt
1. Author ID
2. Author name

FieldsOfStudy.txt
1. Field of study ID
2. Field of study name

Conferences.txt
1. Conference ID
2. Short name (abbreviation)
3. Full name
4. Location
5. Official conference URL
6. Conference start date
7. Conference end date
8. Conference abstract registration date
9. Conference submission deadline date
10. Conference notification due date
11. Conference final version due date

PaperAuthorAffiliations.txt
1. Paper ID
2. Author ID
3. Affiliation ID
4. Original affiliation name
5. Normalized affiliation name
6. Author sequence number

PaperUrls.txt
1. Paper ID
2. URL

PaperKeywords.txt
1. Paper ID
2. Keyword name
3. Field of study ID mapped to keyword
Scholarly Data Collection

- **Database**
  - Digital library: DBLP, Open access library

- **Web crawling and Document extraction**
  - Crawler
    - Downloader, Queue, Scheduler
  - Document Extracting/Harvesting

- **PDF**
  - Information Extractor:
    - Metadata i.e. Citations, abstracts, figures..

- **Filtering and Categorizing**
  - Link metadata, author disambiguation, remove duplication etc.

- **Index**
  - Store
  - Repositories

- **Data Discovery and Sharing Interface**
  - Data analysis

- **WWW:** scholar homepage; Academic social network

- **For example, DBLP can be used for name disambiguation**
Exploring Big Scholarly Data

Data Collection
- Academic social networks
  - academia.edu
  - Mendeley
  - LinkedIn
  - ResearchGate
- Academic search engine and digital library
  - dblp
  - Computer science bibliography
  - Microsoft Academic Search
  - Google Scholar

Raw Data
- Author profiles
- Publications
- Citations
- Figures
- Tables

Data Storage
- Data Sharing/Indexing

Data Analysis
- Statistical analysis
- Social network analysis
- Content analysis

Application
- Scientific impact evaluation
- Academic recommendation
- Expert finding
- Research trend prediction
Academic Social Networks

Authors

Coauthor Network

Citation Network

Articles

Bibliographic Coupling Network

Co-citation Network

Co-word Network
# Academic Social Network Analysis Tools

<table>
<thead>
<tr>
<th>Software</th>
<th>Platforms</th>
<th>Language</th>
<th>Access</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>CiteSpace</td>
<td>Windows/iOS</td>
<td>Java</td>
<td>Free</td>
<td>Visualizing and analyzing trends and patterns in scientific literature; knowledge domain visualization</td>
</tr>
<tr>
<td>Gephi</td>
<td>Windows/Linux/iOS</td>
<td>Java</td>
<td>Free</td>
<td>Exploratory Data Analysis; Social Network Analysis; Link Analysis</td>
</tr>
<tr>
<td>igraph</td>
<td>Windows/iOS</td>
<td>C/R/Python/Perl</td>
<td>Free</td>
<td>A collection of network analysis tools with the emphasis on efficiency, portability and ease of use</td>
</tr>
<tr>
<td>NetworkX</td>
<td>Windows/iOS</td>
<td>Python</td>
<td>Free</td>
<td>Creation, manipulation, and investigation of the structures, dynamics, and functions of complex networks</td>
</tr>
<tr>
<td>Pajek</td>
<td>Windows/iOS</td>
<td>C/R</td>
<td>Free</td>
<td>Analysis and visualization of large networks having some thousands or even millions of vertices</td>
</tr>
</tbody>
</table>

**Nodes:**
- Average path length
- Clustering coefficient
- Closeness centrality
- Degree centrality
- Betweenness centrality
Key Techniques

Statistical Techniques
- regression models
- variable selection
- principal component analysis
- factor analysis
- cluster analysis

Network Science
- degree distribution
- small-world property
- preferential attachment

Data Mining
- discovering meaning patterns in scholarly networks will lead to some advantages.

Machine Learning
- supervised learning
- unsupervised learning
- reinforcement learning
Potentials

- Researchers/Scholars
- Research Institutions
- Governments (Funding Agencies)
- Other Users? (Companies?)
Scholarly Data (Statistical) Analysis
As a generalization of article level metrics, altmetrics can assess the popularity or social impact of publications based on data collected by social media platforms.

Compared with traditional citation-based metrics, altmetrics can reduce the delay for accumulation and cover new forms of scholarly content (e.g., datasets, software, and research blogs) to achieve broader, more diversiform and rapid impact analysis.
In this case, we assess the *Nature* publications over the period between January 2010 and June 2015 based on altmetrics.

- Based on a comprehensive scientific magazine *Nature*
- Consider the impact of publication year and discipline for the analysis
- Relatively long time (2010-2015)
- Consider the Twitter user type at the first time
We analyze the distribution of academic information about *Nature* articles on Twitter and Facebook, to evaluate the validity of Twitter and Facebook as data sources for altmetrics.

1. Distribution

   The mention rate is used to examine the impact of a *Nature* article on a social media platform.

3. Mention Rate

The coverage is used to evaluate the concern degree of social users on a *Nature* article and the development of the social media platform on the academic field.

2. Coverage

4. Relationship

   We also analyze the relationship between tweets and citations for *Nature* publications to determine whether both types of metrics measure similar concepts.
We can find that both Twitter users and Facebook users are interested in a few *Nature* articles published in 2010. As Twitter and Facebook evolve, social users increasingly focus on the scholarly documents, and thus Twitter and Facebook coverages show an increasing trend over the publication time. **Twitter develops more rapidly than Facebook for the academic field.**

**Twitter coverage for biology sciences is significantly higher than other disciplines** and Twitter coverage for other three disciplines show a similar lower growth trends. For *Nature* articles published after 2012, Twitter coverage for all disciplines approaches 100 percent.
Data Analysis

Facebook coverage by publication year and discipline

Twitter coverage by user type and discipline

Compared with Twitter coverage, the Facebook coverage differences among distinct disciplines are relatively larger. For the articles which are not published in 2014, the Facebook has a lower coverage for chemical sciences than other disciplines and a relatively high coverage for biology sciences and earth & environment sciences.

For all disciplines, members of the public have the highest concern degree. Practitioners have the lowest concern degree. Biology sciences draw more concern degree of four user types.
There is a **continuous growth** for both Twitter and Facebook mention rates because of the development of social media platforms. **Compared with Twitter, the growth of Facebook mention rate is relatively slow.**

For all disciplines, **there is a highest impact on members of the public.** For members of the public, scientists and science communicators, the impact of the articles about chemical sciences is much lower than the articles of other three disciplines. Moreover, for all disciplines, **there is a relatively small impact on practitioners and science communicators.**
Twitter and Facebook mention rate by publication year and discipline

There is an **ascending trend** of both Twitter and Facebook mention rates for articles about all disciplines. For all articles published from 2010 to 2015, we also can see that the articles about **biology sciences and earth & environment sciences** have higher Twitter and Facebook mention rates than the other two disciplines.
The development of social media platforms makes people more interested in academic information.

Twitter users have a higher and faster-growing concern degree on the Nature articles compared with Facebook users.

Nature articles have higher and faster-growing impact on Twitter than on Facebook.

Although tweets and citations are somewhat related, they mostly measure different type of impact.

The correlation between tweets and citations for Nature articles is positive and appears quite sensitive to the publication date, discipline and Twitter user type.
Quantifying the Impact of Scholarly Papers
Identifying Anomalous Citations for Objective Evaluation of Scholarly Article Impact

Xiaomei Bai, Feng Xia, Ivan Lee, Jun Zhang, Zhaolong Ning. PLoS ONE, 2016, DOI: 10.1371/journal.pone.0162364
Scholarly impact is one of the strongest currencies in the academia.

Citation is generally used to measure scholarly impact.

Quantifying the impact of scholarly papers is a cornerstone for evaluating the impact of scholars, journals, institutions, and other scholarly entities.
Illustrative example of COI relationship between different authors.

- **limitations**: citation-based evaluation approaches have been widely used. However, these approaches face limitations, e.g., in identifying anomalous citations.

- **Solution**: Conflict of Interest (COI) relationships between citing and cited papers are taken into consideration in order to better quantify the impact of scholarly papers.
We define COI and suspected COI relationship between citing and cited authors.

Based on PageRank and HITS algorithms, we leverage COI and suspected COI relationships to weight paper citation network.

We use collaboration times, the time span of collaboration, citing times, and the time span of citing to weaken the COI and suspected COI relationships.

Recommendation Intensity is used to evaluate the performance of our method.
**Experiments**

**Dataset**: APS dataset contains the papers from 1970 to 2013. Each article includes its title, DOI, author(s), publication date, affiliation(s) and publisher.

We analyze the citation of all the papers from co-authors and the same institution. The result shows that the anomalous citations are ubiquitous.

Ten most cited articles in the PhysRevC dataset.
Results

The results confirm that COIRank outperforms CAJTRank in terms of RI.

The main reason behind is shown as follows: On the one hand, the COIRank algorithm can benefit from weighted PageRank and HITS algorithms; On the other hand, capturing the dynamic evolutionary nature of citation networks is useful for rank calculation.

Comparing the results of time-weighted method with the corresponding no time-weighted method, it proves that time-weighted method can give a further improvement to the result.

Illustration of the probability of Recommendation Intensity based on different algorithms.
Identifying Academic Rising Stars
CocaRank: A Collaboration Caliber-based Method for Finding Academic Rising Stars

Jun Zhang, Feng Xia, Wei Wang, Xiaomei Bai, Shuo Yu, Teshome Megersa Bekele and Zhong Peng. WWW/BigScholar, 2016
**Significance:**
Finding rising stars can shed light on a lot of scientific questions, such as providing candidates for peer reviews, searching potential cooperators, building basis for foundation and awards application etc.

**Limitations of existing solutions:**
e.g. in considering scholars’ capacities of collaboration, the effect of scholars’ different academic ages, and finding the appropriate time interval for the evaluation of rising stars.
• We propose a novel indicator, the collaboration caliber, to capture a scholar’s ability of cooperating with other people.
• We use the time of the first paper being published as the start line of rising star’s evaluation to avoid the unfair situations for younger researchers when they are compared to seniors.
• In order to find the appropriate time interval for the evaluation of rising stars, we select several time intervals for comparison.
Researchers who are presently not outstanding among peers or with low research profile, but will grow into influential or authoritative scholars in the future.

**What is rising star?**
- We first use the concept of entropy to calculate the caliber of collaboration, for each scholar.
- Then calculate the impact of scholars in heterogeneous academic networks.
- At last, we combine the values of the above two steps and compute the final score for each scholar.

**CocaRank method**
- It consists of three sub-networks, e.g. paper citation network, paper-author network and paper-journal network.

**Evaluation**
- Citation counts and Spearman Correlation Coefficient are used to evaluate the performance of our method.
Experiments

**Dataset:** APS dataset contains the papers from 1970 to 2013. Each article includes its title, DOI, author(s), publication date, affiliation(s) and publisher.
- We choose researchers who published their first articles in 1993, and their academic career is not ended until 2013.
- We choose 4 time intervals, which are 3, 5, 7 and 10 years.

**Benchmark:**
- CocaRank-PaHit: is the variant of our proposed method CocaRank, which only considers the combination results under heterogeneous networks to evaluate the impacts of rising stars.
- StarRank: is introduced in [1], and we choose it as the benchmark for comparison.

Results

- The average citation counts by our proposed CocaRank is the highest among the three methods in the four time intervals.
- It is also observed that the growth rate of average citation counts is the highest between 3 and 5 years.
- In other words, the first 3 to 5 years is a very crucial stage in a scholar's whole scientific career, during which research capacities improve rapidly in this period.

Comparison of future citation counts in different time intervals
Results

- We also calculate the Spearman Correlation Coefficient to measure the correlation between the CitCounts and the above methods for comparison.
- The value of CocaRank is the highest among the three methods. The results indicate that our method has made an remarkable improvement as compared to other methods.

<table>
<thead>
<tr>
<th>Time Interval</th>
<th>CocaRank</th>
<th>CocaRank-PaHit</th>
<th>StarRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.76391</td>
<td>0.59248</td>
<td>-0.05716</td>
</tr>
<tr>
<td>5</td>
<td>0.84361</td>
<td>0.71729</td>
<td>-0.04962</td>
</tr>
<tr>
<td>7</td>
<td>0.76541</td>
<td>0.73836</td>
<td>0.34887</td>
</tr>
<tr>
<td>10</td>
<td>0.75789</td>
<td>0.69925</td>
<td>0.08722</td>
</tr>
</tbody>
</table>
Academic Recommendation
Socially-Aware Venue Recommendation for Conference Participants


IEEE UIC, 2013. (Best Paper Award)

Academic conferences and workshops do not just serve as platforms to present the research work of participants, but also aim to connect participants in the same domain and foster prospective collaborations. However, The plethora of talks and presentations in multiple and parallel tracks at academic conferences makes it difficult, especially for junior researchers to attend the right presentation sessions and collaborate socially with participants and potential researchers who have similar research interests.
We propose an innovative venue recommendation algorithm to enhance smart conference participation.

Social Aware Recommendation of Venues and Environments (SARVE), computes the Pearson Correlation and social characteristic information of conference participants.

SARVE further incorporates the current context of both the smart conference community and participants in order to model a recommendation process using distributed community detection.
1. SARVE computes the Pearson correlation and social ties of the user and all the presenters to ascertain high levels of similarity and ties strength between them.

2. SARVE further computes degree centrality of participant presenters to determine their popularity status/level at the smart conference.

3. SARVE integrates explicit contextual information of the user, presenters and community, in order to accordingly generate an effective social venue recommendation.
Experiments

We gathered data from 78 students of the School of Software, Dalian University of Technology, China. Explanation was given to students to select/annotate keywords of interest as well as contextual information (available time and present location) in relation to the simulated conference (ICWL 2012).

Details and components of ICWL 2012 dataset
Results

In terms of precision, both social context and social relations recommendations for SARVE were more precise and exact especially at higher recommendation values in accordance to the dataset.

In terms of recall, both social context recommendation and social relations recommendation for SARVE exhibited higher recall values and covered more useful items in accordance to the dataset.

Call for Papers: BigScholar @ WWW 2017

http://TheAlphaLab.org/BigScholar/ (submission due: 20 Jan)

The 4th WWW Workshop on Big Scholarly Data: Towards the Web of Scholars
Perth, Australia
April 4, 2017
THANK YOU!

http://TheAlphaLab.org