

AMiner II — Toward Understanding Big Scholar Data

Jie Tang

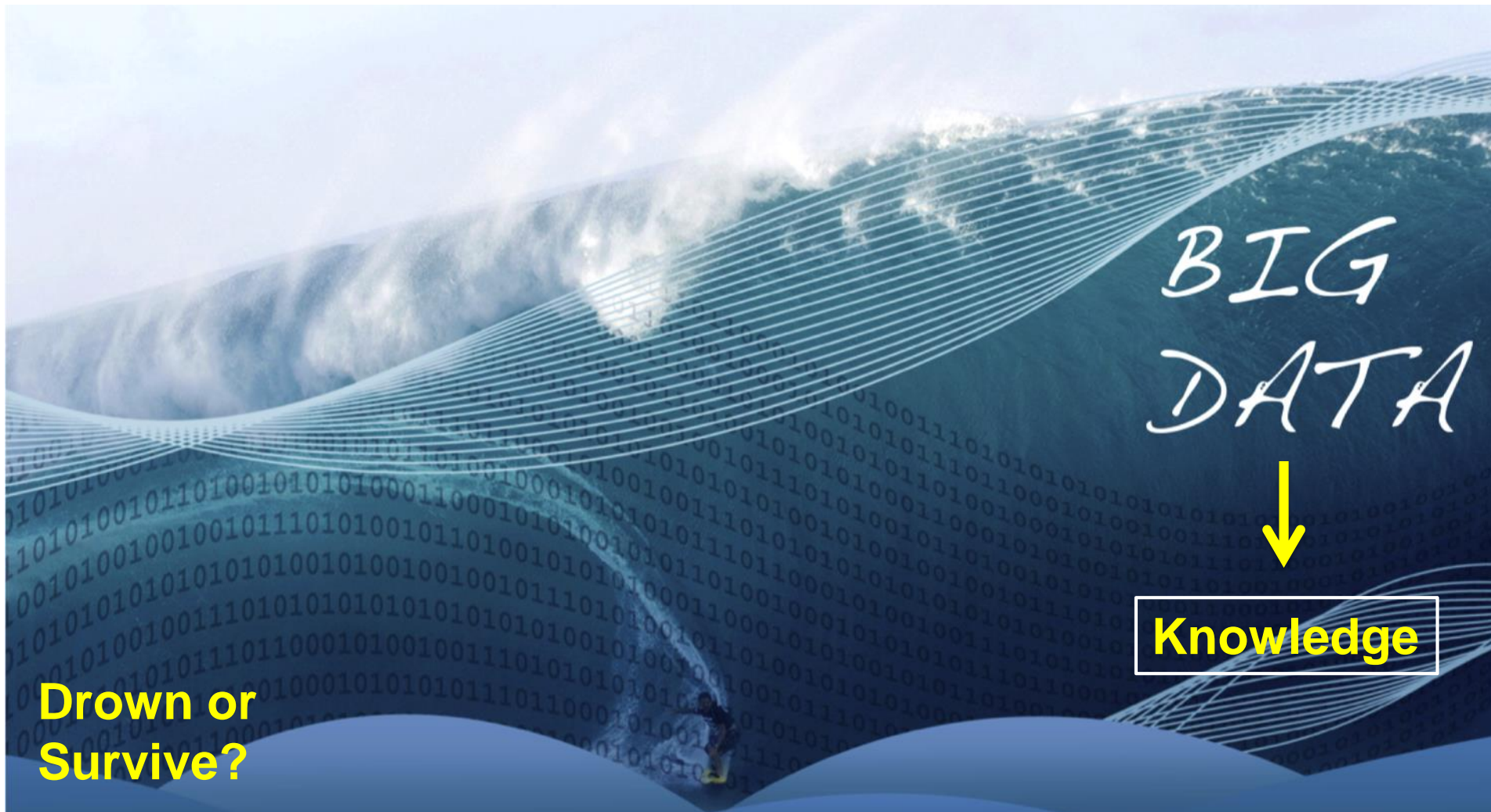
Microsoft Research
Faculty Summit
2015

AMiner II — Toward **Understanding** Big **Scholar** Data

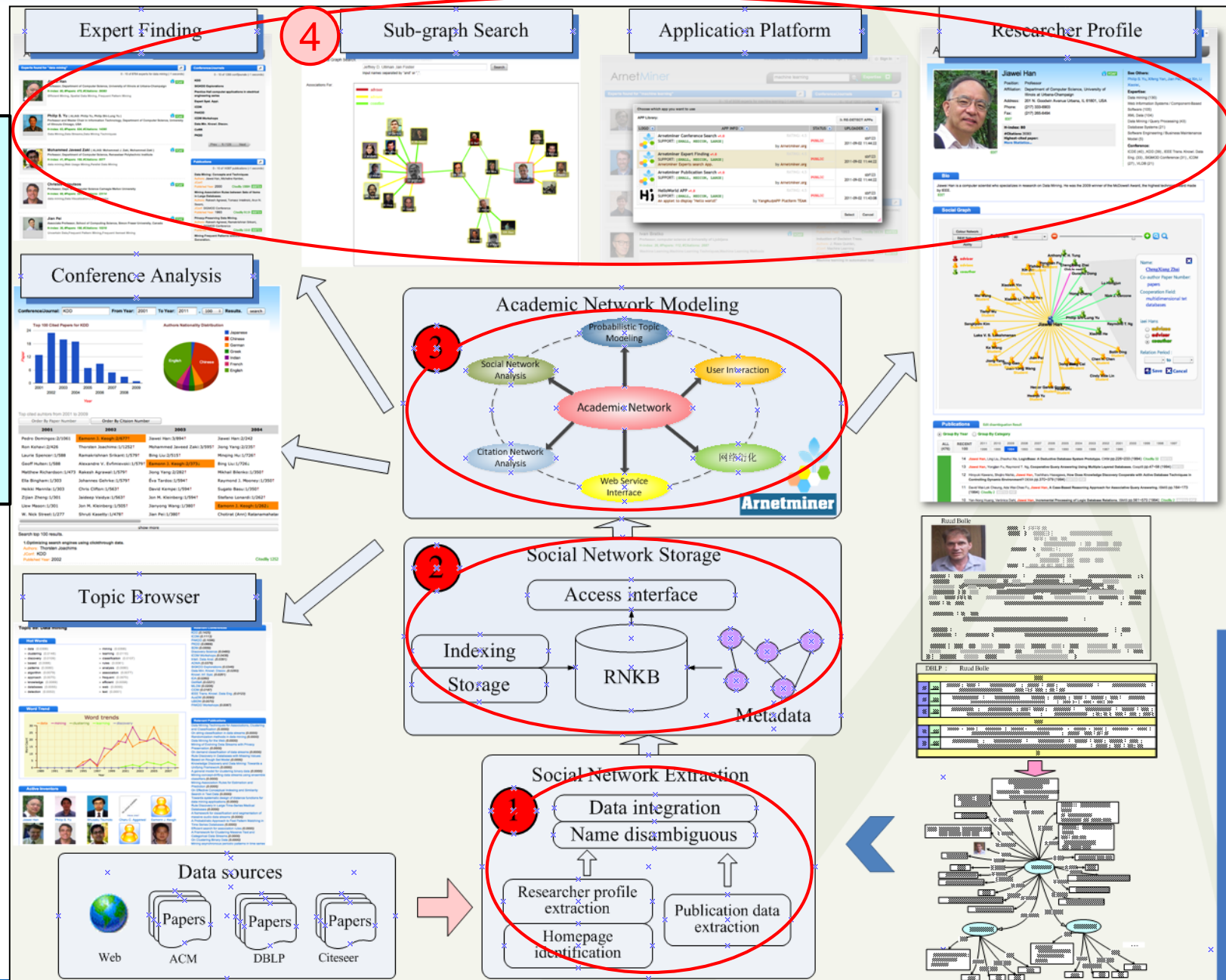
@2006-2015, <http://aminer.org>

Jie Tang
Tsinghua University

Mining Knowledge from Big Data



- Researcher profile extraction
- Expert finding
- Social network search
- Topic browser
- Conference analysis
- ArnetApp platform



Person Search



Jiawei Han (韩家炜)

Follow

Department of Computer Science, University of Illinois at Urbana-Champaign

Professor

(217) 333-6903

hanj@cs.uiuc.edu

<http://www.cs.uiuc.edu/~hanj/>

External Links

Update

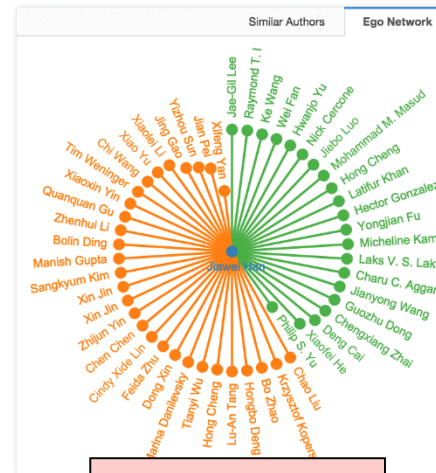
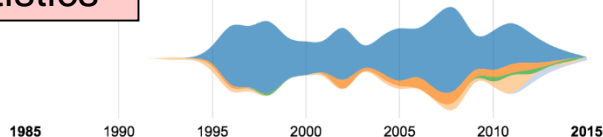
Basic Info.




Research Interests

- Data Mining
- Machine Learning
- Information Extraction
- Text Mining
- Data Analysis

Citation statistics




Ego network



Bringing Structure to Text: Mining Phrases, Entity Concepts, Topics, and Hierarchies
20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), New York 2014

Watch +



Mining Massive RFID, Trajectory, and Traffic Data Sets
14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Las Vegas 2008

Watch +

About

- Papers **790**
- Lectures **13**
- Patents **1**

1

Systems and Methods for Detecting a Novel Data Class

Mohammad Mehedy Masud, Latifur Rahman Khan, Bhavani Marienne Thuraishingam, Qing Chen, Jing Gao, Jiawei Han

Publication-date: 2012-03-01 Application-date: 2011-08-22

Add Paper Remove Paper By Year By Citation

| | | | | | | | | | | | | | | | | | | | | | | | |
|-----------|-------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|--|
| All (790) | Recent (20) | 2015 | 2014 | 2013 | 2012 | 2011 | 2010 | 2009 | 2008 | 2007 | 2006 | 2005 | 2004 | 2003 | 2002 | 2001 | 2000 | 1999 | 1998 | 1997 | 1996 | 1995 | |
| | | 1994 | 1993 | 1992 | 1991 | 1990 | 1989 | 1988 | 1987 | 1986 | 1985 | | | | | | | | | | | | |

790

Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions Cited by 4

Wei Shen, Jianyong Wang, **Jiawei Han**
Knowledge and Data Engineering, IEEE Transactions (2015)
<http://dx.doi.org/10.1109/TKDE.2014.2327028>

Patterns. Cited by 20

eng Yan, **Jiawei Han**
Mining (2014)
http://dx.doi.org/10.1007/978-1-4419-6045-0_12

Troubleshooting interactive complexity bugs in wireless sensor networks using data mining techniques Cited by 5

Mohammad Maffi Hasan Khan, Hieu Khac Le, Hossein Ahmadi, Tarek F. Abdelzaher, **Jiawei Han**
ACM Transactions on Sensor Networks (TOSN) (2014)
Bibtex <http://dx.doi.org/10.1145/2530290>

787

Expert Search

Finding experts, for "data mining"

data mining

Expert Paper

Search Results for data mining

Demographics: gender, language, location, etc.

| | | | | | | | | | | |
|------------|---------------|---------------|-------------|----------------|-------------|---------------|--------------|-------------|----------------|----------------|
| H-Index : | >=60 (31) | 50-60 (19) | 40-50 (52) | 30-40 (107) | 20-30 (167) | 10-20 (403) | <10 (199) | | | |
| Gender : | Male (934) | Female (44) | | | | | | | | |
| Language : | Chinese (290) | English (194) | Greek (37) | German (27) | French (23) | Japanese (20) | Korean (14) | Indian (12) | | |
| Location : | USA (219) | China (141) | Taiwan (34) | Australia (33) | Canada (29) | Japan (25) | Germany (24) | Italy (20) | Hong Kong (20) | Singapore (20) |

Relevance H-Index A-Index Activity Diversity Rising Star #Citation #Paper



Jiawei Han (韩家炜)

H-Index: 126 | #Paper: 790 | #Citation: 90481

Department of Computer Science, University of Illinois at Urbana-Champaign

Professor

Similar authors

similar authors

2567 views

Philip S. Yu

H-Index: 124 | #Paper: 838 | #Citation: 69439

Department of Computer Science, University of Illinois Chicago

Professor and Wexler Chair in Information Technology

Similar

321 views

Hillol Kargupta

H-Index: 40 | #Paper: 141 | #Citation: 6192

Department of Computer Science and Electrical Engineering University of Maryland Baltimore County

Associate Professor

Similar

118 views

Xindong Wu

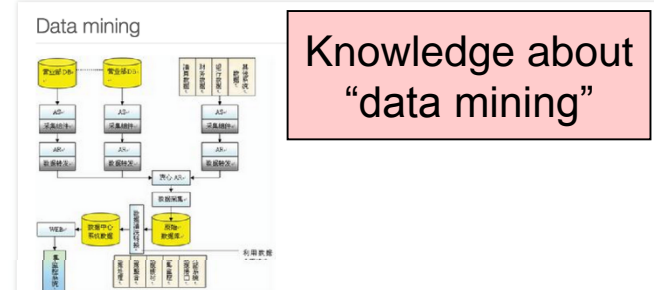
H-Index: 45 | #Paper: 331 | #Citation: 9644

Department of Computer Science, University of Vermont

Professor

Similar

35 views



Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a relatively young and interdisciplinary field of computer science, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

- Super Concepts:**
- Data analysis
 - Data mining
 - Formal sciences
 - Applied sciences
 - Networks
 - Artificial intelligence
- Related Concepts:**
- Data compression
 - Data visualization
 - Natural language processing
 - Data cleansing
 - Distributed computing
 - Infomatization
 - Speech recognition
 - Business intelligence
 - Pattern recognition
 - Spatial database
 - Full text search
 - Metadata
 - Computer vision
 - ISAM
 - Biological neural network
 - Database
 - Grid computing
 - Database marketing
 - Parallel computing

数据挖掘

数据挖掘(Data Mining)是通过分析每个数据,从大量数据中寻找其规律的技术,主要有数据准备、规律寻找和规律表示3个步骤。数据挖掘的任务有关联分析、聚类分析、分类分析、异常分析、特异群组分析和演变分析等。

上位词:

资料分析 计算机科学基础理论 决策支持系统 信息管理术语 数据挖掘 形式科学

Organization Ranking



Computer Science

Statistics Visualization

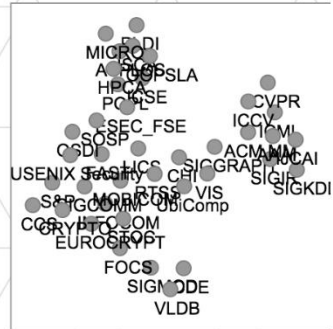
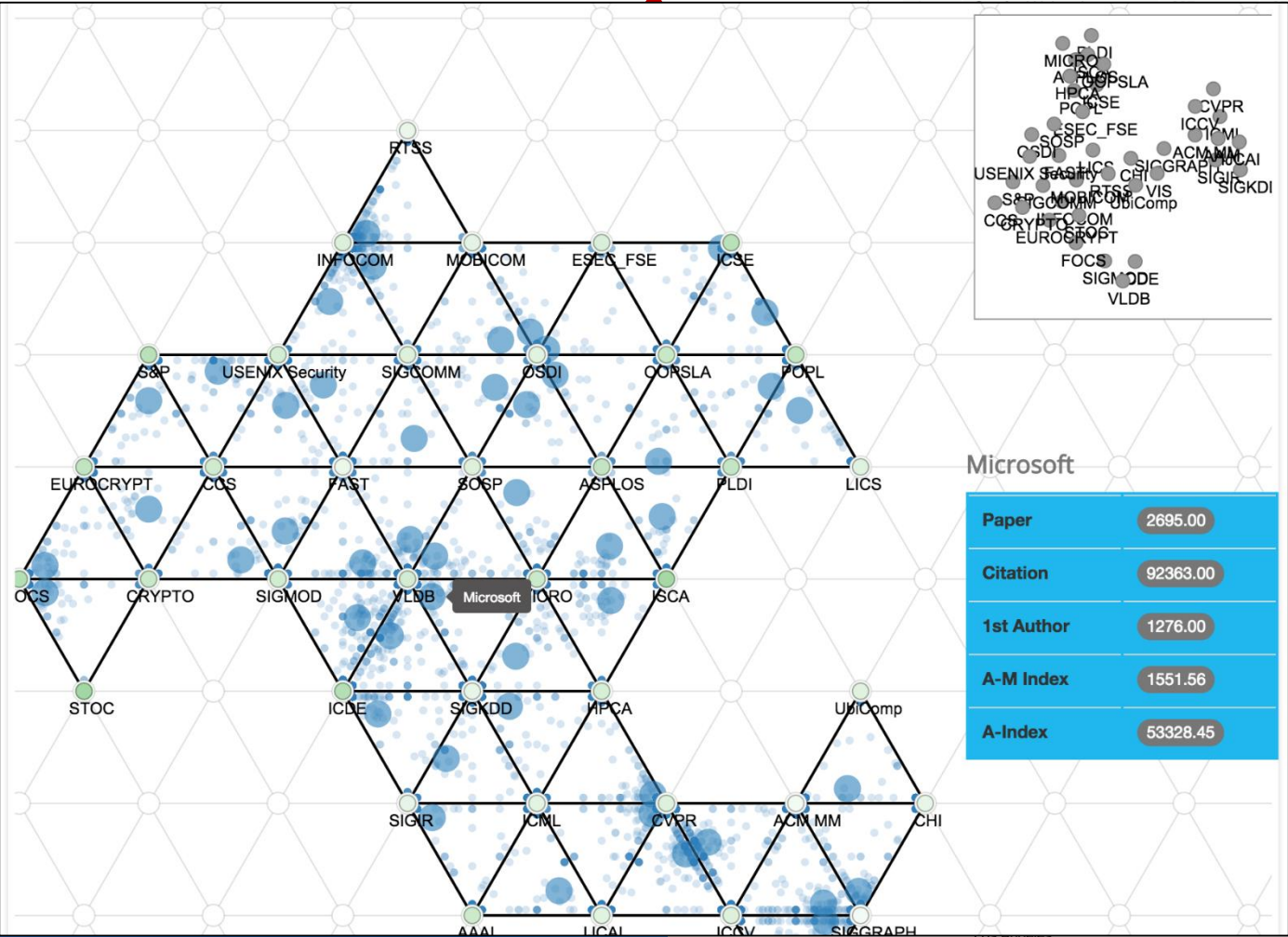
Search for Organization... Go!

- All
- High Performance Computing
- Computer Network
- Net and Information Security

All University Research institute Company

By different metrics

| Rank | Organization | #Paper | #FirstAuthor | Credit | #Citations | Weighted_Credit | A-Index | AM-Index |
|------|---------------------------------------|--------|--------------|----------|------------|-----------------|-----------|----------|
| 1 | Microsoft | 2,695 | 1,276 | 1,451.82 | 92,363 | 52,978.29 | 53,328.45 | 1,551.56 |
| 2 | Carnegie Mellon University | 1,456 | 1,081 | 983.48 | 52,700 | 33,540.08 | 34,854.62 | 888.69 |
| 3 | Massachusetts Institute of Technology | 1,110 | 780 | 748.87 | 50,019 | 33,984.45 | 34,356.75 | 823.14 |
| 4 | | | | 577.01 | 50,746 | 30,910.64 | 32,000.09 | 769.55 |
| 5 | | | | 882.87 | 49,432 | 28,605.32 | 28,968.54 | 724.06 |
| 6 | | | | 589.39 | 42,246 | 25,936.45 | 26,413.88 | 661.09 |
| 7 | | | | 723.26 | 37,566 | 26,671.27 | 27,062.57 | 646.47 |
| 8 | | | | 494.02 | 32,806 | 21,883.74 | 22,296.53 | 597.85 |
| 9 | | | | 499.99 | 22,065 | 15,803.23 | 16,064.56 | 431.86 |
| 10 | | | | 484.63 | 23,233 | 15,669.74 | 15,877.43 | 400.24 |
| 11 | | | | 363.71 | 21,406 | 14,344.10 | 14,576.56 | 397.31 |
| 12 | | | | 421.99 | 18,035 | 13,690.39 | 13,879.77 | 395.41 |
| 13 | | | | 475.18 | 20,106 | 12,130.08 | 12,426.60 | 385.67 |
| 14 | | | | 374.13 | 22,283 | 14,180.29 | 14,588.01 | 380.17 |
| 15 | | | | 484.95 | 19,742 | 13,084.09 | 13,232.71 | 369.72 |
| 16 | | | | 342.06 | 19,561 | 13,801.29 | 14,134.56 | 361.08 |
| 17 | | | | 432.66 | 18,332 | 12,104.18 | 12,315.72 | 343.56 |
| 18 | | | | 406.87 | 15,740 | 11,095.00 | 11,279.12 | 318.02 |
| 19 | | | | 231.29 | 18,441 | 11,764.28 | 11,821.74 | 300.50 |
| 20 | | | | 234.81 | 27,424 | 13,648.39 | 13,945.38 | 288.10 |
| 21 | | | | 270.80 | 16,841 | 12,821.13 | 12,796.51 | 284.73 |
| 22 | | | | 503.02 | 13,288 | 8,929.52 | 8,977.09 | 269.57 |
| 23 | | | | 242.03 | 17,496 | 10,056.54 | 10,345.91 | 269.57 |
| 24 | | | | 257.77 | 16,979 | 9,058.36 | 9,071.13 | 258.45 |
| 25 | | | | 245.24 | 14,406 | 8,208.11 | 8,169.25 | 254.69 |
| 26 | | | | 349.30 | 11,493 | 5,888.79 | 6,129.02 | 243.98 |
| 27 | | | | 330.87 | 17,799 | 9,487.88 | 9,628.10 | 243.48 |
| 28 | North Carolina State University | 417 | 320 | 290.43 | 12,593 | 8,032.33 | 8,238.18 | 230.03 |



Microsoft

| | |
|------------|----------|
| Paper | 2695.00 |
| Citation | 92363.00 |
| 1st Author | 1276.00 |
| A-M Index | 1551.56 |
| A-Index | 53328.45 |

Conference Ranking



Computer Science

- All
- High Performance Computing
- Computer Network
- Net and Information Security

| Rank | Conference (Full Name) | Short Name | Impact Factor |
|------|--|----------------|---------------|
| 1 | Science | | 162.00 |
| 2 | Nucleic Acids Research | NAR | 128.00 |
| 3 | IEEE Conference on Computer Vision and Pattern Recognition | CVPR | 112.00 |
| 4 | IEEE Transactions on Pattern Analysis and Machine Intelligence | TPAMI | 101.00 |
| 5 | NeuroImage | NeuroImage | 99.00 |
| 6 | IEEE Transactions on Industrial Electronics | TIE | 80.00 |
| 7 | Bioinformatics | Bioinformatics | 79.00 |
| 8 | Communications Magazine | ICM | 74.00 |
| 9 | Transactions on Signal Processing | TSP | 73.00 |
| 10 | Conference on Human Factors in Computing Systems | CHI | 71.00 |
| 11 | arXiv ePrint Archive | | 71.00 |
| 12 | Transactions on Image Processing | TIP | 69.00 |
| 13 | Automatica | Automatica | 69.00 |
| 14 | International Conference on Computer Communications | INFOCOM | 69.00 |
| 15 | Communications of the ACM | Commun. ACM | 69.00 |
| 16 | Transactions on Automatic Control | TAC | 67.00 |
| 17 | World Wide Web Conferences | WWW | 66.00 |
| 18 | Proceedings of the IEEE | Proc. IEEE | 64.00 |

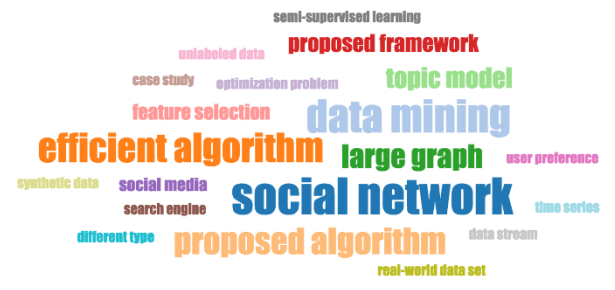
ACM Knowledge Discovery and Data Mining

Search for Conference

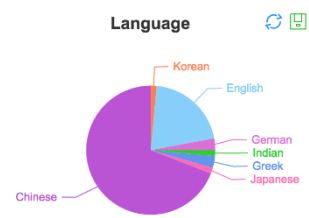
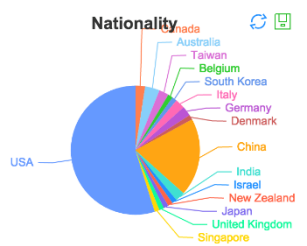
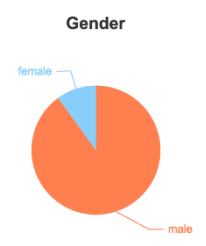
Conference/Journal

From Year

To Year



Author Distribution



Top cited authors

| 2008 | 2009 | 2010 | 2011 | 2012 | 2013 |
|----------------------------|-------------------------|-----------------------|---|------------------------------|-------------------------------|
| Yehuda Koren:1/924 | Jure Leskovec:1/778† | Wei Chen:1/437† | Eunjoon Cho:1/541† | Jing Yuan:1/137† | Chris Thornton:1/64† |
| Jure Leskovec:1/474 | Wei Chen:1/578† | Yu Wang:1/168† | Dashun Wang:1/208† | Thanawin Rakthanmanon:1/108† | Bin Liu:1/33† |
| Jie Tang:1/440 | Jie Tang:1/384 | Hongning Wang:1/162† | Jing Yuan:1/171† | Alan Ritter:1/103† | Yu Zheng:1/33† |
| Yabo Xu:1/423 | Mohsen Jamali:1/312† | Deng Cai:1/135† | Rainer Gemulla:1/167† | Isabelle Stanton:1/86† | Hongzhi Yin:1/31† |
| Victor S. Sheng:1/420 | Justin Ma:1/259† | Maayan Roth:1/131† | Salvatore Scellato:1/164† | Ashton Anderson:1/78† | Arjun Mukherjee:1/31† |
| David Crandall:1/404 | Albert Bifet:1/249† | Liang Xiang:1/128† | Marco Pennacchiotti:1/106† | Ling-Yin Wei:1/75† | Charalampos Tsourakakis:1/27† |
| Aris Anagnostopoulos:1/392 | Theodoros Lappas:2/235† | Arik Friedman:1/125† | Noman Mohammed:1/93† | Rui Li:1/75† | Quan Yuan:1/26† |
| Huanhuan Cao:1/313 | Anna Monreale:1/229† | Min-Ling Zhang:1/113† | Mao Ye:1/90† | Jie Tang:2/68† | Reza Zafarani:1/24† |
| David J. Crandall:1/218 | Sayali Kulkarni:1/228† | Yong Ge:1/104† | Wei Liu:1/89† | Jiliang Tang:2/67† | Madhav Jha:1/23† |
| Ian Porteous:1/216 | Prem Melville:1/207† | Michael Jahrer:1/103† | Robson Leonardo Ferreira Cordeiro:1/74† | Xiwang Yang:1/65† | Wook-Shin Han:1/22† |

Reviewer Suggestion

Interest matching
COI avoiding
Load balancing
Forecast review quality



Title
ArnetMiner: extraction and mining of academic social networks.

Authors (Seperated by comma)
Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, Zhong Su

Abstracts
This paper addresses several key issues in the ArnetMiner system, which aims at extracting and mining academic social networks. Specifically, the system focuses on: 1) Extracting researcher profiles automatically from the Web; 2) Integrating the publication data into the network from existing digital libraries; 3) Modeling the entire academic network; and 4) Providing search services for the academic network. So far, 448,470 researcher profiles have been extracted using a unified tagging approach. We integrate publications from online Web databases and propose a probabilistic framework to deal with the name ambiguity problem. Furthermore, we propose a unified modeling approach to

Conference (Journal)
KDD

SEARCH

Keywords: input keyword +

academic social network x Providing search service x expertise search x
people association search x

Relevant Conferences/Journals: input jour./conf. +

ICDM x PAKDD x PKDD x SDM x Discovery Science x

The recommended reviewers: BM25 +T +A +S All

View All Relevant Publications

Yajun Wang
Microsoft Research Asia
H-index: 14, #Papers: 49, #Citations: 677
social network, principal component analysis, Shortest Path
+ Relevant Publications: [thumbs up] [thumbs down] [share] [house]

Ming-Syan Chen (ALIAS: Ming-Syan Syan Chen)
National Taiwan University
H-index: 42, #Papers: 278, #Citations: 9978
Data Mining, Data Streams, Data Replication
+ Relevant Publications: [thumbs up] [thumbs down] [share] [house]

Michael R. Berthold (ALIAS: Michael Berthold)
KNIME.com, University of Konstanz
H-index: 17, #Papers: 77, #Citations: 1180
International Symposium, Data Analysis, Second International Symposium
+ Relevant Publications: [thumbs up] [thumbs down] [share] [house]

Christoph Lingenfelder
German Software Development Lab, IBM
H-index: 4, #Papers: 12, #Citations: 35
Knowledge-Based Methods, Proof Transformation, Der rechtliche Schutz von
+ Relevant Publications: [thumbs up] [thumbs down] [share] [house]

Michael Zeller
Zementis
H-index: 4, #Papers: 7, #Citations: 68
cloud computing, neural networks, open standard, predictive analytics, data mining, predictive model markup language, pmm1
+ Relevant Publications: [thumbs up] [thumbs down] [share] [house]

Theodoros Lappas
H-index: 4, #Papers: 12, #Citations: 124
social network, interactive recommendation, Efficient Confident Search
+ Relevant Publications: [thumbs up] [thumbs down] [share] [house]

Reviewer Suggestion

HTML PDF Supplementary Files Abstract Cover Letter External Searches

Reviewer List

| Order | Name | Status | History | Remove |
|-------|--|--|---|-------------------------------------|
| 1 | Yangarber, Roman | Invited Response: <input type="text" value="Select..."/> <input checked="" type="checkbox"/> Save | Invited: 07-Jul-2015 view full history | <input checked="" type="checkbox"/> |
| 2 | Toussaint, Yannick | Invited Response: <input type="text" value="Select..."/> <input checked="" type="checkbox"/> Save | Invited: 07-Jul-2015 view full history | <input checked="" type="checkbox"/> |
| 3 | Nie, Zaigang Microsoft Research Asia, Web Search and Mining Group | Invited Response: <input type="text" value="Select..."/> <input checked="" type="checkbox"/> Save | Invited: 07-Jul-2015 view full history | <input checked="" type="checkbox"/> |
| 4 | Jia, Yunde | Declined | Invited: 07-Jul-2015 Declined : 07-Jul-2015 view full history | <input checked="" type="checkbox"/> |
| 5 | Weninger, Tim | Invited Response: <input type="text" value="Select..."/> <input checked="" type="checkbox"/> Save | Invited: 07-Jul-2015 view full history | <input checked="" type="checkbox"/> |
| 6 | Jung, Kyomin Computer Science | Declined | Invited: 07-Jul-2015 Declined : 07-Jul-2015 view full history | <input checked="" type="checkbox"/> |

Alternates Save

Author's Preferred / Non-Preferred Reviewers

| Name, Keywords, Institution/Organization, Roles | Current / Past 12 Months | Days Since Last Review | Average R-Score | Add |
|---|--------------------------|------------------------|-----------------|-----|
| 0 people entered. | | | | |

Progress

reviews required to make decision: 3

active selections: 4

invited: 4

agreed: 0

declined: 2

returned: 0

Save

Create Reviewer Account

req Salutation:

req First (Given) Name:

req Last (Family) Name:

req E-Mail Address:

Create and Add

Version History

TKDE-2015-06-0517

Submitted on 22-Jun-2015

Reviewer Recommender

Keywords (Separated by ", "):
Text processing , Applications , Pattern Recognition , Computing Methodologies , Evolutionary computing and genetic algorithms , Miscellaneous , Artificial Intelligence , Computing Methodologies , Text mining ,

H-index: <10 10~20 20~30 >30

Location: Language:

Result (20 of 207 experts)

I-Jeng Wang
(Assistant Research Professor, Department of Computer Science)
H-index: 11
Expertise: Probabilistic Models, Wireless Sensor Networks, Sensor Networks, Distributed Systems, Sensor Network
E-mail: ijwang@cs.jhu.edu

Select

Steffi Beckhaus
(Professor, University of Hamburg (W1, Juniorprofessor), head of im/vc group)
H-index: 10
Expertise: Virtual Environment, Virtual Reality, Collision Avoidance, Virtual Environments, Individual Tutorial
E-mail: steffi.beckhaus@uni-hamburg.de

Select

Dick De Ridder
(, Information & Communication Theory Group Delft University of Technology)
H-index: 18
Expertise: Pattern Recognition, Neural Networks, Artificial Neural Networks, Neural Network, Protein Function Prediction
E-mail: <http://genlab.tudelft.nl/~dick/pic/email.gif>

Select

Powered by [AMiner](#)

AMiner II (ArnetMiner)

Academic Social Network Analysis and Mining system—AMiner (<http://aminer.org>)

- Online since 2006
- >38 million researcher profiles
- >76 million publication papers
- >241 million requests
- >12.35 Terabyte data
- 100K IP access from 170 countries per month
- 10% increase of visits per month

Deep analysis, mining, and search

Jiawei Han
 Position: Professor
 Affiliation: Department of Computer Science, University of Illinois at Urbana-Champaign
 Address: 201 N. Goodwin Avenue, Urbana, IL 61801, USA.
 Phone: (217) 333-6503
 Fax: (217) 265-6434
 Email: han@jcs.uiuc.edu
 Links: [Social Media Icons]

STATISTIC

| | | | | | |
|-------------|-------|------------|-------|--------------------|--------|
| H-index: | 96 | Uptrend: | 30.46 | Diversity: | 0.71 |
| #Papers: | 553 | Activity: | 32.94 | Sociability: | 726.64 |
| #Citations: | 55885 | Longevity: | 26 | More Statistics... | |

Bio
 Jiawei Han is computer scientist who specializes in research on Data Mining. He was the 2009 winner of the MacDowell Award, the highest technical award made by IEEE. He is currently a professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign. Previously he was a professor in the School of Computing Science at Simon Fraser University. He is an ACM fellow and an IEEE fellow.

Research Interest
 [Spatial Data Mining] [Frequent Pattern Mining] [More]

See Others:
 Philip S. Yu (Coauthor-Count: 50, H-index: 86)
 Xifeng Yan (Coauthor-Count: 54, H-index: 32)
 Jian Pei (Coauthor-Count: 37, H-index: 49)

Expertise:
 Data mining (146)
 Mobile Robot / Hybrid Control (119)
 XML, Data (110)
 Data Mining / Query Processing (44)
 Database Systems (21)
 Information Retrieval / Probabilistic Indexing (14)

Conference:
 KDD (49) ICDE (44)
 IEEE Trans. Knowl. Data Eng. (31)
 SIGMOD (33)
 SIGBIOD Conference (12)
 VLDB (23)

Experts found for "data mining" (0 - 10 of 1748 experts for data mining (0 seconds))

| | |
|--|---|
| Jiawei Han Professor, Department of Computer Science, University of Illinois at Urbana-Champaign H-index: 96 #Papers: 553 #Citations: 55885 | Philip S. Yu (ALIAS: Philip Yu, Philip Shi-Lang Yu) Professor and Distinguished Chair in Information Technology, Department of Computer Science, University of Illinois Chicago H-index: 86 #Papers: 493 #Citations: 30504 |
| Mohammed Javeed Zaki (ALIAS: Mohammed J. Zaki, Mohammed Zaki) Professor, Department of Computer Science, Rensselaer Polytechnic Institute H-index: 49 #Papers: 162 #Citations: 3690 | Christos Faloutsos Professor, Dept. of Computer Science, Carnegie Mellon University H-index: 41 #Papers: 237 #Citations: 3154 |
| Jian Pei Professor, School of Computing Science, Simon Fraser University H-index: 43 #Papers: 219 #Citations: 15467 | H. Mannila (ALIAS: Heikki Mannila) Professor, Helsinki University of Technology H-index: 52 #Papers: 192 #Citations: 14762 |
| Charu C. Aggarwal (ALIAS: Charu Chandr, Charu Chandr Aggarwal, Charu Aggarwal) Research Scientist, IBM T. J. Watson Research Center H-index: 48 #Papers: 269 #Citations: 11620 | |

Conference/Journals
 SIGMOD Extensions
 Data Mining Workshops
 Expert Syst. Appl.
 PAKDD
 ICDM Workshops
 COIS
 Data Min. Knowl. Discov.
 PKDD

Social Graph

Relations: [Filter]

Publications

Group By: Year | Group By: Category

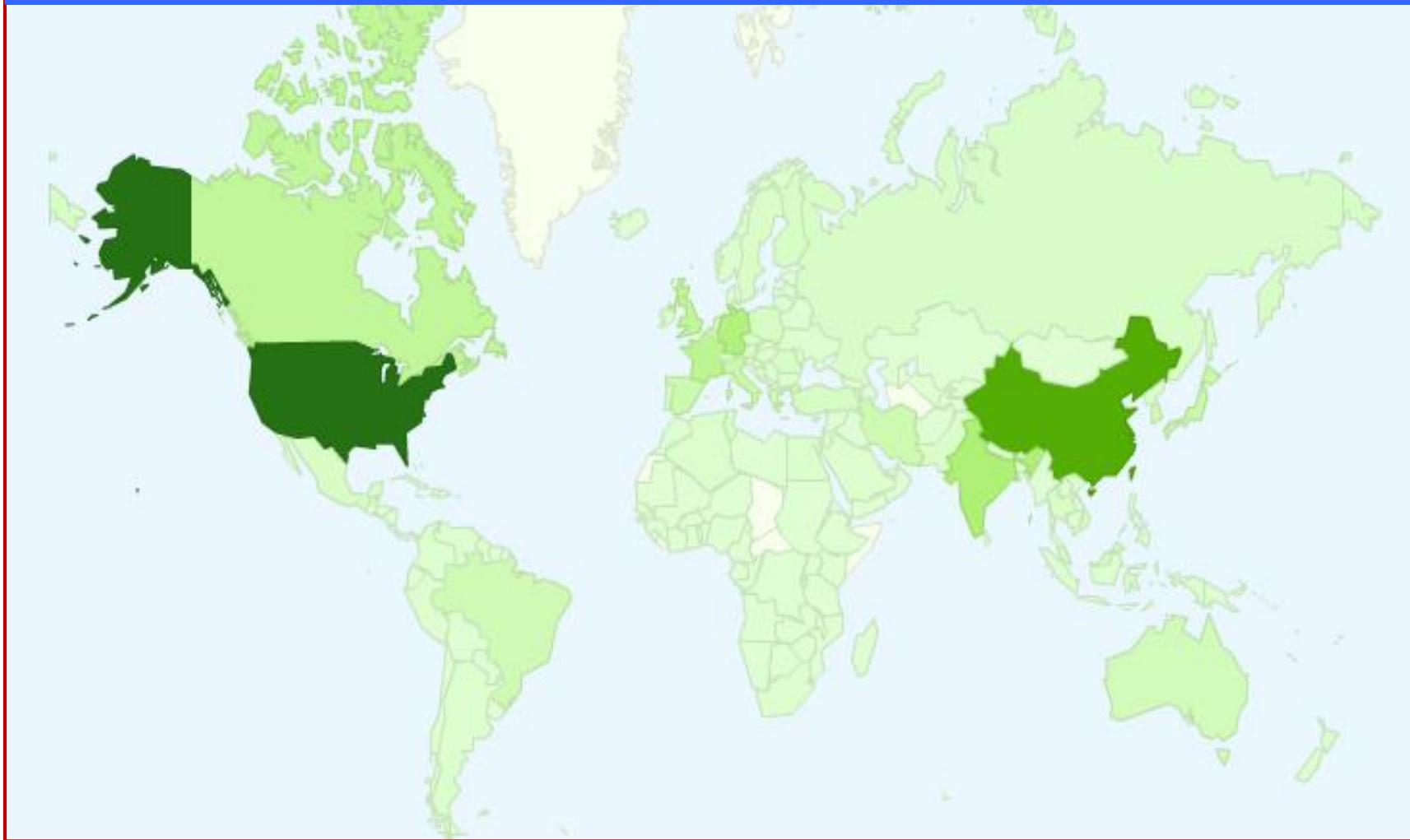
101 Query-driven Discovery of Semantically Similar Substructures in Heterogeneous Networks
 2012, KDD, pp. 1000-1009

103 Integrating Multi-Path Selection with User-Guided Object Clustering in Heterogeneous Information Networks
 2012, SIGMOD, pp. 1340-1356

105 Mining periodic behaviors of abstract environments for animal and biological sustainability studies
 2012, Data Min. Knowl. Discov., pp. 601-608

User Distribution

6.32 million IP from 220 countries/regions



User Distribution

6.32 million IP from 220 countries/regions



Top 10 countries

- | | |
|------------|-----------|
| 1. USA | 6. Canada |
| 2. China | 7. Japan |
| 3. Germany | 8. Spain |
| 4. India | 9. France |
| 5. UK | 10. Italy |

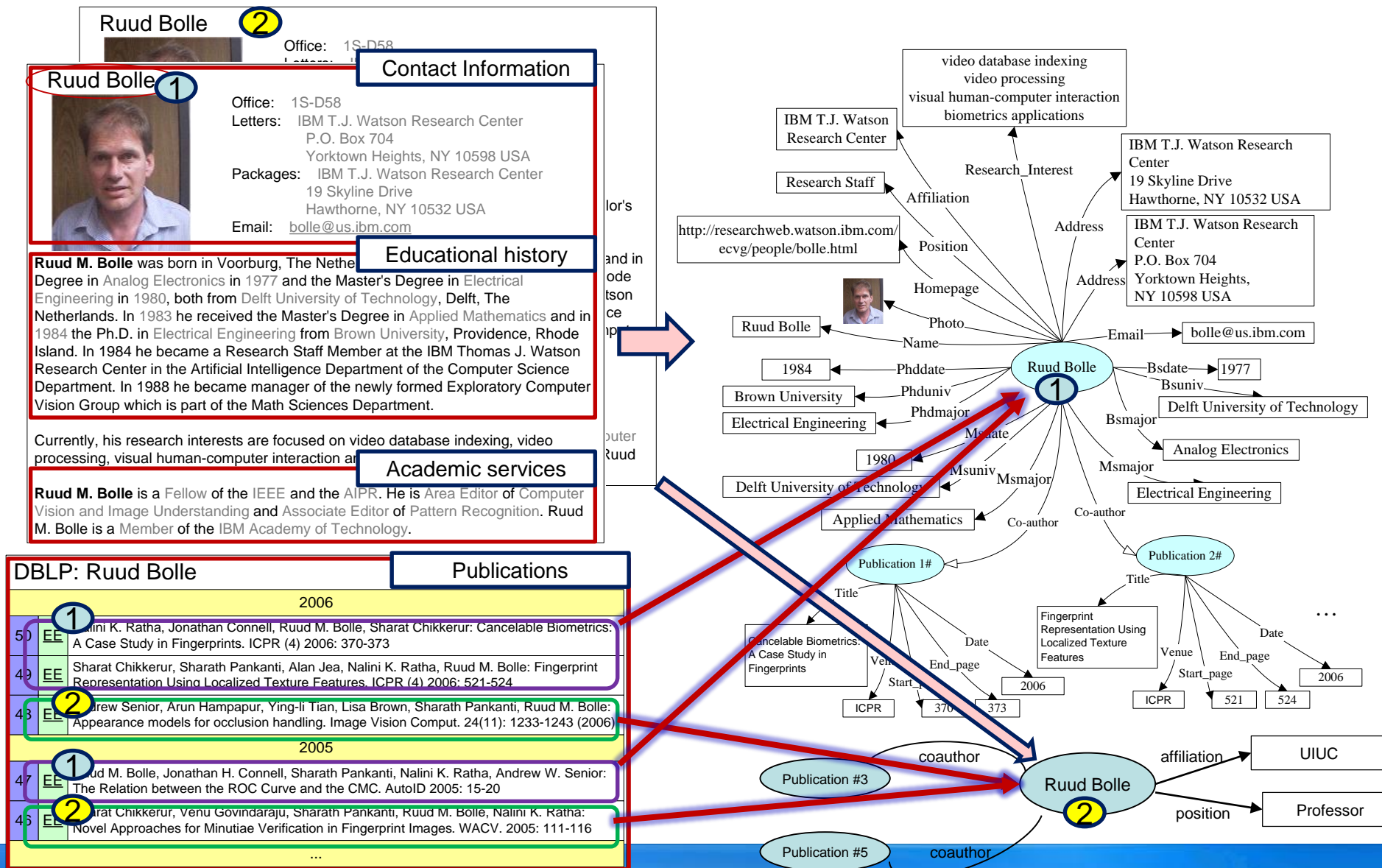
Technologies

—Toward understanding big scholar data

Recent progress...

Knowledge Acquisition from the Web

(ACM TKDD, WWW'12, ISWC'06, ICDM'07, ACL'07)



Researcher Profile Database^[1]

Extracted more than 1,000,000 researcher profiles from the Web

Jiawei Han
 Position: Professor
 Affiliation: Department of Computer Science, University of Illinois at Urbana-Champaign
 Address: 201 N. Goodwin Avenue, Urbana, IL 61801, USA.
 Phone: (217) 333-6903
 Fax: (217) 265-6494
 Email: hanj[at]cs.uiuc.edu
 Links: [Social Media Icons]

STATISTIC

| | | | | | |
|-------------|-------|------------|-------|------------------------------------|--------|
| H-index: | 96 | Uptrend: | 30.46 | Diversity: | 0.71 |
| #Papers: | 553 | Activity: | 32.04 | Sociability: | 726.64 |
| #Citations: | 55885 | Longevity: | 26 | More Statistics... | |

Bio
 Jiawei Han is computer scientist who specializes in research on Data Mining. He was the 2009 winner of the McDowell Award, the highest technical award made by IEEE. He is currently a professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign. Previously he was a professor in the School of Computing Science at Simon Fraser University. He is an ACM fellow and an IEEE fellow.

Research Interest
[E-commerce Mining](#) [Spatial Data Mining](#) [Frequent Pattern Mining](#)

Education
 Phd University

M. I. Jordan
 ALIAS: Michael I. Jordan, Michael Jordan, Michael Irwin Jordan
 Position: Professor
 Affiliation: Department of EECS Department of Statistics University of California, Berkeley
 Address: University of California, Berkeley EECS Department 731 Soda Hall #1776 Berkeley, CA 94720-1776
 Phone: (510) 642-3806
 Fax: (510) 642-5775
 Email: jordan@stat.berkeley.edu
 Links: [Social Media Icons]

STATISTIC

| | | | | | |
|-------------|-------|------------|-------|------------------------------------|--------|
| H-index: | 75 | Uptrend: | 7.2 | Diversity: | 0.03 |
| #Papers: | 242 | Activity: | 11.12 | Sociability: | 331.69 |
| #Citations: | 44312 | Longevity: | 23 | More Statistics... | |

H. Garcia
 ALIAS: H. Garcia Molina, H. Garcia Molina, Hector Garcia Molina, Hector Garcia Molina
 Position: Professor
 Affiliation: Departments of Computer Science and Electrical Engineering, Stanford University
 Address: Department of Computer Science Stanford University Gates Hall 4A, Room 434 Stanford, CA 94305-9040 USA
 Phone: (650) 723-0685
 Fax: (650) 725-2598
 Email: hector@cs.stanford.edu

See Others:
 Andreas Paepcke (Coauthor-Count: 32, H-index: 37)
 Jennifer Widom (Coauthor-Count: 24, H-index: 79)
 D. Barbara (Coauthor-Count: 26, H-index: 0)

Expertise:
 Data (115)
 Base Systems (60)
 Time Systems / Automated rare Test Data (30)
 Mining (23)
 Mobile Robot / Hybrid Control (22)
 Digital Library / Information Access

Conference:
 VLDB (27)
 SIGMOD Conference (25)
 ICIS (23)
 Data Eng. Bull. (13)
 DB (11)

Research Interest
[Database Systems](#) [Data Management](#) [Data Warehousing](#)

Scott
 Position: Professor
 Affiliation: Department of Computer Science, Stanford University
 Address: Department of Computer Science Stanford University Gates Hall 4A, Room 434 Stanford, CA 94305-9040 USA
 Phone: (650) 723-0685
 Fax: (650) 725-2598
 Email: hector@cs.stanford.edu

STATISTIC

| | | | | | |
|-------------|-------|------------|-------|------------------------------------|--------|
| H-index: | 96 | Uptrend: | -4.04 | Diversity: | 0.22 |
| #Papers: | 195 | Activity: | 4.86 | Sociability: | 407.19 |
| #Citations: | 57908 | Longevity: | 25 | More Statistics... | |

Expertise:
 Wireless network / End-to-end Routing Behavior (80)
 ATM Networks (21)

Research Interest
[Database Systems](#) [Data Management](#) [Data Warehousing](#)

Bio
 Monterrey, Mexico, in 1974. From Stanford University, Stanford, California, he received in 1975 a MS in electrical engineering and a PhD in computer science in 1979. He holds an honorary PhD from ETH Zurich (2007). Garcia-Molina is a Fellow of the Association for Computing Machinery and of the American Academy of Arts and Sciences; is a member of the National Academy of Engineering; received the 1999 ACM SIGMOD Innovations Award, is a Venture Advisor for Onset Ventures, and is a member of the Board of Directors of Oracle.

[1] J. Tang, L. Yao, D. Zhang, and J. Zhang. A Combination Approach to Web User Profiling. ACM Transactions on Knowledge Discovery from Data (TKDD), (vol. 5 no. 1), Article 2 (December 2010), 44 pages.

Is this Enough?

A Miner

Home |



Jeannette Wing 🔒

Computer Science Department Carnegie Mellon University

Professor of Computer Science

412-260-8926 (cell)

wing@cs.cmu.edu

<http://www.cs.cmu.edu/~wing/>

External Links Update

[in](#)

Research Interests

● Formal Methods
● Software Engineering
● Formal Verification

● Embedded Systems
● Specification Languages



Similar Authors
Ego Network





Required Semantics are distributed in Multiple Sources



LinkedIn

Jeannette Wing 2nd
Corporate Vice President at Microsoft
Redmond, Washington | Computer Software

Current Microsoft, Carnegie Mellon University, Computer Science
Previous Microsoft, Carnegie Mellon University, National Science Foundation/Computer and Information Science and Engineering Directorate
Education Massachusetts Institute of Technology

[Connect](#) [Send Jeannette InMail](#) 446 connections

<https://www.linkedin.com/pub/jeannette-wing/3/8a6/1b8> [Contact Info](#)

Background

Experience

Corporate Vice President
Microsoft
July 2013 – Present (2 years 1 month) | Redmond
In charge of core research labs at Microsoft Research.

President's Professor of Computer Science
Carnegie Mellon University, Computer Science
1985 – Present (30 years)

Wikipedia

Jeannette Wing

From Wikinedia, the free encyclopedia
Pittsburgh

Jeannette Marie Wing is Corporate Vice President of [Microsoft Research](#) with oversight of its core research laboratories around the world and Microsoft Research Connections.^{[2][3]} Prior to 2013, she was the President's Professor of [Computer Science](#) at [Carnegie Mellon University](#), [Pittsburgh](#), [Pennsylvania](#), [United States](#). She also served as assistant director for Computer and Information Science and Engineering at the [NSF](#) from 2007 to 2010.^{[4][5][6][7][8][9][10][11][12][13]}

Contents [\[hide\]](#)

- Education
- Career and research
- References
- External links

Education

[\[edit\]](#)

Wing earned her S.B. and S.M. in Electrical Engineering and Computer Science at [MIT](#) in June 1979. Her advisers were [Ronald Rivest](#) and [John Reiser](#). In 1983, she earned her Ph.D. in Computer Science at MIT under [John Guttag](#).^[1]

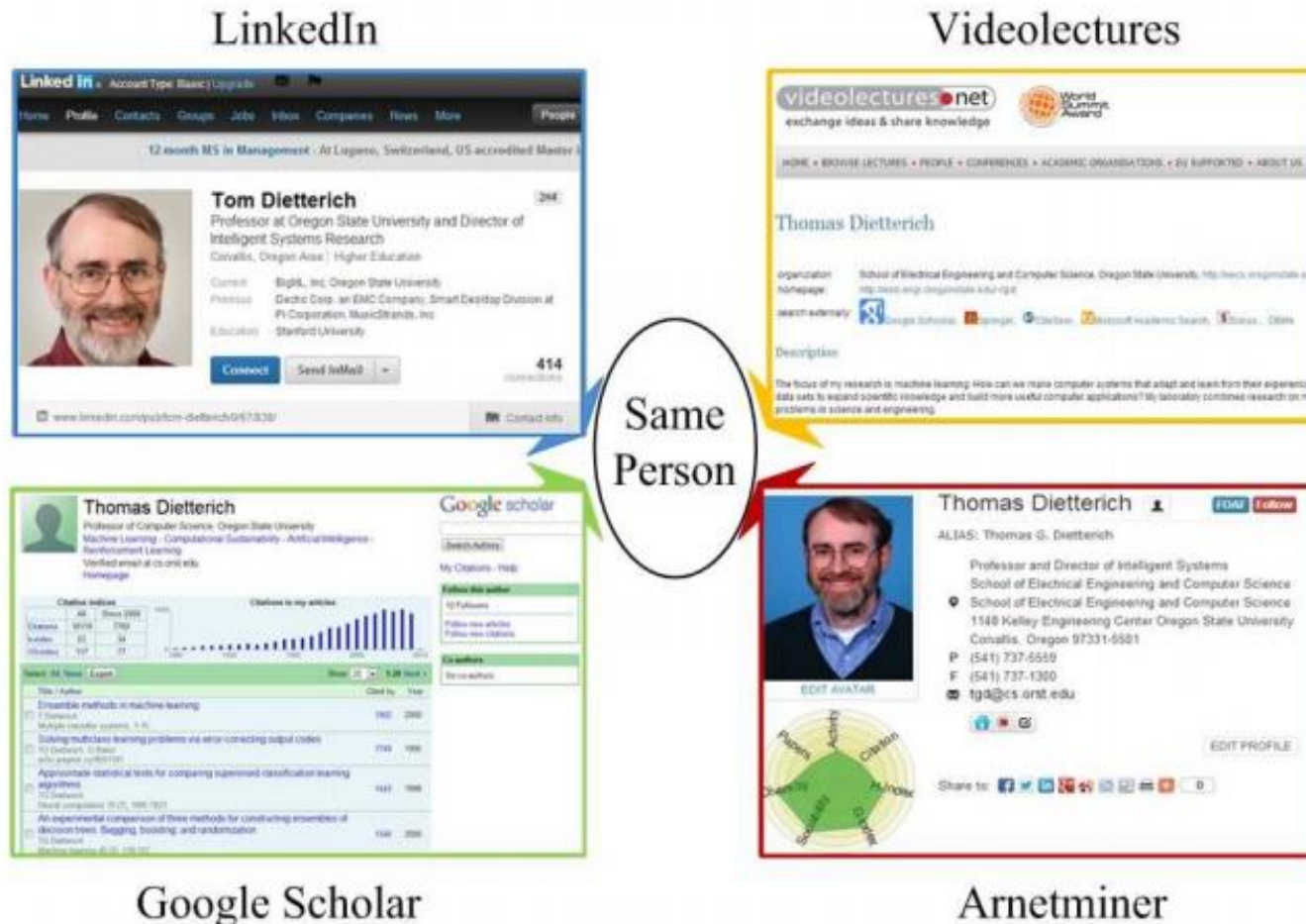
Jeannette Wing

Speaking at the [World Economic Forum](#) in [Davos](#), Switzerland, on January 26, 2013.

Born Jeannette Marie Wing
Nationality [American](#)
Fields [Computer science](#)
Institutions [Carnegie Mellon University](#)
Alma mater [Massachusetts Institute of Technology](#)
Thesis [A Two-Tiered Approach to Specifying Programs](#) (1983)

Network Integration

- Identifying users from multiple heterogeneous networks and integrating semantics from the different networks together.



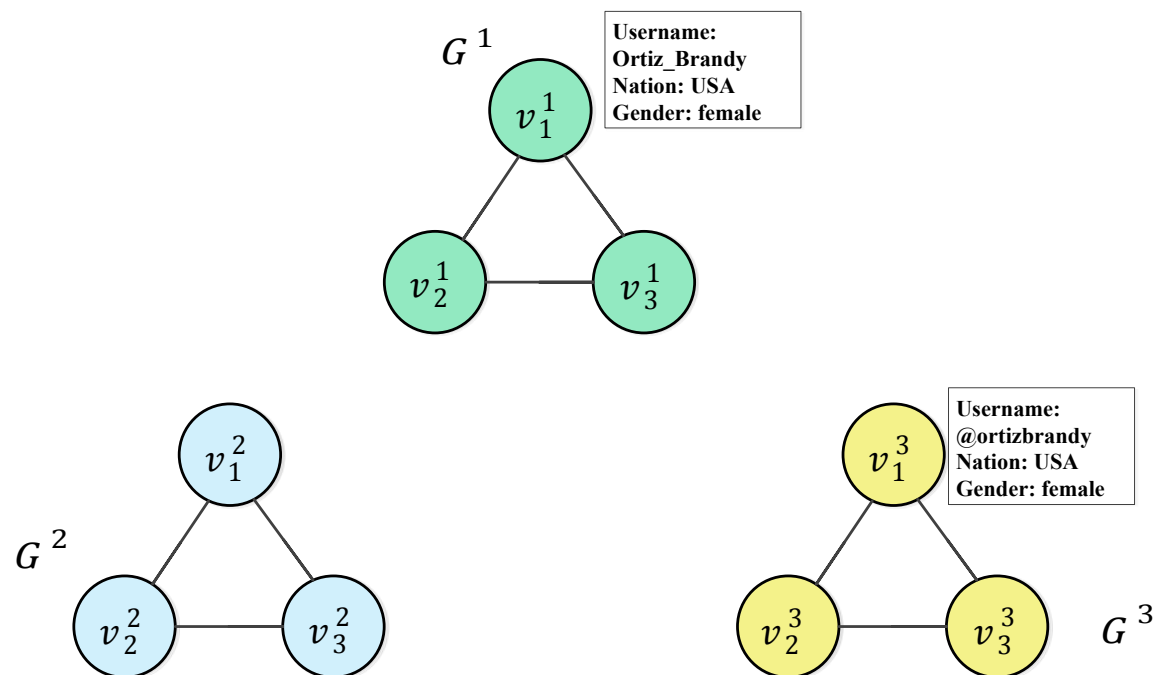
COSNET: Connecting Social Networks with Local and Global Consistency

- **Input:** $\mathbf{G} = \{G^1, G^2, \dots, G^m\}$, with $G^k = (V^k, E^k, R^k)$
- **Formalization:** $\mathbf{X} = \{x_i\}$, all possible pairwise matchings and each corresponds to $y_i \hat{\in} \{1, 0\}$
- **COSNET:** an energy-based model

$$Y^* = \arg \max E(Y, X)$$

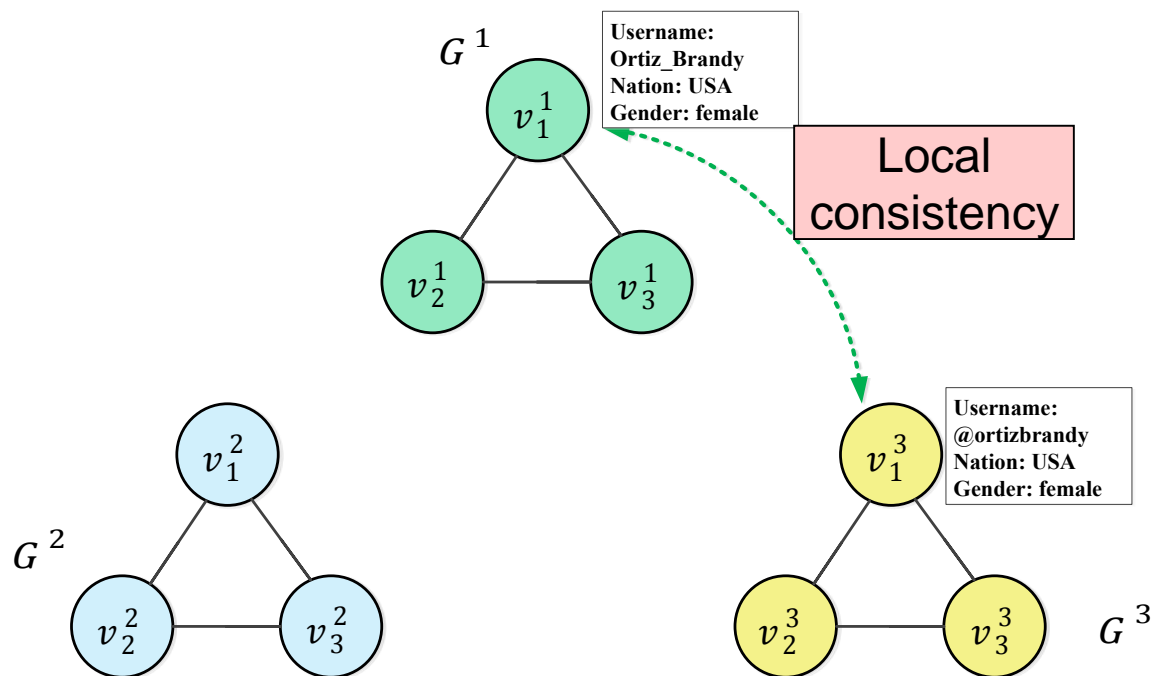
Local vs. Global consistency

- Given three networks,



Local vs. Global consistency

- Local matching: matching users by profiles



Pairwise similarity features

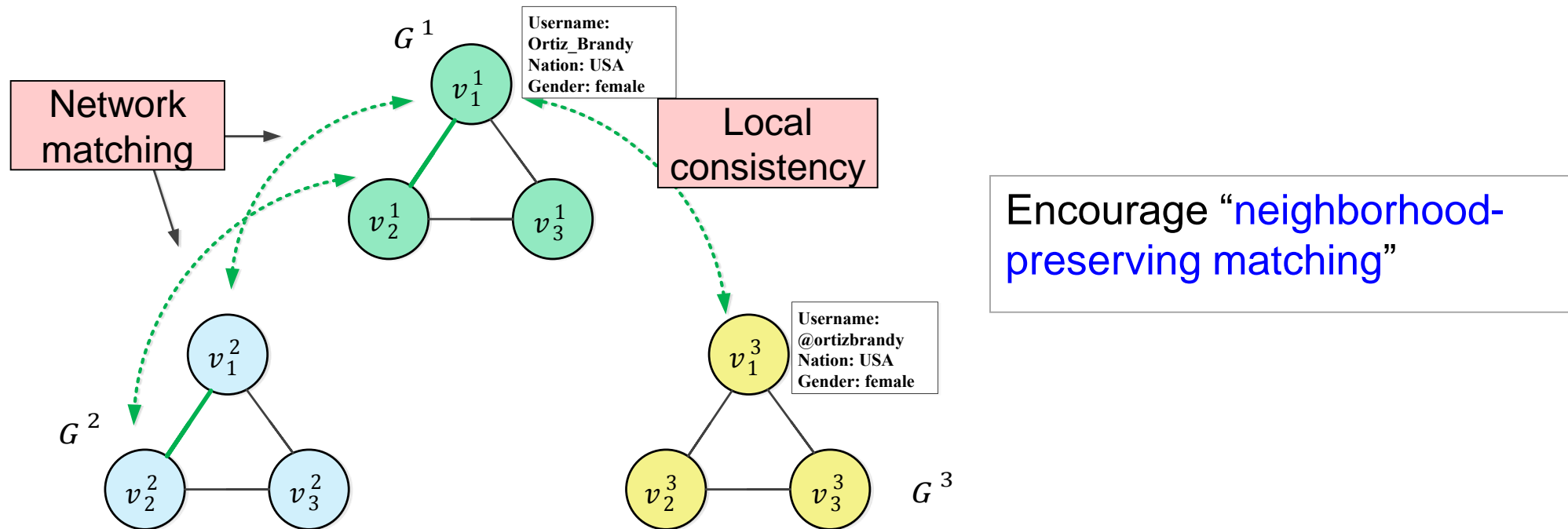
- Username similarity and uniqueness
- Profile content similarity
- Ego network similarity
- Social status

Energy function

$$E_l(Y, X) = \sum_i \mathbf{w}_l^T \mathbf{g}_l(\mathbf{x}_i, y_i)$$

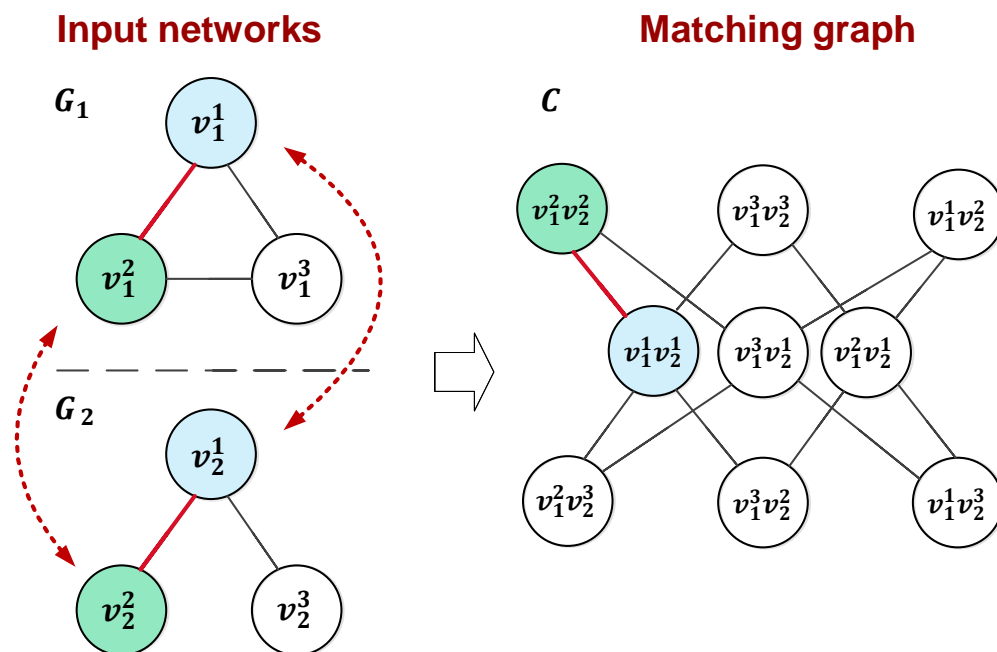
Local vs. Global consistency

- Network matching: matching users' ego networks



Network Matching

- Network matching: matching users' ego networks



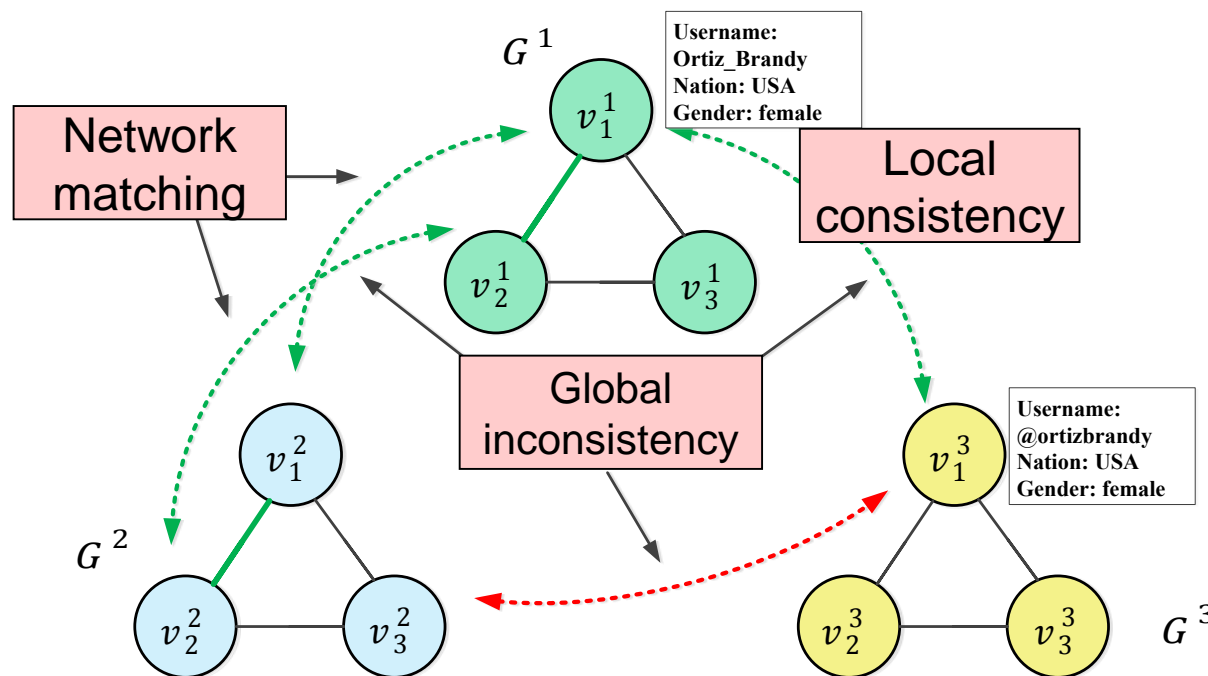
Energy function

$$E_e(Y, X) = \sum_{\langle \mathbf{x}_i, \mathbf{x}_j \rangle \in E_{MG}} \mathbf{w}_e^T \mathbf{f}_e(y_i, y_j)$$

$$\mathbf{f}_e(y_i, y_j) = \begin{cases} (1, 0, 0)^T & \text{if } y_i = y_j = 0 \\ (0, 1, 0)^T & \text{if } y_i + y_j = 1 \\ (0, 0, 1)^T & \text{if } y_i = y_j = 1 \end{cases}$$

Local vs. Global consistency

- Global consistency: matching users by avoiding global inconsistency



DEFINITION 2 (GLOBAL INCONSISTENCY). Given a set of social networks \mathbf{G} , a set of user pairs X and the corresponding labels Y , if there exists a sequence of user pairs $\langle \mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \dots, \mathbf{x}_{i_n} \rangle$, such that

$$\forall i = i_1, i_2, \dots, i_n, y_i = 1$$

and

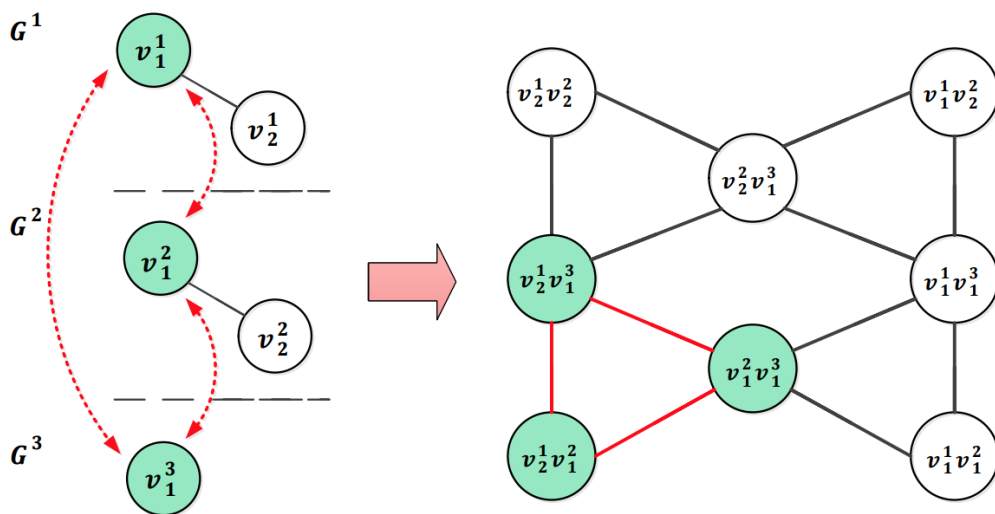
$$\forall k = 1, 2, \dots, n - 1, \mathcal{V}_{i_k}^2 = \mathcal{V}_{i_{k+1}}^1$$

and

For the pair $\langle \mathcal{V}_{i_n}^2, \mathcal{V}_{i_1}^1 \rangle$, if the corresponding label $y_j = 0$ then we say that the assigned labels Y causes global inconsistency given \mathbf{G} and X .

Avoid “global inconsistency”

Avoid global inconsistency

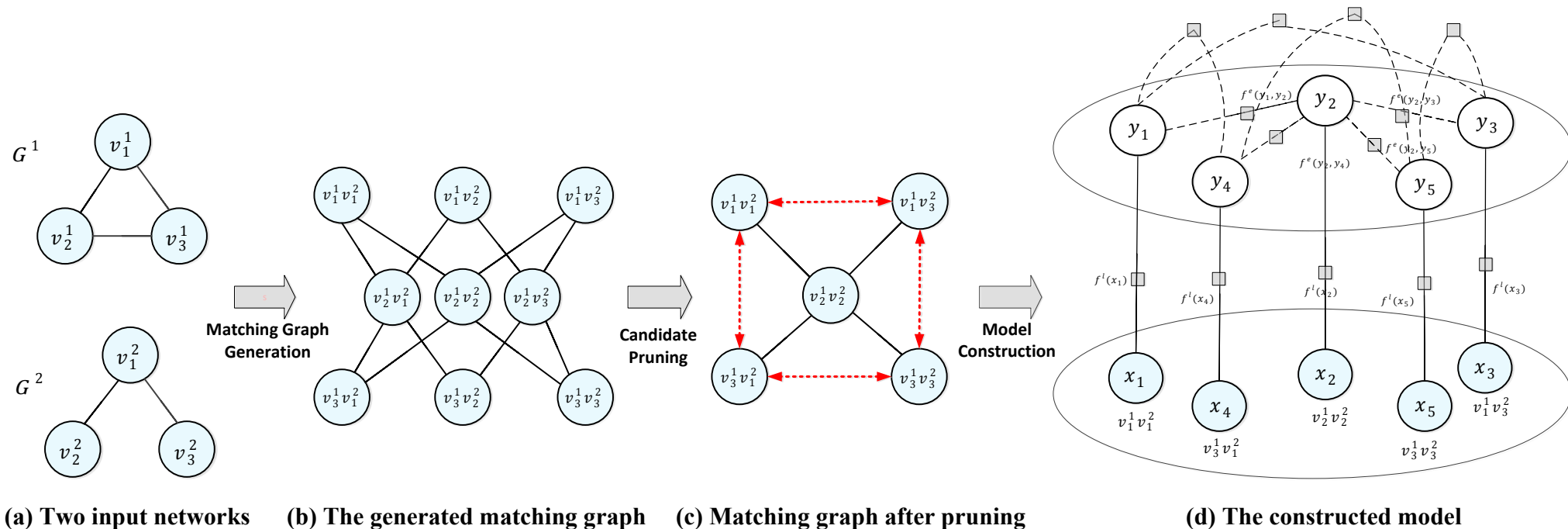


Energy function

$$E_t(Y, X) = \sum_{c \in T_{MG}} \mathbf{w}_t^\top \mathbf{f}_t(Y_c)$$

$$\mathbf{f}_t(y_i, y_j) = \begin{cases} (1, 0, 0, 0)^\top & \text{if } |Y_c| = 0 \\ (0, 1, 0, 0)^\top & \text{if } |Y_c| = 1 \\ (0, 0, 1, 0)^\top & \text{if } |Y_c| = 2 \\ (0, 0, 0, 1)^\top & \text{if } |Y_c| = 3 \end{cases}$$

Model Construction



Objective function by combining all the energy functions

$$\begin{aligned}
 E(Y, X) = & \sum_{\mathbf{x}_i \in V_{MG}} \mathbf{w}_l^T \mathbf{g}_l(\mathbf{x}_i, y_i) + \sum_{\langle \mathbf{x}_i, \mathbf{x}_j \rangle \in E_{MG}} \mathbf{w}_e^T \mathbf{f}_e(y_i, y_j) \\
 & + \sum_{c \in T_{MG}} \mathbf{w}_t^T \mathbf{f}_t(Y_c)
 \end{aligned} \tag{2}$$

Model Learning

- Max-margin learning

$$\begin{aligned} \min_W \quad & \frac{1}{2} \|W\|^2 + \mu\xi \\ \text{s.t.} \quad & E(\hat{Y}, X; W) \leq E(Y, X; W) - \Delta(Y, \hat{Y}) + \xi \end{aligned}$$

- As the original problem is intractable, we use Lagrangian relaxation to decompose the original objective function into a set of easy-to-solve sub-problems

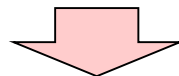
$$\begin{aligned} E(Y, X; W) &= \sum_{f \in \mathcal{F}} E_f(Y_f, X_f; W) \\ &= \sum_{f \in \mathcal{F}} \sum_{\mathbf{x}_i \in X_f} \left(\frac{1}{|\mathcal{F}_i|} \mathbf{w}_i^\top \mathbf{g}_l(\mathbf{x}_i, y_i^f) + \mathbf{w}_f^\top f(Y_f) \right) \\ \text{s.t.} \quad & y_i^f = y_i, \quad \forall f, y_i \in Y_f \end{aligned}$$

Model Learning (cont.)

- Dual decomposition

$$L(Y, X, \lambda; W) = \min_W \sum_{f \in \mathcal{F}} \left(\sum_{y_i \in Y_f} \frac{1}{|\mathcal{F}_i|} \mathbf{w}_l^T \mathbf{g}_l(\mathbf{x}_i, y_i^f) + \mathbf{w}_f^T f(Y_f) \right) + \sum_{f \in \mathcal{F}} \sum_{y_i \in Y_f} \lambda_i^f (y_i - y_i^f)$$

This provides a **lower bound** to the original function



$$\min_{W, \lambda} \frac{1}{2} \|W\|^2 + \mu (E(\hat{Y}, X; W) - \max_{\lambda} L(Y, X, \lambda; W))$$

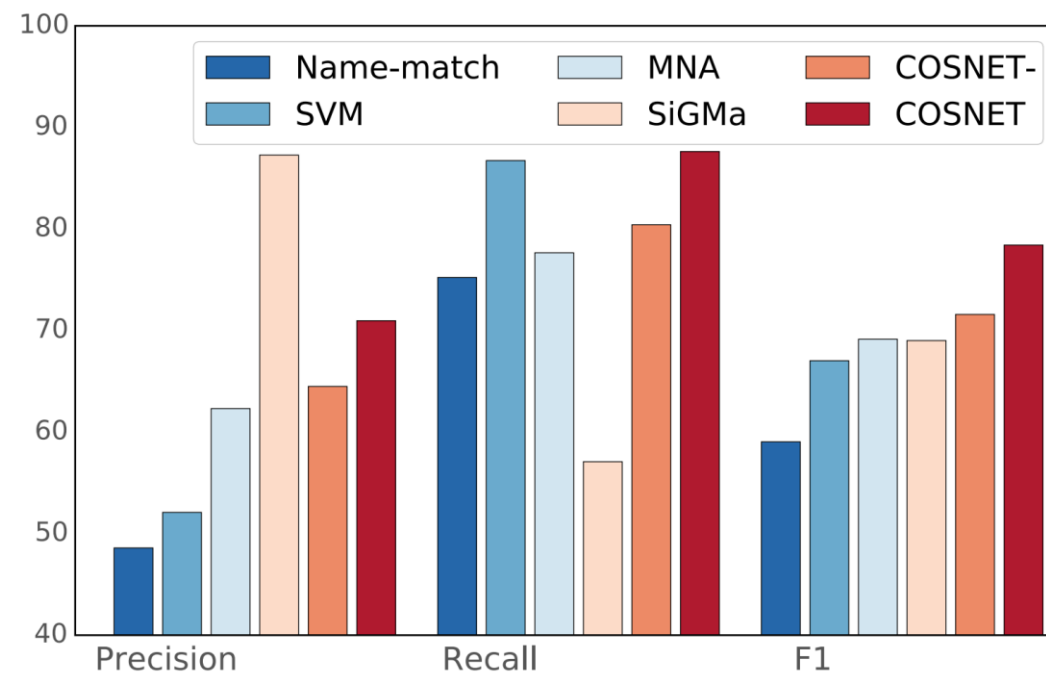
s.t. $\sum_{y_i \in Y_i} \lambda_i^f = 0, \quad \forall f \in \mathcal{F}$

The resulting objective function is convex and non-differentiable, and can be solved by **projected sub-gradient** method

Results

Connecting AMiner with ...

- LinkedIn and VideoLectures



Name-match: match name only;
SVM: use classifier to identify the same user;
MNA: an optimization method;

SiGMa: local propagation;
COSNET: our method;
COSNET-: w/o global consistency.

Person Search



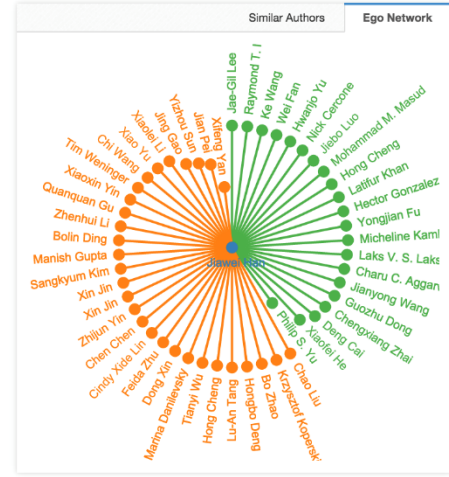
Jiawei Han (韩家炜)

Department of Computer Science, University of Illinois at Urbana-Champaign
 Professor
 (217) 333-6903
 hanj@cs.uiuc.edu
 http://www.cs.uiuc.edu/~hanj/

LinkedIn

External Links:

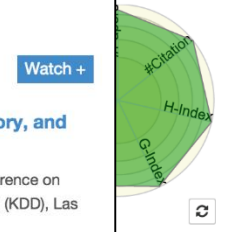
Research Interests: Data Mining, Information Extraction, Data Analysis, Machine Learning, Text Mining



VideoLectures

Bringing Structure to Text: Mining Phrases, Entity Concepts, Topics, and Hierarchies
 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), New York 2014

Mining Massive RFID, Trajectory, and Traffic Data Sets
 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Las Vegas 2008



About

| | |
|----------|-----|
| Papers | 790 |
| Lectures | 13 |
| Patents | 1 |

USPTO

1

Systems and Methods for Detecting a Novel Data Class

Mohammad Mehedy Masud, Latifur Rahman Khan, Bhavani Marienne Thuraisingham, Qing Chen, Jing Gao, Jiawei Han

Publication-date: 2012-03-01 Application-date: 2011-08-22

790

Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions Cited by 4
 Wei Shen, Jianyong Wang, Jiawei Han
 Knowledge and Data Engineering, IEEE Transactions (2015)
<http://dx.doi.org/10.1109/TKDE.2014.2327028>

Patterns. Cited by 20
 ang Yan, Jiawei Han
 Mining (2014)
http://dx.doi.org/10.1007/978-1-4419-6045-0_12

Troubleshooting interactive complexity bugs in wireless sensor networks using data mining techniques Cited by 5
 Mohammad Maffi Hasan Khan, Hieu Khac Le, Hossein Ahmadi, Tarek F. Abdelzaher, Jiawei Han
 ACM Transactions on Sensor Networks (TOSN) (2014)
 Bibtex <http://dx.doi.org/10.1145/2530290>

787

AMiner Today

— A brief summary

ArnetMiner's History

| Date | Version | New Features |
|---------|---------|--|
| 2006/5 | V0.1 | Profile extraction, person/paper/conf. search |
| 2006/8 | V1.0 | Rewritten all codes in Java. |
| 2007/7 | V2.0 | Survey search, research interest, association search |
| 2008/11 | V4.0 | Graph search, topic mining, NSFC/NSF |
| 2009/4 | V5.0 | Bole/course search, profile editing, open resources, |
| 2009/12 | V6.0 | Academic statistics, user feedbacks, refined ranking |
| 2010/5 | V7.0 | Name disambiguation, reviewer assignment, open API |
| 2011/7 | V8.0 | AMiner, location search, conference analysis |
| 2012/3 | V9.0 | New UI, cross-domain collaboration |
| 2013/5 | VII | Knowledge graph, new architecture |
| 2014/10 | VII 2.0 | Organization ranking, conference ranking |
| 2015/4 | VII 3.0 | Network integration, deep learning |

Widely used..

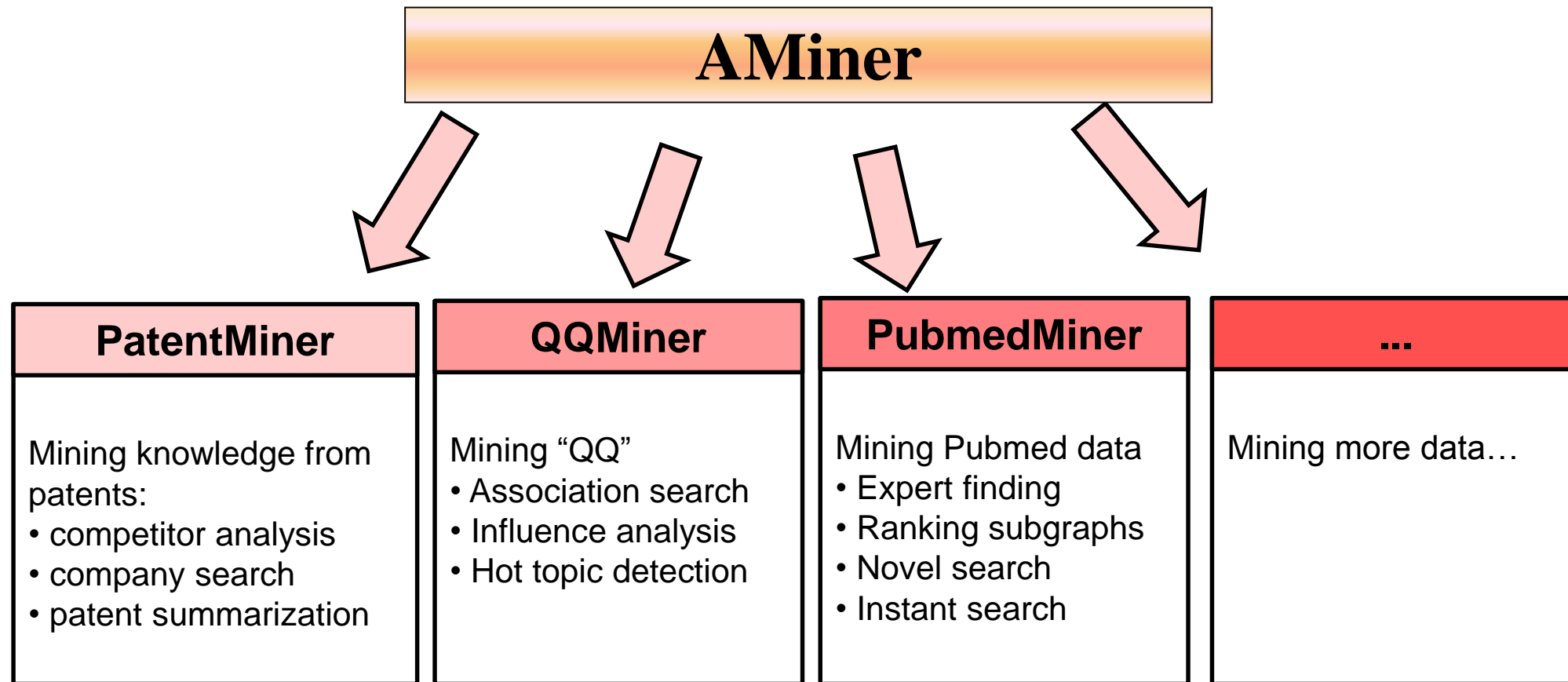
- The largest publisher: Elsevier
- Conferences
 - KDD 2010
 - KDD 2011
 - KDD 2012
 - WSDM 2011
 - ICDM 2011
 - ICDM 2012
 - SocInfo 2011
 - ICMLA 2011
 - WAIM 2011
 - etc.

The screenshot shows the WSDM 2011 website interface. At the top, there's a navigation bar with 'Home', 'About', 'Committee', 'Authors', 'Attendees', 'Program', and 'Sponsors'. Below that, a search bar contains 'data mining' with 'Search' and 'Search Tips' buttons. The search results show 'About 54317 results for: ALL(data mining)'. A sidebar on the left offers 'Refine Results' with 'Limit to' and 'Exclude' options, and 'Content Sources' with checkboxes for Scien, Scopus, Paten, Digita, and MD Ci. The main content area features a banner for 'Organizing Committee' with 'General Chair' Irwin King. Below this is a 'Conference Organizers' section listing:

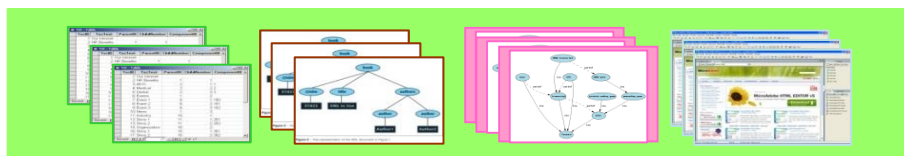
- General Chair:** gc@kdd2011.com, Chidanand Apte (IBM Research Division)
- Program Co-Chairs:** program-chairs@kdd2011.com, Joydeep Ghosh (Electrical and Computer Engineering, University of Texas, Austin), Padhraic Smyth (School of Information and Computer Sciences, Center for Machine Learning and Intelligent Systems, University of California, Irvine)
- Industry Track Co-Chairs:** industrial@kdd2011.com, Ted E. Senator (SAIC), Michael Zeller (Zementis)
- Industry Practice Expo Co-Chairs:** industrial_practice@kdd2011.com, Ying Li (adCenter Labs, Microsoft Corporation), Rajesh Parekh (Groupon)

 On the right side, there are sections for 'QUICK LINKS' (Instructions for Authors, Registration, Accommodations, Travel Grants Information, Tutorial Information, Workshop Information), 'IMPORTANT DATES' (Camera-ready deadline, Google travel grant deadline, Author registration deadline, Early-bird registration deadline, Google travel grant notification, Student travel grant deadline, Group rate hotel reservation deadline, Student travel grant notification), and 'SPONSORS' (Platinum Sponsors: Microsoft Research; Gold Sponsors: Google; Technical Sponsors: Arnetminer).

AMiner as a platform...



Opportunity: exploiting social network and semantic web in the real-world



Web, relational data,
ontological data,
social data

Data Mining and Social Network techniques

Scientific Literature

Users cover >180 countries
>600K researcher
>3M papers

Arnetminer.org
(NSFC, 863)

Social search & mining

Social extraction
Social mining

**IBM US, Tencent
IBM CRL**

Advertisement

Advertisement
Recommendation

Sohu

Mobile Context

Mobile search
& recommendation

Nokia

Energy trend analysis

Energy product
Evolution
Techniques
Trend

Oil Company

Large-scale Mining

Scalable algorithms
for message tagging
and community
Discovery

Google

国家核高基项目 (NSFC)
自然科学基金重点 (NSFC Key)
科技信息资源内容监测与分析服务平台 (中国科技部信息情报研究所)

Search, browsing, complex query, integration, collaboration, trustable analysis, decision support, intelligent services,



Representative Publications

- Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. **KDD'08**. (**Top 6** cited papers among KDD 2008's papers)
- Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. **KDD'09**. (**Top 4** cited papers among KDD 2009's papers)
- Chi Wang, Jiawei Han, Yuntao Jia, Jie Tang, Duo Zhang, Yintao Yu, Jingyi Guo. Mining Advisor-Advisee Relationships from Research Publication Networks. **KDD'10**.
- Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. Cross-domain Collaboration Recommendation. **KDD'12** (Full Presentation & Best Poster Award)
- Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei, and Philip Yu. COSNET: Connecting Heterogeneous Social Networks with Local and Global Consistency. **KDD'15**.
- Jie Tang, Limin Yao, Duo Zhang, and Jing Zhang. A Combination Approach to Web User Profiling. ACM **TKDD**, 2010.
- Jie Tang, Jing Zhang, Ruoming Jin, Zi Yang, Keke Cai, Li Zhang, and Zhong Su. Topic Level Expertise Search over Heterogeneous Networks. **Machine Learning Journal**, 2011.
- Jie Tang, A.C.M. Fong, Bo Wang, and Jing Zhang. A Unified Probabilistic Framework for Name Disambiguation in Digital Library. IEEE **TKDE**, 2012.

Thanks!

AMiner.org

Jie Tang, KEG, Tsinghua U,
Download data & Codes,

<http://keg.cs.tsinghua.edu.cn/jietang>
<http://aminer.org/download>